



МИНОБРНАУКИ РОССИИ

федеральное государственное бюджетное образовательное учреждение
высшего образования

«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ФГБОУ ВО «ИГУ»

Кафедра физико-химической биологии, биоинженерии и биоинформатики



Рабочая программа дисциплины

Наименование дисциплины: Б1.О.37 **«ИСКУСТВЕННЫЙ ИНТЕЛЛЕКТ И
МАШИННОЕ ОБУЧЕНИЕ»**

Специальность: 06.05.01 «Биоинженерия и биоинформатика»

Специализация: «Биоинженерия и биоинформатика»

Квалификация выпускника: биоинженер и биоинформатик

Форма обучения: очная с элементами электронного обучения и дистанционных образовательных технологий

Согласовано с УМК биолого-почвенного
факультета
Протокол № 4 от 20.03.2024
Председатель А. Н. Матвеев

Рекомендовано кафедрой физико-химической
биологии, биоинженерии и биоинформатики
Протокол № 15 от 17.04.2024
Зав. кафедрой В.П. Саловарова

Иркутск 2024 г.

Содержание

	стр.
I. Цель и задачи дисциплины	3
II. Место дисциплины в структуре ОПОП	3
III. Требования к результатам освоения дисциплины	3
IV. Содержание и структура дисциплины	7
4.1 Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов	
4.2 План внеаудиторной самостоятельной работы обучающихся по дисциплине	
4.3 Содержание учебного материала	
4.3.1 Перечень семинарских, практических занятий и лабораторных работ	
4.3.2. Перечень тем (вопросов), выносимых на самостоятельное изучение в рамках самостоятельной работы студентов	
4.4. Методические указания по организации самостоятельной работы студентов	
4.5. Примерная тематика курсовых работ (проектов)	
V. Учебно-методическое и информационное обеспечение дисциплины	14
а) перечень литературы	
б) периодические издания	
в) список авторских методических разработок	
г) базы данных, поисково-справочные и информационные системы.....	
VI. Материально-техническое обеспечение дисциплины	17
6.1. Учебно-лабораторное оборудование	
6.2. Программное обеспечение	
6.3. Технические и электронные средства обучения	
VII. Образовательные технологии	20
VIII. Оценочные материалы для текущего контроля и промежуточной аттестации	21

I. Цель и задачи дисциплины:

Цель: является формирование системного представления, первичных знаний, умений и навыков по основам знаний и области машинного обучения как основному направлению построения систем; искусственного интеллекта. Дать представление практические знания в области использования методов машинного обучения и искусственного интеллекта в области анализ бельгийских данных и предсказания поведения биосистем.

Задачи:

- сформировать способность использовать современные информационные технологии и программные средства, при решении задач профессиональной деятельности с использованием интеллектуальных систем.
- сформировать способность разрабатывать модели биологички процессов и компонентов интеллектуальных систем, включая модели баз знаний;
- изучить спектр алгоритмов и методов машинного обучения и искусственного интеллекта их применением при решении практических задач по моделированию биосистем с помощью библиотек для различных языков программирования и онлайн ресурсов.
- сформировать способность проводить концептуальное, функциональное и логическое проектирование и разрабатывать архитектуру, прототип, дизайн интеллектуальных систем различного масштаба и сложности;
- сформировать способность выполнять работы по разработке, модификации, тестированию, развертыванию, эксплуатации и сопровождению интеллектуальных систем при исследовании различных процессов в области биологии и экологии.
- сформировать способности производить обработку запросов на анализ биологических данных и биосистем с помощью методов машинного обучения и искусственного интеллекта.

II. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

2.1. Учебная дисциплина Б1.О.37 **«Искусственный интеллект и машинное обучение»** относится к обязательной части образовательной программы.

2.2. Для изучения данной учебной дисциплины необходимы знания, умения и навыки, формируемые предшествующими дисциплинами: «Основы программирования», «Математика», «Физика», «Информатика», «Иностранный язык», «Математический анализ», «Специальные главы математики».

2.3. Перечень последующих учебных дисциплин, для которых необходимы знания, умения и навыки, формируемые данной учебной дисциплиной: «Структурно-функциональная биоинформатика», «Молекулярная филогенетика», «Геномный и метагеномный анализ», «ДНК метабаркодинг», выполнение ВКР.

III. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование компетенций в соответствии с ФГОС ВО и ОП ВО по данному направлению подготовки 06.05.01 «Биоинженерия и биоинформатика», профиль «Биоинженерия и биоинформатика»:
ОПК-3: Способен проводить экспериментальную работу с организмами и клетками, использовать физико-химические методы исследования макромолекул, математические методы обработки результатов биологических исследований.

ОПК-5: Способен находить и использовать информацию, накопленную в базах данных по биологическим объектам, включая нуклеиновые кислоты и белки, владеть основными биоинформатическими средствами анализа

Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
ОПК-3 Способен проводить экспериментальную работу с организмами и клетками, использовать физико-химические методы исследования макромолекул, математические методы обработки результатов биологических исследований;	ИДК ОПК-3.1 Проводит экспериментальную работу с организмами и клетками с использованием физико-химических методов исследования макромолекул	Знать: основные математические понятия и методы, применимые для анализа биологички систем и биологических данных с применением алгоритмов с применением машинного обучения. Уметь: адекватно выбрать математический метод и алгоритмы для описания биологической системы и биологического процесса в эволюционной биологии геномики, биоинформатики и экологии с применением машинного обучения. Владеть: основными принципами формализации сложных конвейеров анализа данных с применением всех рассмотренных алгоритмов машинного обучения.
	ИДК ОПК-3.2 Демонстрирует практические навыки математических методов обработки результатов экспериментальных исследований	Знать: цель, основные задачи и области применения алгоритмов биоинформатики в рамках направления подготовки. Уметь: формализовать процесс обработки данных в геномики, эволюционной биологии, экологии и других биологических дисциплинах в виде конвейеров различных вычислительных алгоритмов машинного обучения. Владеть: методами применения разработанных алгоритмов и конвейеров анализа данных при исследовании биологических процессов и биосистем с применением машинного обучения.
	ИДК ОПК-3.3 Владеет опытом применения методов для исследования макромолекул, обработки результатов биологических исследований, прогнозирования перспектив и социальных последствий своей профессиональной деятельности.	Знать: спектр алгоритмов машинного обучения, применяемых при анализе биомолекул. Уметь: выбирать спектр программ и алгоритмов, использующих машинное обучение для исследования биомоллекул для решения поставленных биологических задач. Владеть: методами реконструкции структур биомоллекул с применением методов машинного обучения.
ОПК-5 Способен находить и использовать информацию, накопленную в базах данных по	ИДК ОПК 5.1 Использует информацию, накопленную в базах данных по структуре геномов, белков и другую биологическую информацию	Знать: информацию, накопленную в базах данных по структуре геномов, белков и другую биологическую информацию Уметь: применять информацию, накопленную в базах данных по структуре геномов, белков и другую биологическую информацию

биологическим объектам, включая нуклеиновые кислоты и белки, владеть основными биоинформатическими средствами анализа;		Владеть: методами анализа информации, накопленной в базах данных по структуре геномов, белков и другую биологическую информацию
	Умеет применять биоинформатические методы и полученные знания для анализа геномной, структурной и иной информации	Знать: биоинформатические методы и полученные знания для анализа геномной, структурной и иной информации Уметь: применять биоинформатические методы и полученные знания для анализа геномной, структурной и иной информации Владеть: биоинформатическими методами для анализа геномной, структурной и иной информации
	Демонстрирует навыки владения основными биоинформатическими средствами анализа геномной, структурной и иной информации и способен критически оценивать развитие научных идей	Знать: основные биоинформатические средства анализа геномной, структурной и иной информации и способен критически оценивать развитие научных идей. Уметь: применять основные биоинформатические средства анализа геномной, структурной и иной информации и способен критически оценивать развитие научных идей. Владеть: основными биоинформатическими средствами анализа геномной, структурной и иной информации и способен критически оценивать развитие научных идей.

IV. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Объем дисциплины составляет 7 зачетных единицы, 252 часов.

Из них реализуется с использованием электронного обучения и дистанционных образовательных технологий 28 часов.

Форма промежуточной аттестации: зачет.

4.1 Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

№ п/п	Раздел дисциплины/тема	Семестр	Всего часов	Из них практическая подготовка обучающихся	Виды учебной работы, включая самостоятельную работу обучающихся , практическую подготовку и трудоемкость (в часах)				Форма текущего контроля успеваемости/ Форма промежуточной аттестации (по семестрам)
					Контактная работа преподавателя с обучающимися			Самостоятельн ая работа	
					Лекция	Семинар/ Практическое, лабораторное занятие/	Консультация		
1	2	3	4	5	6	7	8	9	10
1	Тема 1.Введение, что такое машинное обучение.	5	10	4	2	4		6	КСР
2	Тема 2. Обучение с учителем. Ленная регрессия.	5	10	4	2	4		4	КСР
3	Тема 3. Обучение с учителем. Линейная классификация.	5	12	4	2	4		6	КСР
4	Тема 4. Обучение с учителем. Метод к ближайших соседей.	5	12	4	2	4		4	КСР
5	Тема 5. Обучение с учителем. Деревья решений.	5	10	4	2	4		6	КСР
6	Тема 6. Обучение с учителем. Ансамбли	5	12	4	2	4		4	КСР

	деревьев решений.								
7	Тема 7. Обучение с учителем. Метод опорных векторов.	5	10	4	2	4		4	KCP
8	Тема 8. Обучение с учителем. Наивный Байесовский классификатор	5	10	4	2	4		4	KCP
9	Тема 9. Обучение с учителем. Нейронные сети (глубокое обучение).	5	12	4	2	4		6	KCP
10	Тема 10. Обучение без учителя. Типы машинного обучения без учителя.	6	14	4	2	4		8	KCP
11	Тема 11. Обучение без учителя. Предварительная обработка и масштабирование данных.	6	14	4	2	4		8	KCP
12	Тема 12. Обучение без учителя. Анализ методом главных компонент.	6	14	4	2	4		8	KCP
13	Тема 13. Обучение без учителя. Кластеризация k средних.	6	14	4	2	4		8	KCP
14	Тема 14. Обучение без учителя. Иерархическая кластеризация.	6	14	4	2	4		8	KCP
15	Тема 15. Обучение без учителя. Оценки достоверности кластеризации.	6	14	4	2	4		8	KCP
16	Тема 16. Оценка и улучшения качества моделей машинного обучения	6	14	4	2	4		8	KCP
14	Тема 17. Объединение алгоритмов в цепочки и конвейеры.	6	17	4	2	4	1	10	KCP

4.2 План внеаудиторной самостоятельной работы обучающихся по дисциплине

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Трудоемкость (час.)		
5	Тема 1. Введение, что такое машинное обучение.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по теме линейная регрессия в машинном обучении.	1	6	КСР	Раздел 5 а-г
5	Тема 2. Обучение с учителем. Линейная регрессия.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по теме линейная регрессия в машинном обучении.	2	4	КСР	- « -
5	Тема 3. Обучение с учителем. Линейная классификация.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по теме линейная классификация в машинном обучении.	4	6	КСР	- « -
5	Тема 4. Обучение с учителем. Метод k ближайших соседей.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию метода k ближайших соседей в машинном обучении..	5	4	КСР	- « -
5	Тема 5. Обучение с учителем. Деревья решений.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию деревьев в машинном обучении.	6	6	КСР	- « -
5	Тема 6. Обучение с учителем. Ансамбли деревьев решений.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию ансамблей деревьев в машинном обучении.	8	4	КСР	- « -
5	Тема 7. Обучение с учителем. Метод опорных векторов.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию метода опорных векторов в машинном обучении.	9	4	КСР	- « -
5	Тема 8. Обучение с учителем. Наивный Байесовский классификатор	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию наивного Байесовского классификатора.	10	4	КСР	- « -

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Трудоемкость (час.)		
5	Тема 9. Обучение с учителем. Нейронные сети (глубокое обучение).	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по построению нейронных сетей.	11	6	КСР	- « -
6	Тема 10. Обучение без учителя. Типы машинного обучения без учителя.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по теме.	1	8	КСР	
6	Тема 11. Обучение без учителя. Предварительная обработка и масштабирование данных.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию методов масштабирования данных.	2	8	КСР	
6	Тема 12. Обучение без учителя. Анализ методом главных компонент.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий с использованием метода главных компонент в машинном обучении.	4	8	КСР	
6	Тема 13. Обучение без учителя. Кластеризация k средних.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию алгоритма кластеризации k средних.	5	8	КСР	
6	Тема 14. Обучение без учителя. Иерархическая кластеризация.	1. Разбор темы лекции и практического занятия. 2. Решение домашних заданий по использованию методов иерархической кластеризации при анализе данных.	6	8	КСР	

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Трудоемкость (час.)		
6	Тема 15. Обучение без учителя. Оценки достоверности кластеризации.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по анализу достоверности результатов кластеризации.	8	8	КСР	
6	Тема 16. Оценка и улучшения качества моделей машинного обучения	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по анализу качества моделей машинного обучения для анализа заданных массивов данных..	9	8	КСР	
6	Тема 17. Объединение алгоритмов в цепочки и конвейеры.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание разработке простейшего конвейера по анализу данным методами машинного обучения.	11	10	КСР	
Общий объем самостоятельной работы по дисциплине (час) – 110						
Из них объем самостоятельной работы с использованием электронного обучения и дистанционных образовательных технологий 110 часов.						

4.3 Содержание учебного материала

Тема 1. Введение, что такое машинное обучение. В рамках темы дается понятие «машинному обучению» как области искусственного интеллекта, которая изучает методы и алгоритмы, позволяющие вычислительным машинам при использовании специализированного программного обеспечения "обучаться" на данных и выполнять задачи построения моделей, для предсказания поведения биосистем, принятия решения и формулировки выводов.

Тема 2. Обучение с учителем. Ленная регрессия. В рамках темы рассматривается один из самых простых и популярных методов обучения с учителем, используемый для решения задач регрессии. Регрессия — это задача предсказания непрерывной целевой переменной на основе входных признаков. Рассматриваются примеры использования метода при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 3. Обучение с учителем. Линейная классификация. В рамках темы рассматривается один из методов обучения с учителем, который используется для задач классификации. Изучается линейная классификация как конструируемая модель пытается разделить данные на классы с помощью линейной границы (прямой или плоскости) в двухмерном или многомерном пространстве. Рассматриваются примеры использования метода при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 4. Обучение с учителем. Метод k ближайших соседей. В рамках темы рассматривается алгоритм обучения с учителем, используемый для задач классификации и регрессии. Изучается идея алгоритма которая заключается в том, чтобы предсказать значение целевой переменной для нового объекта на основе значений его ближайших соседей в обучающих данных. . Рассматриваются примеры использования метода при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 5. Обучение с учителем. Деревья решений. В рамках темы рассматривается алгоритм обучения с учителем, используемый для задач классификации и регрессии. Изучается принцип работы алгоритма, основанный построении иерархической структуры, где каждый внутренний узел представляет собой условие на основе признаков, каждая ветвь — результат этого условия, а каждый лист — предсказанное значение (класс или численное значение). Рассматриваются примеры использования метода

при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 6. Обучение с учителем. Ансамбли деревьев решений. В рамках темы рассматривается метод обучения с учителем дерева решений и ансамбли деревьев решений как методы, которые широко используются для задач классификации и регрессии. Рассматриваются ансамбли деревьев решений как алгоритм, объединяющий несколько деревьев для улучшения точности и устойчивости модели. Рассматриваются примеры использования метода при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 7. Обучение с учителем. Метод опорных векторов. В рамках темы рассматривается метод обучения с учителем - метод опорных векторов (Support Vector Machines, SVM), используемый для задач классификации и регрессии. Изучается основная идея SVM, которая заключается в нахождении гиперплоскости, которая максимально разделяет классы в пространстве признаков. Рассматриваются примеры использования метода при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 8. Обучение с учителем. Наивный Байесовский классификатор. В рамках темы рассматривается метод обучения с учителем - наивный Байесовский классификатор как алгоритм машинного обучения, основанный на теореме Байеса. Алгоритм широко используется для задач классификации, особенно в текстовой аналитике – в биологии для анализа текстов белгийских последовательностей (ДНК, РНК и белков). Рассматриваются примеры использования метода при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 9. Обучение с учителем. Нейронные сети (глубокое обучение). В рамках темы рассматриваются нейронные сети как один из алгоритмов глубокого обучения. Изучаются вопросы нейронных сетей для решения сложных задач, таких как классификация, регрессия, обработка изображений, обработка биологических текстов. Нейронные сети рассматриваются как системы, состоящие из множества слоев, которые позволяют модели автоматически извлекать иерархические признаки из данных. Рассматриваются примеры использования метода при анализе реальных биологических и экологических данных, при реализации с помощью библиотек языков программирования R или Python.

Тема 10. Обучение без учителя. Типы машинного обучения без учителя. В рамках темы дается общая характеристика классам алгоритмов машинного

обучения машинного обучения, в котором модель обучается на данных без явных меток или целевых переменных. В отличие от обучения с учителем, где модель учится предсказывать выходные значения на основе входных данных, в обучении без учителя цель состоит в том, чтобы обнаружить скрытые структуры, закономерности или группировки в данных.

Тема 11. Обучение без учителя. Предварительная обработка и масштабирование данных. В рамках темы рассматриваются различные алгоритмы масштабирования данных, работа которых заключается в приведении признаков к единой шкале изменения переменных. Рассматриваются вопросы как масштабирование помогает ускорить сходимость алгоритмов, улучшить их производительность и избежать доминирования признаков с большими значениями. Рассматриваются примеры реализаций алгоритмов масштабирования с помощью библиотек языков программирования R или Python.

Тема 12. Обучение без учителя. Анализ методом главных компонент. В рамках темы рассматривается статистический метод — метод главных компонент (PCA, Principal Component Analysis), используемый для уменьшения размерности данных, который помогает выявить основные направления (компоненты), объясняющие наибольшую вариацию в данных. Изучаются основные этапы реализации метода для анализа биологических данных. Рассматриваются примеры реализаций метода главных компонент с помощью библиотек языков программирования R или Python.

Тема 13. Обучение без учителя. Кластеризация k средних. В рамках темы изучается кластеризация методом k-средних (k-means clustering) — это один алгоритмов обучения без учителя, используемый для группировки данных на основе их сходства. Изучается основная цель метода кластеризации k-средних — разделить данные на k кластеров, где каждый кластер представляет собой группу объектов, близких друг к другу в пространстве признаков. Рассматриваются примеры реализаций метода для анализа данных биологии и экологии с помощью библиотек языков программирования R или Python.

Тема 14. Обучение без учителя. Иерархическая кластеризация. В рамках темы рассматривается иерархическая кластеризация — это метод обучения без учителя, который строит иерархию кластеров, представляя данные в виде древовидной структуры (дендрограммы). Изучаются преимущества и недостатки метода при анализе различных типов данных. Рассматриваются примеры реализаций различных методов иерархической кластеризации для анализа данных биологии и экологии с помощью библиотек языков программирования R или Python.

Тема 15. Обучение без учителя. Оценки достоверности кластеризации. В рамках темы рассматриваются различные подходы и алгоритмы, которые применяются для оценки достоверности и качества кластеризации с использованием различных индексов. Рассматриваются примеры реализаций различных методов оценки качества кластеризации при анализа данных биологии и экологии с помощью библиотек языков программирования R или Python.

Тема 16. Оценка и улучшения качества моделей машинного обучения. В рамках темы рассматриваются различные методы и метрики (среднеквадратичная ошибка, средняя абсолютная ошибка, коэффициент детерминации, индекс силуэта, кросс-валидация) оценки качества моделей машинного обучения, которые напрямую влияют на эффективность их применения. Рассматриваются примеры реализаций различных метрик оценки качества моделей машинного обучения при анализа данных биологии и экологии с помощью библиотек языков программирования R или Python.

Тема 17. Объединение алгоритмов в цепочки и конвейеры. В рамках темы рассматривается комплексный подход к анализу в машинном обучении, который позволяет комбинировать несколько этапов обработки данных и моделирования в единый процесс. Это особенно полезно для автоматизации и упрощения работы с данными, а также для повышения воспроизводимости и эффективности моделей. Рассматриваются примеры реализаций конвейеров для при анализа данных биологии и экологии с помощью библиотек языков программирования R или Python.

4.3.1. Перечень семинарских, практических занятий и лабораторных работ

№ п/п	№ раздела и темы	Наименование семинаров, практических и лабораторных работ	Трудоемкость (час.)		Оценочные средства	Формируемые компетенции (индикаторы)*
			Всего часов	Из них практическая подготовка		
1	2	3	4	5	6	7
1	Тема 1	Решение задач по теме введение в языки программирования R и Python как инструменты реализации методов машинного обучения	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
2	Тема 2	Решение задач по теме логистическая регрессия	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3

						ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
3	Тема 3	Решение задач по теме линейная классификация.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
4	Тема 4	Решение задач по теме метод k ближайших соседей.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
5	Тема 5	Решение задач по теме анализ данных с помощью дерева решений.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
6	Тема 6	Решение задач по теме анализ данных с помощью ансамблей деревьев решений	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
7	Тема 7	Решение задач по теме анализ данных методом опорных векторов.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
8	Тема 8	Решение задач по теме наивный Байесовский классификатор	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5

						ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
9	Тема 9	Решение задач по теме конструирование нейронных сетей для анализа биологических данных	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
10	Тема 10	Решение задач по теме введение в языки программирования R и Python как инструменты реализации методов машинного обучения без учителя	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
11	Тема 11	Решение задач по теме масштабирование данных.	4	4	Опрос КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
12	Тема 12	Решение задач по теме анализ методом главных компонент.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
13	Тема 13	Решение задач по теме кластеризация k средних.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
14	Тема 14	Решение задач по теме иерархическая кластеризация.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3

15	Тема 15	Решение задач по теме оценка достоверности кластеризации.	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
16	Тема 16	Решение задач по теме оценка и улучшения качества моделей машинного обучения	4	4	КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3
17	Тема 17	Решение задач по теме разработка цепочек и конвейеров для анализа данных методами машинного обучения.	4	4	Опрос КСР	ОПК-3 ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ОПК-5 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3

4.3.2. Перечень тем (вопросов), выносимых на самостоятельное изучение студентами в рамках самостоятельной работы (СРС)

№ п/п	Тема	Задание	Формируемая компетенция	ИДК
1.	Тема 9. Обучение с учителем. Нейронные сети (глубокое обучение).	Самостоятельное изучение темы – использование нейронных сетей для распознавания образов и идентификации биологических видов.	ОПК-3 ОПК-5	ИДК ОПК-3.1 ИДК ОПК-3.2 ИДК ОПК-3.3 ИДК ОПК-5.1 ИДК ОПК-5.2 ИДК ОПК-5.3

4.4. Методические указания по организации самостоятельной работы студентов

Самостоятельная работа студентов является составной частью учебного процесса и имеет целью закрепление и углубление полученных знаний и навыков, поиск и приобретение новых знаний, а также выполнение учебных заданий, подготовку к предстоящим занятиям, и экзамену по предмету.

Для организации самостоятельной работы по дисциплине «Искусственный интеллект и машинное обучение» используются следующие формы самостоятельной учебной работы:

- Работа по изучению темы с использованием материалов практического занятия.

- Подбор, изучение, анализ рекомендованной литературы.
- Изучения тем занятий, вынесенных на самостоятельное изучение, подготовка отчета по решению задач по темам, выносимы на самостоятельное изучение.

Самостоятельное решения домашних задач по анализу данных на основе опыта, полученного на практических занятиях.

- Подготовка письменных отчетов по решению домашних задач и загрузка отчетов на образовательный портал ИГУ.

Письменный отчет по решению домашних заданий – это отчет о выполнении домашнего задания по темам дисциплины, содержащий следующую информацию:

- Ф.И.О. номер группы студента;
- номер задания;
- формулировка задания;
- описание хода решения задания;
- описание результат решения задания с приведением таблиц и рисунков в соответствии с формулировкой задания.

Критерий оценки отчета по решению домашнего задания:

- Оценка «зачтено». Задание выполнено правильно и в полном объёме, все таблицы и графики согласно формулировке задания предоставлены в отчете.
- Оценка «не зачтено». Задание выполнено неправильно или не в полном объёме, вопрошается на переделку и доработку.

Подготовка к экзамену в виде тестирования. К экзамену в виде тестирования допускаются студенты, получившие зачеты по всем самостоятельным заданиям.

4.5. Примерная тематика курсовых работ (проектов): не предусмотрены учебным планом.

V. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

а) перечень литературы

1. Леск А. Введение в биоинформатику : пер. с англ. / А. М. Леск ; ред.: А. А. Миронов, В. К. Швядаса. - М. : Бинум. Лаборатория знаний, 2009. - 318 с. - ISBN 978-5-94774-501-6 (8 экз.)
2. Приставка А. А. Большой практикум по биоинженерии и биоинформатике [Текст] : учеб.-метод. пособие : в 3 ч. / А. А. Приставка, В. П. Саловарова - Иркутск : Изд-во ИГУ, 2013. - Ч. 1 : Белки. - 2013. - 121 с. - ISBN 978-5-9624-0962-7 (69 экз.)
3. Белькова Н.Л. Большой практикум по биоинженерии и биоинформатике [Текст] : учеб.-метод. пособие : в 3 ч. / Н. Л. Белькова. - Иркутск : Изд-во ИГУ, 2013. - ISBN 978-5-9624-0956-6. Ч. 2 : Нуклеиновые кислоты. - 2014. - 155 с. - ISBN 978-5-9624-1184-2 (39 экз.)

б) дополнительная литература

1. Игнасимуту С. Основы биоинформатики / С. Игнасимуту ; пер. с англ. А. А. Чумичкин. - Ижевск : Регулярная и хаотическая динамика : Ин-т компьютер. исслед., 2007. - 316 с. - ISBN 978-5-93972-620-7 (1 экз.)
2. Каменская М.А. Информационная биология / М.А. Каменская. – М.: Академия, 2006. – 361 с. - - ISBN 5-7695-2580-0 (8 экз.)
3. Компьютеры и суперкомпьютеры в биологии / Под ред. В.Д. Лахно, М.Н. Устинин. – Москва-Ижевск: Институт компьютерных исследований, 2002. – 528 с. - ISBN 5-93972-188-5 (2 экз.)

4. Математические методы для анализа последовательностей ДНК. / Под ред. М.С. Уотермена, перевод с англ. – М.: Мир, 1999. – 349 с. - ISBN 5030025200 (1 экз.)
5. Паун Г. ДНК-компьютер. Новая парадигма вычислений / Г. Паун, Г. Розенберг, А. Саломаа ; Пер. с англ. Д. С. Ананичева, И. С. Киселевой, О. Б. Финогеновой, ред. М. В. Волков. - М. : Мир, 2004. - 527 с. - ISBN 5-03-003480-3 (1 экз.).
6. Структура и функционирование белков: применение методов биоинформатики / пер. с англ.: В. Н. Новоселецкий, Е. Д. Балицкая, Т. В. Науменкова ; ред. В. Н. Новоселецкий. - М. : УРСС : Ленанд, 2014. - 414 с. - ISBN 978-5-9710-0842-2. - ISBN 978-5-453-00057-9 (1 экз.)
7. Шипунов А. Б., Балдин Е. М., Волкова П.А., и др. Наглядная статистика. Используем R! Издательство: ДМК Пресс, 2014 – 300 с. Книга доступна в свободном доступе по ссылке: <http://ashipunov.info/shipunov/school/books/rbook.pdf>
- 8.

в) периодические издания

1. <https://www.matbio.org/> - сайт журнала «Математическая биология и биоинформатика». Содержит большое количество статей в pdf – формате.
2. <https://journal.r-project.org/> - сайт журнала по статистическим методам на R, «The R Journal».

г) базы данных, информационно-справочные и поисковые системы

1. <http://dmb.biophys.msu.ru> - Информационная система «Динамические модели в биологии», рассчитанная на широкий круг пользователей, включает в себя гипертекстовые документы и реляционные базы данных и обеспечивает унифицированный доступ к разнообразной информации по данной предметной области.
2. <http://www.jcabi.ru/> - сайт объединенного центра вычислительной биологии и биоинформатики
3. <http://mathmod.aspu.ru/> - Сайт совместной лаборатории Института математических проблем биологии Российской академии наук и Астраханского государственного университета
4. <http://www.exponenta.ru/> - образовательный математический сайт
5. <http://www.library.biophys.msu.ru/MathMod/BM.HTML> - книга Г.Ю. Ризниченко «Биология математическая»
6. <http://nature.web.ru/db/msg.html?mid=1156624&uri=index.htm> - Бейли Н.. Математика в биологии и медицине. – М.: Мир, 1970.
7. <http://www.biometrika.tomsk.ru/> - электронный журнал «Биометрика» для медиков и биологов – сторонников доказательной биомедицины. Содержит большое количество статей и иных материалов, посвященных математическим моделям в биологии.
8. <http://www.library.biophys.msu.ru/FominBerk/main.htm> - Фомин С.В., Беркинблит М.Б. Математические проблемы в биологии. - М.: Гаука, 1973. - 200 с.
9. <https://www.elibrary.ru> – электронная библиотека научных статей, монографии и материалов конференций, выпущенных Российскими учеными.
10. <https://pubmed.ncbi.nlm.nih.gov/> - международная база данных научных статей и монографий, посвященная различным вопросам биологии.

11. <https://apps.webofknowledge.com> – международная база данных, индексирующая научные публикации в высокорейтинговых изданиях

VI. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Учебно-лабораторное оборудование:

- Аудитория для проведения занятий лабораторного типа. Компьютерный класс (учебная аудитория). Аудитория оборудована: специализированной (учебной) мебелью на 20 посадочных мест, доской меловой; оборудована техническими средствами обучения: Системный блок PentiumG850, Монитор BenQ G252HDA-1 шт.; Системный блок Athlon 2 X2 250, Монитор BenQ G252HDA – 8 шт.; Системный блок PentiumD 3.0GHz, Монитор Samsung 740N – 3 шт.; Моноблок IRU T2105P – 2 шт.; Системный блок Pentium G3250, Монитор BenQG955 – 1 шт.; Системный блок Pentium G3250, Монитор BenQ GL2250 – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T200 HD – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T190N – 1 шт.; Системный блок Pentium G3250, Монитор Samsung 740N – 1 шт.; Проектор BenQ MX503; экран ScreenVtdiaEcot. С неограниченным доступом к сети Интернет и обеспечением доступа в электронную информационно-образовательную среду организации, учебно-наглядными пособиями, обеспечивающими тематические иллюстрации по дисциплине «Моделирование и программирование биопроцессов» в количестве 8 шт., презентации по каждой теме программы.

- Компьютерный класс (учебная аудитория) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, организации самостоятельной работы. Аудитория оборудована: специализированной (учебной) мебелью на 20 посадочных мест, доской меловой; оборудована техническими средствами обучения: Системный блок PentiumG850, Монитор BenQ G252HDA-1 шт.; Системный блок Athlon 2 X2 250, Монитор BenQ G252HDA – 8 шт.; Системный блок PentiumD 3.0GHz, Монитор Samsung 740N – 3 шт.; Моноблок IRU T2105P – 2 шт.; Системный блок Pentium G3250, Монитор BenQG955 – 1 шт.; Системный блок Pentium G3250, Монитор BenQ GL2250 – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T200 HD – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T190N – 1 шт.; Системный блок Pentium G3250, Монитор Samsung 740N – 1 шт.; с неограниченным доступом к сети Интернет; Проектор BenQ MX503; экран ScreenVtdiaEcot. Ноутбук Lenovo G580 – 1 шт. С неограниченным доступом к сети Интернет.

- Помещения для хранения и профилактического обслуживания учебного оборудования. Аудитория оборудована: специализированной мебелью на 11 посадочных мест; Шкаф для документов - 3 шт.; Сейф – 1 шт.; Шкаф-купе - 2 шт.; Принтер цв.Canon LBR-5050 Laser Printer; Принтер Canon LBP-3010; Ноутбук Lenovo G580 – 1 шт.

6.2. Программное обеспечение:

DreamSpark Premium Electronic Software Delivery (3 years) Renewal (Windows 10 Education 32/64-bit (Russian) - Microsoft Imagine, Windows 7 Professional with Service Pack 1 32/64-bit (English) - Microsoft Imagine, Windows Server 2008 Enterprise and Standard without Hyper-V with SP2 32/64-bit (English) - Microsoft Imagine, Access 2016 32/64-bit (Russian) - Microsoft Imagine, Access 2010 32/64-bit (Russian) - Microsoft Imagine). Договор №03-016-14 от 30.10.2014г.

Kaspersky Endpoint Security для бизнеса - Стандартный Russian Edition. 250-499. Форум Контракт №04-114-16 от 14ноября 2016г KES. Счет №РСЦЗ-000147 и АКТ от 23ноября 2016г Лиц.№1B08161103014721370444.

Microsoft Office Enterprise 2007 Russian Academic OPEN No Level. Номер Лицензии Microsoft 43364238.

Microsoft Windows XP Professional Russian Upgrade Academic OPEN No Level.
Номер Лицензии Microsoft 41059241.

Office 365 профессиональный плюс для учащихся. Номер заказа: 36dde53d-7cdb-4cad-a87f-29b2a19c463e.

6.3. Технические и электронные средства:

Презентации по всем темам курса.

VII. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

При реализации различных видов учебной работы дисциплины используются как стандартные методы обучения, так и интерактивные формы проведения занятий.

Стандартные методы обучения:

1. Информационная лекция.
2. Практические занятия, предназначенные для освоения студентами базовых методов анализа данных и использованию математических методов с помощью методов математического анализа
3. Самостоятельная работа студентов (выполнение домашних заданий, выполнения домашних заданий по тема для самостоятельного изучения, подготовка к экзаменационному тесту).
4. Консультации преподавателя.

Дистанционные образовательные технологии. Под дистанционными образовательными технологиями понимаются образовательные технологии, реализуемые в основном с применением информационно-телекоммуникационных сетей - интернет-технология – задействование образовательного портала ИГУ - educa.isu.ru для предоставления письменных отчетов по домашним работам.

Наименование тем занятий с использованием дистанционных образовательных технологий:

№	Тема занятия	Вид занятия	дистанционная образовательная технология	Кол-во часов
1	Тема 1. Введение, что такое машинное обучение.	Подготовка доклада	Загрузка презентации для контроля на образовательный портал ИГУ educa.isu.ru	6
2	Тема 2. Обучение с учителем. Линейная регрессия.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
3	Тема 3. Обучение с учителем. Линейная классификация.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	6
4	Тема 4. Обучение с учителем. Метод k ближайших соседей.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4

			портал ИГУ educa.isu.ru	
5	Тема 5. Обучение с учителем. Деревья решений.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	6
6	Тема 6. Обучение с учителем. Метод опорных векторов.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
7	Тема 7. Обучение с учителем. Деревья решений, ансамбли деревьев решений.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
8	Тема 8. Обучение с учителем. Наивный Байесовский классификатор	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
9	Тема 9. Обучение с учителем. Нейронные сети (глубокое обучение).	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	6
	Тема 10. Обучение без учителя. Типы машинного обучения без учителя.	Подготовка доклада	Загрузка презентации для контроля на образовательный портал ИГУ educa.isu.ru	8
	Тема 11. Обучение без учителя. Предварительная обработка и масштабирование данных.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	8
	Тема 12. Обучение без учителя. Анализ методом главных компонент.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	8
	Тема 13. Обучение без учителя. Кластеризация k средних.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	8
	Тема 14. Обучение без учителя. Иерархическая кластеризация.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ	8

			educa.isu.ru	
	Тема 15. Обучение без учителя. Оценки достоверности кластеризации.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	8
	Тема 16. Оценка и улучшения качества моделей машинного обучения	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	8
	Тема 17. Объединение алгоритмов в цепочки и конвейеры.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	10
Итого часов				110

VIII. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Оценочные материалы текущего контроля

Оценочные материалы текущего контроля формируются в соответствии с ЛНА университета.

В рамках дисциплины «Искусственный интеллект и машинное обучение» используются следующие формы текущего контроля:

- письменная работа по решению самостоятельных заданий (все формулировки заданий для самостоятельного решения с необходимыми сопроводительными материалами выложены на образовательном портале ИГУ в темах курса «Искусственный интеллект и машинное обучение»);

Перечень посменных работ для самостоятельного выполнения по разделам – темам дисциплины.

Задание по теме 1:

Задание по теме 2:

Задание по теме 3:

Задание по теме 4:

Задание по теме 5:

Задание по теме 6:

Задание по теме 7:

Задание по теме 8:

Задание по теме 9:

Задание по теме 10:

Задание по теме 11:

Задание по теме 12:

Задание по теме 13:

Задание по теме 14:

Задание по теме 15:

Задание по теме 16:

Задание по теме 17:

Оценочные средства для промежуточной аттестации

Промежуточная аттестация по результатам 5 семестр проходит в форме зачета, к которому допускаются студенты, выполнившие в полном объеме аудиторную нагрузку, самостоятельную работу. Студенты, имеющие задолженность, должны выполнить все обязательные виды деятельности.

Фонд оценочных средств для промежуточной аттестации включает:

- список контрольных вопросов с ответами по темам дисциплин, изучаемых в 5 семестре.

Назначение оценочных средств: выявить сформированность компетенций ОПК-3, ОПК-5 (см. п. III).

Тестовое задание включает два варианта по 16 вопросов по 9 темам курса, рассматриваемых в 5 семестре. К зачету допускаются студенты, сдавшие все домашние задания и получившие по каждому заданию зачет.

При проведении зачета, обучающийся получает два случайных вопроса из указанного списка. Устный ответ на каждый вопрос оценивается в балльной шкале от 0 до 50 баллов. Положительным считается итог, когда обучающийся по программе получает за ответы на два вопроса в сумме от 61 до 100 баллов.

Критерии оценки ответа на вопрос по зачету:

Оценка за ответ на вопрос от 41 до 50 баллов ставится, если обучающиеся обнаруживает высокий, продвинутый уровень сформированности компетенций, если он глубоко и прочно усвоил программный материал курса, исчерпывающе, последовательно, четко и логически стройно его излагает, умеет тесно увязывать теорию с практикой, свободно справляется с дополнительными вопросами.

Оценка за ответ на вопрос от 31 до 40 баллов ставится, если обучающийся обнаруживает повышенный уровень сформированности компетенций, твердо знает материал курса, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при ответах на дополнительные вопросы.

Оценка за ответ на вопрос от 20 до 30 баллов ставится, если обучающийся обнаруживает пороговый уровень сформированности компетенций, имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, нарушения логической последовательности в изложении

программного материала, испытывает затруднения при ответах на дополнительные вопросы.

Оценка за ответ на вопрос от 0 до 19 баллов ставится, если обучающийся обнаруживает недостаточное освоения порогового уровня сформированности компетенций, не знает значительной части программного материала, допускает существенные ошибки, не может ответить на дополнительные вопросы.

Оценочные материалы для промежуточной аттестации (зачет)

Список вопросов по темам с краткими ответами

Вопрос 1:

Опишите основные предположения линейной регрессии. Почему важно, чтобы эти предположения выполнялись?

Ответ:

- **Линейность:** Предполагается, что между независимыми переменными и зависимой переменной существует линейная связь.
- **Независимость ошибок:** Остатки (ошибки) должны быть независимы друг от друга.
- **Гомоскедастичность:** Дисперсия остатков должна быть постоянной для всех уровней независимых переменных.
- **Нормальность остатков:** Остатки должны быть нормально распределены.

Нарушение этих предположений может привести к неточным оценкам коэффициентов, ненадежным прогнозам и неправильным статистическим выводам.

Вопрос 2:

Что такое метод наименьших квадратов (МНК) и как он используется в линейной регрессии?

Ответ:

Метод наименьших квадратов (МНК) – это метод, используемый для оценки коэффициентов в модели линейной регрессии. Он минимизирует сумму квадратов разностей между наблюдаемыми значениями зависимой переменной и значениями, предсказанными моделью. Другими словами, МНК пытается найти линию, которая лучше всего "подходит" данным, минимизируя суммарную ошибку.

Вопрос 3:

Объясните разницу между логистической регрессией и линейной регрессией. В каких случаях следует использовать логистическую регрессию вместо линейной?

Ответ:

- **Линейная регрессия** используется для предсказания непрерывных значений. Она моделирует линейную связь между независимыми и зависимой переменными.
- **Логистическая регрессия** используется для задач классификации (бинарной или мультиклассовой). Она предсказывает вероятность принадлежности объекта к определенному классу.

Логистическую регрессию следует использовать, когда зависимая переменная является категориальной, а не числовой.

Вопрос 4:

Что такое разделяющая гиперплоскость в линейной классификации? Как она используется для разделения классов?

Ответ:

Разделяющая гиперплоскость – это $(n-1)$ -мерная плоскость (например, линия в 2D или плоскость в 3D), которая используется для разделения данных по классам в n -мерном

пространстве признаков. В линейной классификации она определяется линейной комбинацией входных признаков. Объекты, расположенные по одну сторону гиперплоскости, относятся к одному классу, а объекты по другую сторону – к другому. Положение гиперплоскости определяется весами модели и смещением (bias).

Вопрос 5:

Опишите принцип работы алгоритма k-NN. Как выбирается значение параметра k , и какое влияние оказывает выбор k на результат классификации?

Ответ:

Алгоритм k-NN классифицирует новый объект, основываясь на классах k ближайших к нему соседей в обучающем наборе данных. Расстояние между объектами обычно вычисляется с использованием метрики Евклида или Манхэттена. Значение k выбирается с использованием методов валидации (например, кросс-валидации). Маленькое значение k делает модель более чувствительной к шуму в данных и может привести к переобучению. Большое значение k сглаживает границы классов, но может привести к недообучению и снижению точности, особенно если классы сильно различаются.

Вопрос 6:

Какие основные преимущества и недостатки алгоритма k-NN? В каких случаях он может быть особенно полезен, а в каких – неподходящим?

Ответ:

Преимущества:

- Простой в реализации и понимании.
- Не требует обучения (ленивый алгоритм).
- Может использоваться как для классификации, так и для регрессии.
- Адаптивен к новым данным.

Недостатки:

- Вычислительно затратен, особенно для больших наборов данных (необходим расчет расстояний до всех точек).
- Требуется большого объема памяти для хранения обучающих данных.
- Чувствителен к выбросам и масштабу признаков.
- Плохо работает с многомерными данными (проклятие размерности).

k-NN полезен, когда данные имеют сложную структуру и нет явной линейной зависимости. Он менее подходит для больших наборов данных и задач, требующих высокой скорости предсказания.

Вопрос 7:

Объясните, что такое энтропия и информационный выигрыш, и как они используются при построении дерева решений?

Ответ:

- Энтропия — это мера неопределенности или случайности в наборе данных. В контексте деревьев решений, она измеряет, насколько "неоднородным" является набор данных с точки зрения распределения классов.
- Информационный выигрыш — это мера уменьшения энтропии после разделения набора данных по определенному признаку. Алгоритм построения дерева решений выбирает признак, который обеспечивает наибольший информационный выигрыш, чтобы максимизировать "чистоту" получающихся дочерних узлов.

В процессе построения дерева решений выбирается тот признак, который максимально снижает энтропию (то есть предоставляет наибольший информационный выигрыш), для разделения узла дерева.

Вопрос 8:

Что такое переобучение в контексте деревьев решений? Какие методы можно использовать для борьбы с переобучением при построении дерева решений?

Ответ:

Переобучение возникает, когда дерево решений становится слишком сложным и начинает "запоминать" обучающие данные, включая шум и выбросы. В результате дерево хорошо работает на обучающих данных, но плохо обобщает новые, невидимые данные.

Методы борьбы с переобучением:

- Ограничение глубины дерева (`max_depth`): Не допускает слишком глубокое разрастание дерева.
- Минимальное количество объектов в узле (`min_samples_split`, `min_samples_leaf`): Задаёт минимальное количество объектов, необходимое для разделения узла или для формирования конечного листа.
- Отсечение ветвей (`pruning`): Удаление частей дерева, которые не улучшают обобщающую способность.
- Использование ансамблевых методов (Random Forest, Gradient Boosting): Объединение нескольких деревьев для уменьшения дисперсии и повышения устойчивости.

Вопрос 9:

Опишите основные принципы работы алгоритмов Random Forest и Gradient Boosting. В чем заключаются их различия и когда один из них предпочтительнее другого?

Ответ:

- Random Forest: Строит множество деревьев решений на случайных подмножествах данных и признаков. Каждое дерево обучается независимо, а затем прогнозы усредняются (для регрессии) или определяется класс большинством голосов (для классификации).
- Gradient Boosting: Строит деревья последовательно, каждое новое дерево пытается исправить ошибки предыдущих деревьев. Деревья взвешиваются и объединяются в сильную модель.

Различия: Random Forest обучается параллельно, менее подвержен переобучению, но может быть менее точным, чем Gradient Boosting. Gradient Boosting обучается последовательно, требует более тщательной настройки параметров, сильнее подвержен переобучению, но может достигать более высокой точности.

Предпочтения: Random Forest предпочтительнее, когда важна скорость обучения и устойчивость к переобучению. Gradient Boosting предпочтительнее, когда требуется максимальная точность и есть возможность тщательно настроить параметры.

Вопрос 10:

Что такое feature importance (важность признаков) в ансамблях деревьев решений? Как ее можно использовать для анализа данных и улучшения модели?

Ответ:

Feature importance показывает, насколько каждый признак вносит вклад в предсказания модели. Обычно она вычисляется на основе того, сколько раз признак использовался для разделения узлов дерева и насколько сильно это разделение уменьшало ошибку (например, энтропию или MSE).

Использование:

- Отбор признаков: Позволяет выявить наиболее важные признаки и исключить малозначимые, улучшая обобщающую способность модели и снижая вычислительную сложность.
- Понимание данных: Помогает понять, какие факторы наиболее сильно влияют на целевую переменную, что может быть полезно для анализа данных и принятия решений.

- Улучшение модели: Анализ feature importance может подсказать, какие признаки стоит преобразовать или добавить новые, чтобы повысить точность модели.

Вопрос 11:

Объясните, что такое опорные векторы, разделяющая гиперплоскость и ширина разделяющей полосы в контексте SVM. Как SVM находит оптимальную гиперплоскость?

Ответ:

- Опорные векторы – это подмножество точек обучающего набора данных, которые лежат ближе всего к разделяющей гиперплоскости. Они критичны для определения положения и ориентации этой гиперплоскости.
- Разделяющая гиперплоскость – это гиперплоскость, которая разделяет данные разных классов с максимальным запасом (margin).
- Ширина разделяющей полосы – это расстояние между двумя параллельными гиперплоскостями, которые проходят через опорные векторы каждого класса.

SVM находит оптимальную гиперплоскость путем максимизации ширины разделяющей полосы. Это означает, что алгоритм пытается найти гиперплоскость, которая максимально далеко отстоит от ближайших точек каждого класса (опорных векторов).

Вопрос 12:

Что такое kernel trick (ядро) в SVM? Какие существуют основные типы ядер и для чего они используются?

Ответ:

Kernel trick – это метод, который позволяет SVM решать нелинейные задачи классификации без явного преобразования входных данных в пространство более высокой размерности. Вместо этого ядро вычисляет скалярное произведение векторов в этом пространстве, что значительно упрощает вычисления.

Основные типы ядер:

- Линейное ядро (Linear): Используется для линейно разделимых данных.
- Полиномиальное ядро (Polynomial): Позволяет строить нелинейные границы классов определенной степени.
- Радиальное базисное ядро (RBF): Позволяет строить более сложные нелинейные границы классов. Хорошо подходит для большинства задач, но требует тщательной настройки параметра gamma.
- Сигмоидное ядро (Sigmoid): Иногда используется как замена нейронным сетям, но менее популярно, чем другие ядра.

Выбор ядра зависит от структуры данных и сложности задачи. RBF часто является хорошим выбором по умолчанию, но для больших наборов данных линейное ядро может быть быстрее и эффективнее.

Вопрос 13:

Объясните, в чем заключается предположение "наивности" в Наивном Байесовском классификаторе. Какие преимущества и недостатки это предположение создает?

Ответ:

"Наивность" в Наивном Байесовском классификаторе заключается в предположении о том, что все признаки (атрибуты), используемые для классификации, являются условно независимыми друг от друга при условии известной принадлежности к классу. Иными словами, значение одного признака не влияет на значение другого признака.

Преимущества:

- Упрощает вычисления вероятностей, делая алгоритм быстрым и масштабируемым.
- Требуется небольшого объема обучающих данных.
- Хорошо работает в задачах с высокой размерностью.

Недостатки:

- Предположение о независимости признаков часто не выполняется на практике, что может снизить точность классификации.
- Если какое-то значение признака не встречается в обучающем наборе, алгоритм будет давать нулевую вероятность для этого признака, что может привести к проблемам.

Вопрос 14:

Опишите основные типы Наивных Байесовских классификаторов (Gaussian, Multinomial, Complement, Bernoulli) и укажите, для каких типов данных они обычно используются.

Ответ:

- Gaussian Naive Bayes: Предполагает, что непрерывные признаки распределены нормально (Гаусса). Используется для данных с непрерывными значениями, например, для классификации изображений (после извлечения признаков).
- Multinomial Naive Bayes: Используется для дискретных данных, таких как частоты слов в тексте (count data). Хорошо подходит для классификации текстов, где признаки представляют собой количество вхождений слов.
- Complement Naive Bayes: Модификация Multinomial Naive Bayes, которая особенно хорошо работает при несбалансированных классах (когда один класс представлен значительно больше, чем другие). Используется для классификации текстов.
- Bernoulli Naive Bayes: Используется для бинарных данных или признаков (да/нет, 0/1). Подходит для классификации текстов, где признаки указывают на наличие или отсутствие определенного слова.

Вопрос 15:

Опишите, что такое прямое и обратное распространение (forward propagation и backpropagation) в нейронной сети. Какова роль каждого из этих процессов в обучении нейронной сети?

Ответ:

- Прямое распространение (Forward propagation) - это процесс передачи входных данных через нейронную сеть от входного слоя к выходному, последовательно обрабатывая данные в каждом слое с помощью функций активации. В результате получается предсказание сети.
- Обратное распространение (Backpropagation) - это процесс вычисления градиентов функции потерь по отношению к весам и смещениям нейронной сети. Этот процесс начинается с вычисления ошибки на выходном слое и затем "распространяется" обратно по сети слой за слоем.

Роль:

- Прямое распространение генерирует предсказание сети.
- Обратное распространение вычисляет, как нужно изменить веса и смещения, чтобы уменьшить ошибку и улучшить предсказания.

Вместе forward propagation и backpropagation образуют основной цикл обучения нейронной сети.

Вопрос 16:

Объясните, что такое функция активации в нейронной сети. Назовите несколько распространенных функций активации и опишите их основные особенности. Почему функция активации нужна?

Ответ:

Функция активации (activation function) — это функция, которая применяется к выходной сумме нейрона (взвешенной сумме входов плюс смещение) и определяет, будет ли нейрон "активирован" (выдаст сигнал) или нет. Она вносит нелинейность в модель, позволяя нейронной сети аппроксимировать сложные нелинейные функции.

Примеры функций активации:

- ReLU (Rectified Linear Unit): Выдает 0 для отрицательных значений и само значение для положительных ($f(x) = \max(0, x)$). Проста в вычислении, но может генерировать "мертвые" нейроны.
- Sigmoid: Выдает значения в диапазоне от 0 до 1 ($f(x) = 1 / (1 + \exp(-x))$). Использовалась раньше, но страдает от затухания градиента.
- Tanh (Hyperbolic Tangent): Выдает значения в диапазоне от -1 до 1 ($f(x) = \tanh(x)$). Похожа на Sigmoid, но имеет значения от -1 до 1.

Необходимость:

Без функций активации нейронная сеть представляла бы собой просто линейную регрессию, независимо от количества слоев. Функции активации позволяют моделировать сложные нелинейные зависимости в данных.

Промежуточная аттестация в 6 семестре проводится в форме экзамена. К экзамену допускаются студенты, выполнившие в полном объеме аудиторную нагрузку, самостоятельную работу. Студенты, имеющие задолженность, должны выполнить все обязательные виды деятельности.

Фонд оценочных средств для промежуточной аттестации включает:

- тестовые задания для экзамена.

Назначение оценочных средств: выявить сформированность компетенций ОПК-3, ОПК-5 (см. п. III).

Тестовое задание включает два варианта по 20 вопросов по всем темам курса. К тесту допускаются студенты, задавшие все домашние задания и получившие по каждому заданию зачет.

Критерий оценивания тестового задания

№	Тип задания	Критерии оценки	Результат оценивания
1	Задание закрытого типа на установление соответствия	Считается верным, если правильно установлены все соответствия (позиции одного столбца верно соотнесены с позициями другого столбца)	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов
2	Задание закрытого типа на установление последовательности	Считается верным, если правильно указана вся последовательность цифр	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов
3	Задание комбинированного типа с выбором одного верного ответа из четырех предложенных и обоснованием выбора	Считается верным, если правильно указана цифра (буква) правильного ответа и приведены корректные аргументы, используемые при выборе ответа	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов
4	Задание комбинированного типа с выбором нескольких верных ответов из четырех предложенных и обоснованием выбора	Считается верным, если правильно указаны цифры (буквы) правильного ответа и приведены корректные	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов

		аргументы, используемые при выборе ответа	
5	Задание открытого типа с развернутым ответом	Считается верным, если ответ совпадает с эталонным ответом по содержанию и полноте	Полное соответствие эталонному ответу – 1 балл Все остальные случаи – 0 баллов

Система получения баллов за тестирование

Оценка	критерий
отлично	18 и более баллов
хорошо	16 – 17 баллов
удовлетворительно	15 – 13 баллов
неудовлетворительно	12 баллов и менее

Оценочные материалы для промежуточной аттестации (экзамена)

Тестирование (Вариант 1).

Индекс и содержание формируемой компетенции	Индикаторы компетенций	Тестовые задания для промежуточной аттестации
ОПК-2 Способен использовать специализированные знания фундаментальных разделов математики, физики, химии и биологии для проведения исследований в области биоинженерии, биоинформатики и смежных дисциплин (модулей);	ИДК ОПК-2.1 Демонстрирует специализированные знания в области фундаментальных разделов математики, физики, химии, биологии и перспективы междисциплинарных исследований	<p>Задание комбинированного типа с выбором одного или нескольких верных ответов из четырех предложенных с аргументацией выбора</p> <p>Вопрос 1. Что такое машинное обучение? а) Изучение алгоритмов, которые не используют данные б) Написание статических правил для решения задач в) Метод, позволяющий системам обучаться на данных и делать прогнозы г) Система сбора информации в базу данных Ответ _____ Правильный ответ: в Аргументация: Машинное обучение — это область ИИ, в которой системы автоматически извлекают знания из данных и делают предсказания, без явного программирования логики.</p> <p>Вопрос 2. В чем основная задача линейной регрессии? а) Разделить данные на группы б) Предсказать категорию объекта в) Предсказать непрерывную переменную на основе признаков г) Сжать данные для хранения Ответ _____ Правильный ответ: в Аргументация: Линейная регрессия используется для количественного предсказания, например, температуры, стоимости, роста и пр., на основе линейной зависимости между признаками.</p> <p>Вопрос 3. Что характеризует линейную классификацию? а) Использование древовидной структуры б) Объединение нескольких моделей в) Разделение классов линейной границей г) Преобразование признаков в латентное пространство Ответ _____</p>
	ИДК ОПК-2.2 Умеет использовать навыки проведения исследований в области биоинженерии, биоинформатики с учетом специализированных фундаментальных знаний	
	ИДК ОПК-2.3 Владеет методами химии, физики и математического	

	моделирования для проведения исследований в области биоинженерии, биоинформатики	<p>Правильный ответ: в</p> <p>Аргументация: Линейные классификаторы (например, логистическая регрессия) используют прямые или гиперплоскости для разделения классов в признаковом пространстве.</p> <p>Вопрос 4.</p> <p>Как алгоритм k ближайших соседей делает предсказание?</p> <p>а) Использует деревья решений</p> <p>б) На основе гиперплоскости</p> <p>в) Среднее значение ближайших соседей</p> <p>г) Случайный выбор класса</p> <p>Ответ _____</p> <p>Правильный ответ: в</p> <p>Аргументация: Алгоритм ищет K объектов, ближайших по расстоянию, и принимает решение (среднее — в регрессии, голосование — в классификации) на их основе.</p>
<p>ОПК-3</p> <p>Способен проводить экспериментальную работу с организмами и клетками, использовать физико-химические методы исследования макромолекул, математические методы обработки результатов биологических исследований;</p>	<p>ИДК ОПК-3.1</p> <p>Проводит экспериментальную работу с организмами и клетками с использованием физико-химических методов исследования макромолекул</p>	<p>Вопрос 5.</p> <p>Что определяет узел в дереве решений?</p> <p>а) Число соседей</p> <p>б) Условие на значение признака</p> <p>в) Глубину модели</p> <p>г) Название класса</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: Каждый узел в дереве представляет собой логическое условие (например, "если возраст > 30"), определяющее путь по дереву.</p> <p>Вопрос 6.</p> <p>В чем преимущество ансамблей деревьев решений?</p> <p>а) Повышают точность и устойчивость модели</p> <p>б) Упрощают модель</p> <p>в) Уменьшают глубину дерева</p> <p>г) Увеличивают скорость предсказания</p> <p>Ответ _____</p> <p>Правильный ответ: а</p> <p>Аргументация: Ансамбли (Random Forest, Gradient Boosting) объединяют множество деревьев, снижая переобучение и повышая обобщающую способность.</p> <p>Вопрос 7.</p>
	<p>ИДК ОПК-3.2</p> <p>Демонстрирует практические навыки математических методов обработки результатов экспериментальных исследований</p>	
	<p>ИДК ОПК-3.3</p> <p>Владеет опытом применения методов для исследования макромолекул,</p>	

	<p>обработки результатов биологических исследований, прогнозирования перспектив и социальных последствий своей профессиональной деятельности.</p>	<p>Что ищет метод опорных векторов (SVM)?</p> <p>а) Сеть связей между точками б) Гиперплоскость, максимально разделяющую классы в) Центр масс каждого класса г) Максимальную плотность кластера</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: SVM создает разделяющую границу, максимизирующую отступ между классами, что помогает улучшить обобщение.</p> <p>Вопрос 8.</p> <p>На чем основан наивный Байесовский классификатор?</p> <p>а) Метод главных компонент б) Теорема Байеса и независимость признаков в) Расчет расстояний между точками г) Строительство деревьев</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: Классификатор применяет формулу Байеса при допущении, что признаки объекта независимы между собой — это и делает его "наивным".</p> <p>Вопрос 9.</p> <p>Что характерно для нейронных сетей?</p> <p>а) Построение плоской регрессии б) Иерархическая структура из узлов и ветвей в) Слои, автоматически извлекающие признаки г) Случайное разделение классов</p> <p>Правильный ответ: в</p> <p>Аргументация: Нейросети состоят из нескольких слоев, которые учатся представлять данные через более абстрактные признаки.</p> <p>Вопрос 10.</p> <p>Что отличает обучение без учителя?</p> <p>а) Использование размеченных данных б) Отсутствие целевых переменных в) Построение деревьев решений г) Регрессия как основа анализа</p> <p>Ответ _____</p>
--	---	--

		<p>Правильный ответ: б</p> <p>Аргументация: В обучении без учителя модели работают только с входными данными, не имея меток классов или выходных значений.</p> <p>Вопрос 11.</p> <p>Для чего проводится масштабирование данных?</p> <p>а) Повысить дисперсию</p> <p>б) Стандартизировать признаки</p> <p>в) Сделать данные линейными</p> <p>г) Переименовать переменные</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: Масштабирование делает признаки сопоставимыми по шкале, что особенно важно для алгоритмов, чувствительных к расстояниям.</p> <p>Вопрос 12.</p> <p>Зачем применяют метод главных компонент (PCA)?</p> <p>а) Построение дерева решений</p> <p>б) Повышение дисперсии данных</p> <p>в) Уменьшение размерности</p> <p>г) Увеличение числа признаков</p> <p>Ответ _____</p> <p>Правильный ответ: в</p> <p>Аргументация: PCA позволяет упростить данные, уменьшив число признаков с сохранением максимальной дисперсии, облегчая анализ.</p> <p>Вопрос 13.</p> <p>Что делает метод k-средних?</p> <p>а) Предсказывает значения</p> <p>б) Делит данные на k групп по близости</p> <p>в) Создает дерево</p> <p>г) Определяет вес признаков</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: Это метод кластеризации, который находит центры кластеров и распределяет объекты по ближайшим центрам.</p> <p>Задание закрытого типа на установление соответствия</p>
--	--	---

		<p>Вопрос 14. Соотнесите понятия и их описания в контексте машинного обучения:</p> <p>a) Регрессия b) Классификация c) Кластеризация d) PCA (метод главных компонент)</p> <p>Варианты:</p> <ol style="list-style-type: none"> 1. Деление объектов на группы без меток 2. Прогнозирование непрерывной переменной 3. Уменьшение размерности данных 4. Прогнозирование категориальной переменной <p>Ответ: _____ Правильный ответ: a — 2 (Регрессия — прогнозирует число) b — 4 (Классификация — определяет класс) c — 1 (Кластеризация — без учителя, делит на группы) d — 3 (PCA — уменьшает размерность)</p> <p>Вопрос 15. Соотнесите методы машинного обучения с их основными особенностями:</p> <p>a) Деревья решений b) Метод опорных векторов (SVM) c) Нейронные сети d) Наивный Байес</p> <p>Варианты:</p> <ol style="list-style-type: none"> 1. Использует теорему Байеса и предполагает независимость признаков 2. Использует многоуровневые слои и обучается на сложных признаках 3. Строит дерево с условиями на каждом узле 4. Строит разделяющую гиперплоскость <p>Ответ: _____ Правильный ответ: a — 3 (Деревья решений — структура "если-то") b — 4 (SVM — разделение гиперплоскостью) c — 2 (Нейросеть — слои, обучение признаков) d — 1 (Наивный Байес — простота и независимость)</p> <p>Задание закрытого типа на установление последовательности</p>
--	--	---

		<p>Вопрос 16.</p> <p>Расположите следующие шаги регрессионного анализа в машинном обучении в правильной последовательности:</p> <ol style="list-style-type: none"> 1. Оценка модели: Используйте метрики, такие как MSE, RMSE, R-squared, для оценки производительности модели на тестовом наборе данных. 2. Подготовка данных: Очистите, преобразуйте и разделите данные на обучающий и тестовый наборы. Обработайте отсутствующие значения и закодируйте категориальные признаки, если необходимо. 3. Выбор модели: Выберите подходящий алгоритм регрессии (например, линейная регрессия, полиномиальная регрессия, регрессия с использованием дерева решений, случайный лес, градиентный бустинг). 4. Сбор данных: Получите данные, необходимые для вашего регрессионного анализа. 5. Обучение модели: Обучите выбранную модель на обучающем наборе данных. 6. Визуализация и интерпретация результатов: Визуализируйте результаты, оцените коэффициенты (если применимо), и сделайте выводы на основе анализа. 7. Настройка гиперпараметров: Оптимизируйте гиперпараметры модели, используя методы, такие как перекрестная проверка или поиск по сетке (Grid Search), для повышения производительности. 8. Прогнозирование: Используйте обученную модель для прогнозирования значений целевой переменной на новых данных. <p>Правильный ответ: 4, 2, 3, 5, 7, 8, 1, 6</p> <p>Вопрос 17.</p> <p>Расположите следующие шаги линейной классификации в машинном обучении в правильной последовательности:</p> <ol style="list-style-type: none"> 1. Обучение модели: Обучите выбранную модель линейной классификации (например, логистическая регрессия, SVM с линейным ядром) на обучающем наборе данных, находя оптимальные параметры, разделяющие классы. 2. Подготовка данных: Очистите, преобразуйте и разделите данные на обучающий и тестовый наборы. Обработайте отсутствующие значения, закодируйте категориальные признаки и масштабируйте числовые признаки. 3. Сбор данных: Получите данные, необходимые для вашей задачи классификации. 4. Оценка модели: Используйте метрики, такие как точность, полнота, F1-мера, AUC-ROC, для оценки производительности модели на тестовом наборе данных. 5. Выбор модели: Выберите подходящий алгоритм линейной классификации (например, логистическая регрессия, линейная SVM, перцептрон). 6. Настройка гиперпараметров: Оптимизируйте гиперпараметры модели, используя методы, такие как перекрестная проверка или поиск по сетке (Grid Search), для повышения производительности и предотвращения переобучения. 7. Прогнозирование: Используйте обученную модель для прогнозирования классов для новых данных. 8. Анализ и интерпретация результатов: Визуализируйте результаты классификации (например, матрицу ошибок), оцените важность признаков и сделайте выводы на основе анализа.
--	--	---

		<p>Правильный ответ: 3, 2, 5, 1, 6, 7, 4, 8</p> <p>Задание открытого типа с развернутым ответом</p> <p>Вопрос 18.</p> <p>Принцип работы наивного Байесовского классификатора в машинном обучении?</p> <p>Ответ:</p> <p>Наивный байесовский классификатор - это простой и эффективный алгоритм машинного обучения, который использует теорему Байеса с "наивным" предположением о независимости признаков для решения задач классификации. Вот его основные принципы работы:</p> <p>1. Теорема Байеса:</p> <p>В основе алгоритма лежит теорема Байеса, которая определяет условную вероятность события на основе предварительных знаний об условиях, связанных с этим событием. Формула теоремы Байеса выглядит так:</p> $P(A B) = [P(B A) * P(A)] / P(B)$ <p>Где: * $P(A B)$ - апостериорная вероятность события A при условии, что произошло событие B (то, что мы хотим вычислить). * $P(B A)$ - вероятность события B при условии, что произошло событие A (правдоподобие). * $P(A)$ - априорная вероятность события A (вероятность события A до наблюдения события B). * $P(B)$ - априорная вероятность события B (вероятность наблюдения события B).</p> <p>В контексте классификации:</p> <ul style="list-style-type: none"> • A - это класс (например, "спам" или "не спам"). • B - это набор признаков (например, слова в электронном письме). <p>2. "Наивное" предположение о независимости:</p> <p>"Наивность" алгоритма заключается в предположении, что все признаки независимы друг от друга, при условии принадлежности к определенному классу. Это означает, что наличие одного признака никак не влияет на наличие другого признака. Хотя это предположение часто не соответствует действительности в реальных данных, оно значительно упрощает вычисления и позволяет алгоритму работать быстро и эффективно, особенно на больших наборах данных.</p> $P(B A) = P(\text{feature1} A) * P(\text{feature2} A) * \dots * P(\text{featureN} A)$ <p>Где: * feature1, feature2, ..., featureN - это отдельные признаки.</p> <p>3. Обучение (Estimation of Probabilities):</p> <p>Во время обучения наивный байесовский классификатор оценивает следующие вероятности на основе обучающих данных:</p> <ul style="list-style-type: none"> • $P(A)$: Априорные вероятности каждого класса. Это просто доля каждого класса в обучающем наборе данных. • $P(\text{feature}_i A)$: Условные вероятности каждого признака для каждого класса. Способ оценки этих вероятностей зависит от выбранного типа наивного Байеса (см. ниже). <p>4. Классификация (Prediction):</p> <p>Для классификации нового экземпляра данных алгоритм вычисляет апостериорную вероятность принадлежности к каждому классу, используя теорему Байеса и "наивное" предположение. Затем экземпляр классифицируется в</p>
--	--	---

		<p>класс с наивысшей апостериорной вероятностью. $\text{predicted_class} = \operatorname{argmax}_A P(A B)$</p> <p>Типы наивных байесовских классификаторов: Существуют различные типы наивных байесовских классификаторов, каждый из которых подходит для разных типов данных и предположений о распределении признаков:</p> <ul style="list-style-type: none"> • Gaussian Naive Bayes (Гауссовский наивный Байес): Предполагает, что числовые признаки имеют гауссовское (нормальное) распределение. Используется для непрерывных данных. • Multinomial Naive Bayes (Мультиномиальный наивный Байес): Предполагает, что признаки представляют собой частоты или счетчики (например, количество раз, когда слово появляется в документе). Используется для задач классификации текста и других задач, связанных со счетными данными. • Bernoulli Naive Bayes (Наивный Байес Бернулли): Предполагает, что признаки являются бинарными (например, присутствует или отсутствует слово в документе). Используется для задач классификации, где признаки представлены булевыми значениями. <p>Преимущества наивного байесовского классификатора:</p> <ul style="list-style-type: none"> • Простота реализации. • Быстрая скорость обучения и классификации. • Хорошо работает с высокоразмерными данными. • Эффективен даже при небольшом объеме обучающих данных. <p>Недостатки наивного байесовского классификатора:</p> <ul style="list-style-type: none"> • "Наивное" предположение о независимости признаков часто не выполняется в реальных данных. • Проблема нулевой частоты: если признак не встречается в обучающих данных для определенного класса, условная вероятность этого признака для этого класса будет равна нулю, что может привести к неправильным предсказаниям. Эта проблема решается с помощью сглаживания (например, добавление небольшого значения к каждому счетчику). <p>Вопрос 19. Принцип работы линейной классификация в машинном обучении? Ответ: Линейная классификация — это метод машинного обучения, который использует линейную функцию для разделения данных на различные классы. Основная идея заключается в том, чтобы найти оптимальную гиперплоскость (или линию в двумерном пространстве), которая наилучшим образом разделяет данные на основе их признаков. Вот основные принципы работы линейной классификации:</p> <p>1. Линейная модель: В линейной классификации используется линейная функция для предсказания принадлежности к классу. Эта функция имеет вид: $f(x) = w^T x + b$ Где: * x — вектор признаков объекта (данные, которые используются для предсказания). * w — вектор весов</p>
--	--	--

		<p>(параметры модели), который определяет ориентацию гиперплоскости. * b — смещение (bias), которое определяет положение гиперплоскости. * $w^T x$ — скалярное произведение векторов w и x.</p> <p>2. Гиперплоскость: Уравнение $w^T x + b = 0$ описывает гиперплоскость в пространстве признаков. Эта гиперплоскость разделяет пространство на две области:</p> <ul style="list-style-type: none"> • Объекты, для которых $w^T x + b > 0$, относятся к одному классу. • Объекты, для которых $w^T x + b < 0$, относятся к другому классу. <p>3. Обучение модели: Цель обучения линейной модели — найти оптимальные значения весов w и смещения b, которые наилучшим образом разделяют данные на классы. Для этого используются различные алгоритмы оптимизации, которые минимизируют функцию потерь (loss function). Функция потерь измеряет, насколько хорошо модель предсказывает классы для обучающих данных.</p> <p>4. Функция потерь: Существует несколько различных функций потерь, которые используются в линейной классификации. Некоторые из наиболее распространенных:</p> <ul style="list-style-type: none"> • Hinge Loss (для SVM): Используется в Support Vector Machines (SVM). Она штрафует за неправильную классификацию и за объекты, которые находятся слишком близко к разделяющей гиперплоскости. • Logistic Loss (Cross-Entropy Loss): Используется в логистической регрессии. Она измеряет разницу между предсказанными вероятностями и фактическими классами. • Perceptron Loss: Используется в алгоритме персептрона. Штрафует только за неправильную классификацию. <p>5. Алгоритмы оптимизации: После выбора функции потерь необходимо использовать алгоритм оптимизации, чтобы найти оптимальные значения w и b. Некоторые из наиболее распространенных алгоритмов:</p> <ul style="list-style-type: none"> • Gradient Descent (Градиентный спуск): Итеративно обновляет веса в направлении, противоположном градиенту функции потерь. • Stochastic Gradient Descent (SGD): Вариант градиентного спуска, который обновляет веса на основе одного случайного объекта из обучающего набора. Это делает его более быстрым, но менее стабильным, чем обычный градиентный спуск. • Newton's Method: Более сложный алгоритм, который использует вторую производную функции потерь для более точного определения направления обновления весов. <p>6. Классификация новых объектов: После обучения модели, для классификации нового объекта x, вычисляется значение линейной функции $f(x) = w^T x + b$. Знак этого значения определяет класс, к которому относится объект:</p> <ul style="list-style-type: none"> • Если $f(x) > 0$, то объект относится к классу 1 (или положительному классу). • Если $f(x) < 0$, то объект относится к классу 0 (или отрицательному классу). <p>7. Примеры линейных классификаторов:</p> <ul style="list-style-type: none"> • Логистическая регрессия: Использует логистическую функцию для предсказания вероятности
--	--	---

		<p>принадлежности к классу. Выдает вероятности, а не просто метки классов.</p> <ul style="list-style-type: none"> • Support Vector Machines (SVM): Стараются найти гиперплоскость, которая максимизирует зазор (margin) между классами. Эффективна в высокоразмерных пространствах. • Линейный дискриминантный анализ (LDA): Предполагает, что данные имеют гауссовское распределение и использует статистические свойства данных для разделения классов. • Персептрон: Простейший алгоритм линейной классификации, который итеративно корректирует веса на основе ошибок классификации. <p>Преимущества линейной классификации:</p> <ul style="list-style-type: none"> • Простота и интерпретируемость: Легко понять, как признаки влияют на предсказания. • Вычислительная эффективность: Быстро обучаются и предсказывают, что делает их подходящими для больших наборов данных. • Хорошо работают для линейно разделимых данных: Когда классы можно разделить прямой линией или гиперплоскостью. <p>Недостатки линейной классификации:</p> <ul style="list-style-type: none"> • Ограниченная выразительность: Не могут моделировать сложные нелинейные отношения между признаками. • Плохая производительность для нелинейно разделимых данных: Требуются более сложные модели (например, нейронные сети) для решения таких задач. • Чувствительность к выбросам: Выбросы могут сильно повлиять на положение разделяющей гиперплоскости. <p>Вопрос 20. Принцип работы анализа методом главных компонент в машинном обучении? Ответ: Анализ методом главных компонент (Principal Component Analysis, PCA) — это метод уменьшения размерности данных, который используется в машинном обучении, статистике и других областях. Его цель состоит в том, чтобы преобразовать исходные данные в новый набор некоррелированных переменных, называемых главными компонентами (principal components), которые отражают наибольшую дисперсию в данных. Вот ключевые принципы работы PCA:</p> <p>1. Цель PCA: Основная цель PCA - это:</p> <ul style="list-style-type: none"> • Уменьшение размерности: Сокращение количества признаков в наборе данных, сохраняя при этом как можно больше информации. • Выделение главных компонент: Определение наиболее важных направлений (главных компонент) в данных, вдоль которых наблюдается наибольшая дисперсия. • Удаление корреляции: Преобразование данных так, чтобы новые признаки (главные компоненты) были некоррелированы друг с другом. • Упрощение модели: Уменьшение сложности модели машинного обучения и предотвращение переобучения.
--	--	---

		<ul style="list-style-type: none"> Визуализация данных: Представление данных в меньшем количестве измерений (например, 2D или 3D) для визуализации. <p>2. Этапы работы PCA:</p> <p>PCA состоит из нескольких основных этапов:</p> <ul style="list-style-type: none"> 2.1 Стандартизация данных: <ul style="list-style-type: none"> * Перед применением PCA обычно необходимо стандартизировать данные, чтобы все признаки имели одинаковый масштаб. Это делается путем вычитания среднего значения и деления на стандартное отклонение для каждого признака. Это необходимо, чтобы признаки с большими значениями не доминировали при определении главных компонент. * Формула стандартизации: $x_standardized = (x - \text{mean}(x)) / \text{std}(x)$ 2.2 Вычисление ковариационной матрицы (или корреляционной матрицы): <ul style="list-style-type: none"> * Ковариационная матрица отражает взаимосвязь между всеми парами признаков. Каждый элемент (i, j) ковариационной матрицы показывает, насколько признаки i и j изменяются вместе. Если признаки сильно коррелированы, то ковариация будет большой. * Если данные были стандартизованы, можно использовать корреляционную матрицу, которая является ковариационной матрицей стандартизованных данных. * Формула для ковариационной матрицы: $\text{Cov}(X) = (1/(n-1)) * (X - \text{mean}(X))^T * (X - \text{mean}(X))$ 2.3 Вычисление собственных векторов и собственных значений: <ul style="list-style-type: none"> * Собственные векторы ковариационной матрицы представляют направления главных компонент. Собственные значения показывают, сколько дисперсии в данных объясняет каждая главная компонента. * Чем больше собственное значение, тем больше дисперсии объясняет соответствующий собственный вектор (главная компонента). * Решение уравнения: $\text{Cov}(X) * v = \lambda * v$, где v - собственный вектор, а λ - собственное значение. 2.4 Сортировка собственных значений и выбор главных компонент: <ul style="list-style-type: none"> * Собственные значения сортируются в порядке убывания. * Выбирается определенное количество главных компонент (собственных векторов), соответствующих наибольшим собственным значениям. Количество выбираемых компонент зависит от того, сколько дисперсии вы хотите сохранить. Часто используется кумулятивная дисперсия, чтобы определить, сколько компонент необходимо сохранить. * Например, можно выбрать первые k компонент, которые объясняют, скажем, 95% общей дисперсии. 2.5 Преобразование данных: <ul style="list-style-type: none"> * Исходные данные проецируются на выбранные главные компоненты, чтобы получить новый набор данных в уменьшенном количестве измерений. * Формула преобразования: $X_pca = X * W$, где W - матрица, столбцами которой являются выбранные собственные векторы. <p>3. Математическое обоснование:</p> <p>PCA стремится найти ортогональные (некоррелированные) линейные комбинации исходных признаков, которые</p>
--	--	--

		<p>максимизируют дисперсию. Это можно интерпретировать как поворот системы координат таким образом, чтобы наибольшая дисперсия данных приходилась на первую координату (первая главная компонента), вторая по величине дисперсия - на вторую координату (вторая главная компонента) и так далее.</p> <p>4. Кумулятивная объясненная дисперсия: Кумулятивная объясненная дисперсия (cumulative explained variance) — это метрика, которая показывает, какой процент общей дисперсии в данных объясняется первыми k главными компонентами. Она используется для выбора оптимального количества главных компонент.</p> <p>5. Примеры использования PCA:</p> <ul style="list-style-type: none"> • Уменьшение размерности изображений: PCA можно использовать для уменьшения количества признаков в изображениях, сохраняя при этом важную визуальную информацию. • Подавление шума: PCA может помочь отфильтровать шум из данных, поскольку шум часто вносит вклад в главные компоненты с наименьшей дисперсией. • Визуализация данных: PCA позволяет визуализировать многомерные данные в 2D или 3D пространстве. • Предобработка данных для машинного обучения: PCA может улучшить производительность моделей машинного обучения, уменьшая размерность данных и удаляя корреляцию между признаками. <p>Преимущества PCA:</p> <ul style="list-style-type: none"> • Уменьшение размерности: Сокращение количества признаков и упрощение моделей. • Удаление корреляции: Улучшение интерпретируемости данных и повышение стабильности моделей. • Визуализация данных: Представление данных в более наглядной форме. <p>Недостатки PCA:</p> <ul style="list-style-type: none"> • Линейность: PCA является линейным методом и не может улавливать нелинейные зависимости в данных. • Потеря информации: При уменьшении размерности всегда происходит некоторая потеря информации. • Интерпретация: Главные компоненты могут быть трудно интерпретировать, поскольку они являются линейными комбинациями исходных признаков. • Чувствительность к масштабу: Необходима стандартизация данных.
--	--	--

Тестирование (Вариант 2).

Индекс и содержание формируемой компетенции	Индикаторы компетенций	Тестовые задания для промежуточной аттестации
ОПК-2 Способен использовать	ИДК ОПК-2.1 Демонстрирует специализированные	Задание комбинированного типа с выбором одного или нескольких верных ответов из четырех предложенных и аргументацией выбора

специализированные знания фундаментальных разделов математики, физики, химии и биологии для проведения исследований в области биоинженерии, биоинформатики и смежных дисциплин (модулей);	знания в области фундаментальных разделов математики, физики, химии, биологии и перспективы междисциплинарных исследований	<p>Вопрос 1. Что изучает машинное обучение?</p> <p>а) Только регрессионные модели б) Алгоритмы ручного кодирования в) Алгоритмы, обучающиеся на данных г) Исключительно биологические данные</p> <p>Ответ _____</p> <p>Правильный ответ: в</p> <p>Аргументация: Суть машинного обучения — в создании моделей, которые обучаются закономерностям в данных без явного задания правил.</p>
	ИДК ОПК-2.2 Умеет использовать навыки проведения исследований в области биоинженерии, биоинформатики с учетом специализированных фундаментальных знаний	<p>Вопрос 2. Какой тип переменной предсказывает линейная регрессия?</p> <p>а) Категориальную б) Непрерывную в) Логическую г) Векторную</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: В линейной регрессии модель предсказывает количественное значение (например, цена, температура).</p>
	ИДК ОПК-2.3 Владеет методами химии, физики и математического моделирования для проведения исследований в области биоинженерии, биоинформатики	<p>Вопрос 3. Что делает линейная классификация?</p> <p>а) Ищет центры кластеров б) Объединяет модели в) Делит данные с помощью гиперплоскости г) Ищет признаки для РСА</p> <p>Ответ _____</p> <p>Правильный ответ: в</p> <p>Аргументация: Такие алгоритмы (логистическая регрессия, SVM) разделяют классы прямой линией (в 2D) или гиперплоскостью (в многомерности).</p>
ОПК-3 Способен проводить	ИДК ОПК-3.1 Проводит	<p>Вопрос 4. Особенность метода k ближайших соседей:</p> <p>а) Использует деревья решений б) Делает предсказания на основе ближайших точек</p>

экспериментальную работу с организмами и клетками, использовать физико-химические методы исследования макромолекул, математические методы обработки результатов биологических исследований;	экспериментальную работу с организмами и клетками с использованием физико-химических методов исследования макромолекул	<p>в) Требуется тренировочной выборки без классов</p> <p>г) Использует гиперплоскость</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: KNN анализирует соседей по расстоянию, определяя класс или значение объекта на основе похожих примеров.</p>
	ИДК ОПК-3.2 Демонстрирует практические навыки математических методов обработки результатов экспериментальных исследований	<p>Вопрос 5.</p> <p>Что представляют собой листья в дереве решений?</p> <p>а) Ветви алгоритма</p> <p>б) Веса признаков</p> <p>в) Предсказания</p> <p>г) Коэффициенты регрессии</p> <p>Ответ _____</p> <p>Правильный ответ: в</p> <p>Аргументация: В листьях дерева содержатся результаты — класс или значение, к которому приводит путь по дереву.</p>
	ИДК ОПК-3.3 Владеет опытом применения методов для исследования макромолекул, обработки результатов биологических исследований, прогнозирования перспектив и социальных последствий своей	<p>Вопрос 6.</p> <p>Что объединяет ансамбль деревьев решений?</p> <p>а) Одно большое дерево</p> <p>б) Кластер точек</p> <p>в) Несколько деревьев для повышения точности</p> <p>г) Гиперплоскости</p> <p>Ответ _____</p> <p>Правильный ответ: в</p> <p>Аргументация: Такие ансамбли (например, случайный лес) объединяют прогнозы многих деревьев для повышения надежности.</p> <p>Вопрос 7.</p> <p>Что делает SVM?</p> <p>а) Делит дерево пополам</p> <p>б) Находит максимально разделяющую плоскость</p> <p>в) Создает центроиды</p> <p>г) Предсказывает непрерывные переменные</p> <p>Ответ _____</p> <p>Правильный ответ: б</p> <p>Аргументация: Метод строит гиперплоскость, разделяющую классы с максимальным зазором (margin),</p>

	<p>профессиональной деятельности.</p>	<p>минимизируя риск ошибок.</p> <p>Вопрос 8. Почему наивный Байес называется "наивным"? а) Использует простые формулы б) Предполагает независимость признаков в) Не требует данных г) Обучается без учителя Ответ _____ Правильный ответ: б Аргументация: "Наивность" в том, что он считает признаки независимыми, что не всегда так, но даёт простоту и скорость.</p> <p>Вопрос 9. В чем сила нейронных сетей? а) В построении дендрограммы б) В плотности кластеров в) В способности извлекать сложные признаки г) В использовании линейных функций Ответ _____ Правильный ответ: в Аргументация: Нейросети учатся представлять сложные закономерности в данных, что делает их мощными для изображений, текста, речи.</p> <p>Вопрос 10. Что отличает обучение без учителя от обучения с учителем? а) Использование деревьев б) Отсутствие меток классов в) Анализ наивного Байеса г) Построение регрессий Ответ _____ Правильный ответ: б Аргументация: В отличие от обучения с учителем, здесь нет "правильного ответа", и алгоритм ищет структуру сам (например, кластеры).</p> <p>Вопрос 11. Масштабирование данных нужно для: а) Удаления пропусков</p>
--	---------------------------------------	---

		<p>б) Приведения всех признаков к одной шкале в) Увеличения размерности г) Построения графиков Ответ _____ Правильный ответ: б Аргументация: Разные шкалы могут исказить результаты, особенно в алгоритмах, использующих расстояния (например, KNN, SVM).</p> <p>Вопрос 12. Метод главных компонент (РСА) позволяет: а) Строить деревья решений б) Найти наиболее значимые направления в) Делить данные на классы г) Увеличить точность регрессии Ответ _____ Правильный ответ: б Аргументация: РСА уменьшает размерность, выделяя главные компоненты — оси с максимальной дисперсией.</p> <p>Вопрос 13. Что делает кластеризация k-средних? а) Разбивает данные на группы по признакам б) Создает гиперплоскости в) Удаляет шумы г) Делает предсказания категорий Ответ _____ Правильный ответ: а Аргументация: Метод ищет k центров кластеров, к которым по близости прикрепляются объекты — типичный алгоритм без учителя.</p> <p>Задание закрытого типа на установление соответствия</p> <p>Вопрос 14. Соотнесите алгоритмы с их применением: а) k-средних б) Линейная регрессия в) Логистическая регрессия г) К ближайших соседей Варианты: 1. Используется для задач классификации и определяет вероятность класса</p>
--	--	---

		<ol style="list-style-type: none"> 2. Делит данные на группы по расстоянию до центра 3. Используется для предсказания непрерывных значений 4. Делает предсказания на основе ближайших наблюдений <p>Ответ: _____</p> <p>Правильный ответ:</p> <p>a — 2 (K-средних — кластеризация)</p> <p>b — 3 (Линейная регрессия — прогноз чисел)</p> <p>c — 1 (Логистическая — вероятность принадлежности к классу)</p> <p>d — 4 (KNN — соседство)</p> <p>Вопрос 15.</p> <p>Соотнесите действия с этапами подготовки данных:</p> <p>a) Масштабирование</p> <p>b) Кодирование категориальных переменных</p> <p>c) Удаление выбросов</p> <p>d) Разделение выборки</p> <p>Варианты:</p> <ol style="list-style-type: none"> 1. Приведение данных к одной шкале 2. Преобразование текста в числовой формат 3. Удаление редких или слишком больших значений 4. Деление на обучающую и тестовую части <p>Ответ: _____</p> <p>Правильный ответ:</p> <p>a — 1 (Масштабирование — важный этап)</p> <p>b — 2 (Кодирование — для моделей)</p> <p>c — 3 (Выбросы могут мешать)</p> <p>d — 4 (Нужно для оценки)</p> <p>Задание закрытого типа на установление последовательности</p> <p>Вопрос 16.</p> <p>Расположите следующие шаги при построении и использовании модели дерева решений в машинном обучении в правильной последовательности:</p> <ol style="list-style-type: none"> 1. Построение дерева: Алгоритм рекурсивно разделяет данные на основе выбранных признаков и порогов, создавая древовидную структуру. 2. Выбор признаков для разделения: Определение наилучшего признака и порогового значения для разделения данных на каждом узле дерева (обычно с использованием метрик, таких как энтропия, информационный выигрыш или критерий Джини).
--	--	---

		<ol style="list-style-type: none"> 3. Сбор данных: Получите размеченные данные, необходимые для задачи классификации или регрессии. 4. Предварительная обработка данных: Очистка данных, обработка пропущенных значений, кодирование категориальных признаков (если необходимо). 5. Оценка модели: Оценка производительности дерева решений на тестовом наборе данных с использованием подходящих метрик (например, точность, полнота, F1-мера для классификации; MSE, RMSE, R-squared для регрессии). 6. Разделение данных: Разделите данные на обучающий и тестовый наборы. 7. Визуализация дерева (опционально): Визуализируйте структуру дерева решений для понимания принятых решений. 8. Уменьшение переобучения (прунинг): Применение методов прунинга (например, сокращение ветвей, ограничение глубины дерева, минимальное количество образцов в узле) для предотвращения переобучения и повышения обобщающей способности модели. 9. Прогнозирование: Использование обученного дерева решений для прогнозирования классов или значений для новых, невидимых данных. <p>Правильный ответ: 3, 4, 6, 2, 1, 8, 5, 7, 9</p> <p>Вопрос 17. Расположите следующие шаги при использовании наивного байесовского классификатора в машинном обучении в правильной последовательности:</p> <ol style="list-style-type: none"> 1. Предварительная обработка данных: Очистка данных, обработка пропущенных значений, масштабирование (при необходимости), кодирование категориальных признаков. Важно учитывать, что для различных типов наивного Байеса могут потребоваться разные виды обработки (например, гауссовский требует числовых данных). 2. Разделение данных: Разделите данные на обучающий и тестовый наборы. 3. Сбор данных: Получите размеченные данные, необходимые для задачи классификации. 4. Выбор модели наивного Байеса: Выберите подходящий тип наивного байесовского классификатора (например, гауссовский, мультиномиальный, Бернулли) в зависимости от типа данных и распределения признаков. 5. Оценка модели: Оценка производительности классификатора на тестовом наборе данных с использованием подходящих метрик (например, точность, полнота, F1-мера, AUC-ROC). 6. Обучение модели: Обучите выбранный наивный байесовский классификатор на обучающем наборе данных, оценивая вероятности признаков для каждого класса. 7. Настройка гиперпараметров (при необходимости): Оптимизируйте сглаживающие параметры (например, параметр α для мультиномиального наивного Байеса) для улучшения производительности и предотвращения переобучения. 8. Прогнозирование: Используйте обученный наивный байесовский классификатор для прогнозирования классов для новых, невидимых данных. <p>Правильный ответ: 3, 1, 2, 4, 6, 7, 8, 5</p>
--	--	---

		<p>Задание открытого типа с развернутым ответом</p> <p>Вопрос 18. Принцип работы линейной регрессии в машинном обучении? Ответ: Линейная регрессия — это один из самых простых и широко используемых алгоритмов машинного обучения для решения задач регрессии, то есть задач прогнозирования числовых значений. Она пытается установить линейную связь между независимыми переменными (признаками) и зависимой переменной (целевой переменной). Вот основные принципы работы линейной регрессии:</p> <p>1. Линейная модель: Основная идея линейной регрессии заключается в том, чтобы построить линейную модель, которая наилучшим образом описывает зависимость между входными признаками и целевой переменной. Эта модель представляется в виде уравнения: $y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_n \cdot x_n$ Где: * y — целевая переменная (значение, которое мы хотим предсказать). * x_1, x_2, \dots, x_n — независимые переменные (признаки). * b_0 — свободный член (intercept), который представляет собой значение y, когда все x равны нулю. * b_1, b_2, \dots, b_n — коэффициенты (coefficients), которые показывают, насколько изменяется y при изменении соответствующего x на единицу. В матричной форме уравнение можно записать так: $y = Xb$ Где: * y - вектор целевых переменных * X - матрица признаков (каждая строка представляет собой объект, каждый столбец - признак) * b - вектор коэффициентов, включая свободный член</p> <p>2. Обучение модели (нахождение оптимальных коэффициентов): Цель обучения линейной регрессии — найти оптимальные значения коэффициентов b_0, b_1, \dots, b_n, которые минимизируют разницу между предсказанными значениями (\hat{y}) и фактическими значениями (y) целевой переменной на обучающем наборе данных. Для этого используется функция потерь (loss function).</p> <p>3. Функция потерь: Наиболее часто используемая функция потерь в линейной регрессии — это среднеквадратичная ошибка (Mean Squared Error, MSE): $MSE = (1/n) \cdot \sum (y_i - \hat{y}_i)^2$ Где: * n — количество объектов в обучающем наборе данных. * y_i — фактическое значение целевой переменной для i-го объекта. * \hat{y}_i — предсказанное значение целевой переменной для i-го объекта. MSE измеряет среднее значение квадратов разностей между фактическими и предсказанными значениями. Чем меньше MSE, тем лучше модель соответствует данным.</p> <p>4. Методы оптимизации: Существуют различные методы оптимизации для минимизации функции потерь и нахождения оптимальных коэффициентов:</p>
--	--	--

		<ul style="list-style-type: none"> • Метод наименьших квадратов (Ordinary Least Squares, OLS): • * Это аналитический метод, который позволяет найти оптимальные коэффициенты напрямую, решая систему линейных уравнений. • * Формула для вычисления коэффициентов: $\hat{b} = (X^T * X)^{-1} * X^T * y$ • * OLS подходит для задач с небольшим количеством признаков и объектов, но может быть вычислительно дорогим для больших наборов данных. • Градиентный спуск (Gradient Descent): • * Это итеративный метод, который обновляет коэффициенты в направлении, противоположном градиенту функции потерь. • * Градиент показывает направление наибольшего возрастания функции потерь, поэтому, двигаясь в противоположном направлении, мы постепенно приближаемся к минимуму. • * Градиентный спуск может быть более эффективным, чем OLS, для больших наборов данных. • Стохастический градиентный спуск (Stochastic Gradient Descent, SGD): • * Это вариант градиентного спуска, который обновляет коэффициенты на основе одного случайно выбранного объекта из обучающего набора данных. • * SGD более быстрый, чем обычный градиентный спуск, но может быть менее стабильным. <p>5. Оценка качества модели:</p> <p>После обучения модели необходимо оценить ее качество. Для этого используются различные метрики, такие как:</p> <ul style="list-style-type: none"> • R-квадрат (R-squared): • * Показывает, какую долю дисперсии целевой переменной объясняет модель. • * Значение R-квадрат находится в диапазоне от 0 до 1. Чем ближе R-квадрат к 1, тем лучше модель соответствует данным. • Среднеквадратичная ошибка (MSE): (уже описана выше) • Средняя абсолютная ошибка (Mean Absolute Error, MAE): • * Измеряет среднее значение абсолютных разностей между фактическими и предсказанными значениями. • * $MAE = (1/n) * \sum y_i - \hat{y}_i$ <p>6. Предсказание:</p> <p>После обучения модели можно использовать ее для предсказания значений целевой переменной для новых объектов. Для этого необходимо подставить значения признаков нового объекта в уравнение линейной регрессии и вычислить значение \hat{y}.</p> <p>7. Assumptions of Linear Regression (Предположения линейной регрессии):</p> <p>Важно помнить, что линейная регрессия основана на определенных предположениях о данных:</p> <ul style="list-style-type: none"> • Линейность: Должна существовать линейная связь между признаками и целевой переменной. • Независимость ошибок: Ошибки (разности между фактическими и предсказанными значениями) должны быть независимы друг от друга. • Гомоскедастичность: Дисперсия ошибок должна быть постоянной для всех значений признаков. • Нормальность ошибок: Ошибки должны быть нормально распределены.
--	--	--

		<ul style="list-style-type: none"> Отсутствие мультиколлинеарности: Признаки не должны быть сильно коррелированы друг с другом. Нарушение этих предположений может привести к неточным или ненадежным результатам. <p>8. Расширения линейной регрессии:</p> <p>Существуют различные расширения линейной регрессии, которые позволяют справиться с более сложными задачами:</p> <ul style="list-style-type: none"> Полиномиальная регрессия: Использует полиномиальные признаки для моделирования нелинейных зависимостей. Регуляризованная линейная регрессия (Ridge, Lasso, Elastic Net): Добавляет штраф к функции потерь, чтобы предотвратить переобучение и уменьшить влияние мультиколлинеарности. <p>Преимущества линейной регрессии:</p> <ul style="list-style-type: none"> Простота и интерпретируемость: Легко понять, как признаки влияют на предсказания. Вычислительная эффективность: Быстро обучается и предсказывает. Широкая доступность: Реализована во многих библиотеках машинного обучения. <p>Недостатки линейной регрессии:</p> <ul style="list-style-type: none"> Ограниченная выразительность: Не может моделировать сложные нелинейные зависимости. Чувствительность к выбросам: Выбросы могут сильно повлиять на результаты. Предположения о данных: Требуется соблюдения определенных предположений о данных. <p>Вопрос 19.</p> <p>Принцип работы линейной регрессии в машинном обучении?</p> <p>Ответ:</p> <p>Кластеризация k-средних (k-means clustering) — это один из самых популярных и простых алгоритмов кластеризации в машинном обучении. Его цель — разделить набор данных на k отдельных кластеров, где каждый объект принадлежит к кластеру с ближайшим средним значением (центроидом). Вот основные принципы работы алгоритма k-средних:</p> <p>1. Цель кластеризации k-средних:</p> <p>Основная цель k-средних — минимизировать суммарное внутрикластерное расстояние (within-cluster sum of squares, WCSS), которое измеряет, насколько объекты внутри каждого кластера близки друг к другу. Другими словами, алгоритм стремится найти такие кластеры, чтобы объекты внутри каждого кластера были максимально похожи, а объекты из разных кластеров — максимально различны.</p> <p>2. Шаги алгоритма k-средних:</p> <p>Алгоритм k-средних состоит из нескольких основных шагов:</p> <ul style="list-style-type: none"> 2.1 Инициализация: <ul style="list-style-type: none"> Определите число кластеров k, которое вы хотите создать. Выберите k начальных центроидов. Существует несколько способов выбрать начальные центроиды: <ul style="list-style-type: none"> Случайный выбор: Случайно выбрать k объектов из набора данных в качестве центроидов. k-means++: Это более продвинутый метод инициализации, который выбирает центроиды таким
--	--	---

		<p>образом, чтобы они были максимально удалены друг от друга. Это помогает избежать ситуации, когда начальные центроиды находятся близко друг к другу, что может привести к плохой кластеризации.</p> <ul style="list-style-type: none"> 2.2 Присвоение кластеров: <ul style="list-style-type: none"> Для каждого объекта в наборе данных вычислите расстояние до каждого из k центроидов. Обычно используется евклидово расстояние. Присвойте каждый объект кластеру с ближайшим центроидом. 2.3 Обновление центроидов: <ul style="list-style-type: none"> Для каждого кластера вычислите новое положение центроида. Новый центроид — это среднее значение всех объектов, принадлежащих этому кластеру. 2.4 Повторение шагов 2 и 3: <ul style="list-style-type: none"> Повторяйте шаги 2 и 3 до тех пор, пока центроиды не перестанут изменяться (или пока не будет достигнуто максимальное количество итераций). Когда центроиды перестают изменяться, это означает, что алгоритм достиг сходимости. <p>3. Функция стоимости (Cost function): Алгоритм k-средних стремится минимизировать функцию стоимости, которая измеряет суммарное внутрикластерное расстояние (WCSS): $WCSS = \sum \sum \ x_i - \mu_j\ ^2$ Где: x_i — объект, принадлежащий кластеру j. μ_j — центроид кластера j. * Суммирование происходит по всем кластерам и всем объектам внутри каждого кластера.</p> <p>4. Выбор оптимального количества кластеров (k): Выбор оптимального количества кластеров k является важной задачей. Существует несколько методов для выбора k:</p> <ul style="list-style-type: none"> Метод локтя (Elbow method): <ul style="list-style-type: none"> Запустите алгоритм k-средних для различных значений k и вычислите WCSS для каждого значения. Постройте график зависимости WCSS от k. Выберите значение k, при котором график WCSS начинает "сгибаться" или замедляться в уменьшении. Этот "изгиб" часто называют "локтем". Метод силуэта (Silhouette method): <ul style="list-style-type: none"> Для каждого объекта вычислите коэффициент силуэта, который измеряет, насколько объект похож на объекты в своем кластере и насколько он отличается от объектов в других кластерах. Вычислите средний коэффициент силуэта для всех объектов. Выберите значение k, которое максимизирует средний коэффициент силуэта. <p>5. Преимущества k-средних:</p> <ul style="list-style-type: none"> Простота реализации: Алгоритм легко понять и реализовать. Вычислительная эффективность: Алгоритм относительно быстрый, особенно для небольших наборов данных.
--	--	--

		<ul style="list-style-type: none"> Широкая применимость: Алгоритм может быть использован для кластеризации данных в различных областях. <p>6. Недостатки k-средних:</p> <ul style="list-style-type: none"> Чувствительность к начальной инициализации: Результаты кластеризации могут зависеть от выбора начальных центроидов. Необходимость определения количества кластеров (k): Требуется предварительно определить количество кластеров, что может быть сложной задачей. Предположение о сферической форме кластеров: Алгоритм лучше всего работает, когда кластеры имеют сферическую форму и примерно одинаковый размер. Чувствительность к выбросам: Выбросы могут сильно повлиять на положение центроидов и результаты кластеризации. Не работает с невыпуклыми кластерами: Плохо справляется с данными, где кластеры имеют сложную, невыпуклую форму. <p>7. Когда использовать k-средних:</p> <ul style="list-style-type: none"> Когда вы хотите разделить данные на несколько отдельных групп. Когда вы не знаете заранее количество кластеров. Когда кластеры имеют примерно сферическую форму и одинаковый размер. Когда важна вычислительная эффективность. <p>Вопрос 20. Принцип работы нейронных сетей в машинном обучении? Ответ: Нейронные сети (Neural Networks, NN) — это мощный и гибкий инструмент машинного обучения. Они используются для решения широкого круга задач, включая классификацию, регрессию, распознавание образов, обработку естественного языка и многие другие. Вот основные принципы работы нейронных сетей:</p> <p>1. Структура нейронной сети: Нейронная сеть состоит из взаимосвязанных узлов, называемых нейронами или персептронами, которые организованы в слои. Существует три основных типа слоев:</p> <ul style="list-style-type: none"> Входной слой (Input Layer): Получает входные данные (признаки) из внешнего мира. Количество нейронов во входном слое соответствует количеству признаков в наборе данных. Скрытые слои (Hidden Layers): Выполняют сложные вычисления над входными данными. Нейронная сеть может иметь один или несколько скрытых слоев. Глубина сети (количество скрытых слоев) определяет ее способность моделировать сложные зависимости в данных. Выходной слой (Output Layer): Выдает результат работы сети. Количество нейронов в выходном слое зависит от типа задачи. Например, для задачи бинарной классификации (два класса) выходной слой может содержать один нейрон, а для задачи многоклассовой классификации (несколько классов) — несколько нейронов.
--	--	--

		<p>2. Работа нейрона (Персептрона):</p> <p>Каждый нейрон выполняет следующие операции:</p> <ul style="list-style-type: none"> • Получение входных данных: Нейрон получает входные данные от нейронов предыдущего слоя. Каждый вход умножается на соответствующий вес (weight), который определяет важность этого входа. • Суммирование взвешенных входов: Нейрон суммирует все взвешенные входы и добавляет смещение (bias). Смещение позволяет нейрону активироваться, даже если все входы равны нулю. • Функция активации (Activation Function): Полученная сумма передается в функцию активации, которая преобразует ее в выходное значение нейрона. Функции активации вводят нелинейность в модель, что позволяет нейронной сети моделировать сложные зависимости в данных. Примеры функций активации: <ul style="list-style-type: none"> ○ Sigmoid: Выдает значение в диапазоне от 0 до 1. Используется для задач бинарной классификации. ○ ReLU (Rectified Linear Unit): Выдает 0 для отрицательных входных значений и входное значение для положительных входных значений. Широко используется в современных нейронных сетях. ○ Tanh (Hyperbolic Tangent): Выдает значение в диапазоне от -1 до 1. • Выдача выходного значения: Выходное значение нейрона передается в нейроны следующего слоя. <p>3. Процесс обучения нейронной сети:</p> <p>Обучение нейронной сети — это процесс настройки весов и смещений, чтобы сеть могла правильно отображать входные данные на выходные. Этот процесс включает в себя два основных этапа:</p> <ul style="list-style-type: none"> • Прямое распространение (Forward Propagation): <ul style="list-style-type: none"> ○ Входные данные передаются через сеть, слой за слоем, пока не будет получен выходной результат. ○ На каждом слое каждый нейрон выполняет вычисления, описанные выше (взвешивание входов, суммирование, применение функции активации). • Обратное распространение (Backpropagation): <ul style="list-style-type: none"> ○ Выходной результат сравнивается с фактическим значением целевой переменной, и вычисляется функция потерь (loss function). Функция потерь измеряет разницу между предсказанным и фактическим значением. ○ Затем, используя алгоритм градиентного спуска (gradient descent), вычисляются градиенты функции потерь по отношению ко всем весам и смещениям в сети. Градиент показывает направление наибольшего возрастания функции потерь. ○ Веса и смещения обновляются в направлении, противоположном градиенту, чтобы минимизировать функцию потерь. Этот процесс повторяется много раз (эпох), пока сеть не научится правильно отображать входные данные на выходные. <p>4. Функция потерь (Loss Function):</p> <p>Функция потерь измеряет разницу между предсказанными и фактическими значениями. Выбор функции потерь зависит от типа задачи:</p> <ul style="list-style-type: none"> • Mean Squared Error (MSE): Используется для задач регрессии. • Binary Cross-Entropy: Используется для задач бинарной классификации. • Categorical Cross-Entropy: Используется для задач многоклассовой классификации.
--	--	---

		<p>5. Градиентный спуск (Gradient Descent):</p> <p>Градиентный спуск — это итеративный алгоритм оптимизации, который используется для минимизации функции потерь. Он обновляет веса и смещения в направлении, противоположном градиенту функции потерь. Существуют различные варианты градиентного спуска, такие как:</p> <ul style="list-style-type: none"> • Batch Gradient Descent: Вычисляет градиент на основе всех объектов в обучающем наборе данных. • Stochastic Gradient Descent (SGD): Вычисляет градиент на основе одного случайно выбранного объекта в обучающем наборе данных. • Mini-Batch Gradient Descent: Вычисляет градиент на основе небольшого подмножества объектов (мини-батча) в обучающем наборе данных. <p>6. Гиперпараметры (Hyperparameters):</p> <p>Гиперпараметры — это параметры, которые не изучаются в процессе обучения, а устанавливаются заранее. Примеры гиперпараметров:</p> <ul style="list-style-type: none"> • Learning Rate (Скорость обучения): Определяет размер шага при обновлении весов и смещений. • Number of Layers (Количество слоев): Определяет глубину сети. • Number of Neurons per Layer (Количество нейронов в слое): Определяет ширину слоя. • Activation Function (Функция активации): Определяет нелинейность, вводимую в сеть. • Batch Size (Размер батча): Определяет количество объектов, используемых для вычисления градиента на каждой итерации. • Epochs (Эпохи): Определяет количество проходов по всему обучающему набору данных. <p>Выбор оптимальных гиперпараметров является важной задачей, которая может существенно повлиять на производительность нейронной сети. Часто используются методы, такие как Grid Search и Random Search для поиска оптимальных значений.</p> <p>7. Типы нейронных сетей:</p> <p>Существуют различные типы нейронных сетей, предназначенные для решения различных задач:</p> <ul style="list-style-type: none"> • Multilayer Perceptron (MLP): Простая нейронная сеть с одним или несколькими скрытыми слоями. Подходит для задач классификации и регрессии, где данные имеют табличную структуру. • Convolutional Neural Network (CNN): Используется для обработки изображений и видео. CNN используют слои свертки (convolutional layers) для извлечения признаков из изображений. • Recurrent Neural Network (RNN): Используется для обработки последовательностей данных, таких как текст и временные ряды. RNN имеют рекуррентные связи, которые позволяют им запоминать информацию о предыдущих элементах последовательности. Примеры RNN: LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit). • Transformers: Используются для обработки последовательностей данных, особенно в задачах обработки естественного языка. Они основаны на механизме внимания (attention mechanism), который позволяет им фокусироваться на наиболее важных частях входной последовательности. <p>Преимущества нейронных сетей:</p> <ul style="list-style-type: none"> • Высокая точность: Могут достигать очень высокой точности в различных задачах.
--	--	---

		<ul style="list-style-type: none"> • Автоматическое извлечение признаков: Могут автоматически извлекать признаки из данных, что избавляет от необходимости ручного проектирования признаков. • Способность моделировать сложные зависимости: Могут моделировать сложные нелинейные зависимости в данных. • Универсальность: Могут быть использованы для решения широкого круга задач. <p>Недостатки нейронных сетей:</p> <ul style="list-style-type: none"> • Требуют больших объемов данных: Для обучения нейронных сетей требуется большое количество данных. • Вычислительно дорогие: Обучение нейронных сетей может быть вычислительно дорогим и требовать мощного оборудования. • Сложность интерпретации: Сложно понять, как нейронная сеть принимает решения (проблема "черного ящика"). • Склонность к переобучению: Нейронные сети могут легко переобучиться на обучающем наборе данных, что приводит к плохой обобщающей способности.
--	--	--

Разработчик:

Букин доцент Букин Ю.С.
(подпись)

Программа составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 06.05.01 «Биоинженерия и биоинформатика»

Программа рассмотрена на заседании кафедры физико-химической биологии, биоинженерии и биоинформатики 17.04.2024 г. протокол № 15.

Зав. кафедрой, д.б.н., профессор В.П. Саловарова Саловарова

Настоящая программа, не может быть воспроизведена ни в какой форме без предварительного письменного разрешения кафедры-разработчика программы.