

1 Цели, задачи и планируемые результаты обучения по дисциплине

Цели дисциплины «Машинное обучение и анализ данных»: ознакомление студентов с современными подходами к анализу и управлению рисками информационной безопасности на основе методов машинного обучения; формирование у студентов системного представления о математическом аппарате, применяемом в машинном обучении; формирование у студентов практических навыков проектирования и разработки моделей машинного обучения; формирование у студентов уверенных навыков выработки решений профессиональных задач на основе методов машинного обучения и оценки их качества.

Задачами дисциплины являются: ознакомление студентов с основами построения моделей машинного обучения; подготовка студентов к разработке программных комплексов для решения прикладных задач на основе методов машинного обучения.

В результате изучения дисциплины студенты должны знать: базовые понятия и методы машинного обучения; принципы работы со специализированными программными продуктами и библиотеками для анализа данных методами машинного обучения.

В результате изучения дисциплины студенты должны уметь: комплексно применять основы математики и информатики для решения задач машинного обучения; разрабатывать прикладное программное обеспечение для решения задач машинного обучения.

В результате изучения дисциплины студенты должны владеть: базовыми навыками проектирования, разработки и поддержания интеллектуальных систем защиты информации.

Обучение по дисциплине «Машинное обучение и анализ данных» направлено на формирование у обучающихся следующих общеобразовательных и профессиональных компетенций:

Код и наименование компетенций	Индикаторы достижения компетенции
ОПК-8. Способен применять методы научных исследований при проведении разработок в области защиты информации в автоматизированных системах.	ИОПК-8.1. Знает методы научных исследований при проведении разработок в области защиты информации в автоматизированных системах. ИОПК-8.2. Умеет применять методы научных исследований при проведении разработок в области защиты информации в автоматизированных системах. ОПК-8.3 Владеет методами научных исследований при проведении разработок в области защиты информации в автоматизированных системах.

2 Место дисциплины в структуре образовательной программы

Дисциплина «Машинное обучение и анализ данных» относится к базовой части блока Б1 «Дисциплины (модули)».

Базируется на знаниях, навыках и умениях, полученных студентами в результате освоения дисциплин: «Языки программирования», «Технологии и методы программирования», «Линейная алгебра и функции нескольких переменных», «Математический анализ», «Теория вероятностей», «Математическая статистика», «Теория информации».

Приобретенные в ходе изучения дисциплины «Машинное обучение и анализ данных» знания, навыки и умения в большей мере требуются для освоения дисциплин: «Защита информации в АСУ ТП», «Управление информационной безопасностью», «Проектная деятельность».

3 Структура и содержание дисциплины

Общая трудоемкость дисциплины «Машинное обучение и анализ данных» составляет 4 зачетные единицы (144 часа).

3.1 Виды учебной работы и трудоемкость (по формам обучения)

3.1.1 Очная форма обучения

№ п/п	Вид учебной работы	Количество часов	Семестры
			8
1	Аудиторные занятия	72	72
	В том числе:		
1.1	Лекции	0	0
1.2	Семинарские/практические занятия	0	0
1.3	Лабораторные занятия	72	72
2	Самостоятельная работа	72	72
3	Промежуточная аттестация		
	Экзамен	–	Экзамен
	Итого:	144	144

3.1.2 Очно-заочная форма обучения

Не предусмотрена программой.

3.1.3 Заочная форма обучения

Не предусмотрена программой.

3.2 Тематический план изучения дисциплины (по формам обучения)

3.2.1 Очная форма обучения

№ п/п	Разделы/темы дисциплины	Трудоемкость, час					
		Всего	Аудиторная работа				Самос тояте льная работ а
			Лек ции	Семинар ские/ практиче ские занятия	Лабор аторн ые заняти я	Практи ческая подгот овка	
1	Тема 1. Методы машинного обучения для решения задач информационной безопасности	0	0	0	4	0	4
2	Тема 2. Предварительное изучение одномерных данных	0	0	0	4	0	4
3	Тема 3. Предварительная обработка многомерных данных	0	0	0	4	0	4

4	Тема 4. Классификация, метод k ближайших соседей	0	0	0	4	0	4
5	Тема 5. Классификация, метод дерева решений	0	0	0	4	0	4
6	Тема 6. Кросс-валидация модели классификации	0	0	0	4	0	4
7	Тема 7. Классификация, наивный метод Баеса	0	0	0	4	0	4
8	Тема 8. Классификация, метод опорных векторов	0	0	0	4	0	4
9	Тема 9. Классификация, многослойный персептрон (нейронная сеть прямого распространения)	0	0	0	4	0	4
10	Тема 10. Метод главных компонент	0	0	0	8	0	8
11	Тема 11. Кластеризация, метод k средних	0	0	0	4	0	4
12	Тема 12. Подбор оптимального числа кластеров, методы локтя и силуэта	0	0	0	4	0	4
13	Тема 13. Кластеризация, метод DBSCAN	0	0	0	4	0	4
14	Тема 14. Кластеризация, метод агломеративной кластеризации	0	0	0	4	0	4
15	Тема 15. Кластеризация, сеть Кохонена	0	0	0	4	0	4
16	Тема 16. Кластеризация, карта Кохонена	0	0	0	4	0	4
17	Тема 17. Удаление выбросов и выявление аномалий в многомерных данных	0	0	0	4	0	4
Итого		144	0	0	72	0	72

3.2.2 Очно-заочная форма обучения

Не предусмотрена программой.

3.2.3 Заочная форма обучения

Не предусмотрена программой.

3.3 Содержание дисциплины

Тема 1. Методы машинного обучения для решения задач информационной безопасности

Предмет машинного обучения, место машинного обучения в искусственном интеллекте. Понятие реляционной модели данных. Понятие векторного пространства. Метрика, аксиомы расстояния. Понятие больших данных. Обзор задач информационной безопасности, решаемых методами машинного обучения.

Тема 2. Предварительное изучение одномерных данных

Переменная (и случайная) величина. Генеральная совокупность, выборка. Математическое ожидание, дисперсия, среднее квадратическое отклонение. Правило «трех

сигм», доверительный интервал (неравенство Чебышева). Гистограмма (и плотность вероятности). Нормальное распределение (кривая Гаусса). Мода, сегментация данных.

Тема 3. Предварительная обработка многомерных данных

Количественные и качественные признаки. Нормализация количественных признаков. Преобразование качественных признаков. Удаление выбросов. Исключение неинформативных признаков (на основе диаграмм частот и гистограмм). Прореживание больших данных. Выравнивание объемов классов.

Тема 4. Классификация, метод k ближайших соседей

Машинное обучение с учителем. Расстояние Евклида. Алгоритм построения и порядок применения модели классификации. Разбор учебных примеров.

Тема 5. Классификация, метод дерева решений

Энтропия. Алгоритм построения и порядок применения модели классификации. Разбор учебных примеров.

Тема 6. Кросс-валидация модели классификации

Метод Монте-Карло. Алгоритмы и применение кросс-валидации. Разбор учебных примеров.

Тема 7. Классификация, наивный метод Баеса

Условная вероятность, теорема Баеса. Алгоритм построения и порядок применения модели классификации. Разбор учебных примеров.

Тема 8. Классификация, метод опорных векторов

Скалярное произведение векторов. Уравнение гиперплоскости. Алгоритмы построения и порядок применения моделей классификации (бинарной и многоклассовой). Разбор учебных примеров.

Тема 9. Классификация, многослойный персептрон (нейронная сеть прямого распространения)

Производная и градиент функции в точке. Искусственный нейрон. Функция активации (ее назначение). Полновесная нейронная сеть. Дельта-правило (метод градиентного спуска). Алгоритм построения и порядок применения модели классификации. Разбор учебных примеров.

Тема 10. Метод главных компонент

Базис векторного пространства. Корреляционная матрица. Собственный вектор. Алгоритм метода и варианты его использования. Тепловая карта. Разбор учебных примеров.

Тема 11. Кластеризация, метод k средних

Машинное обучение без учителя. Центр масс. Алгоритм построения и порядок применения модели кластеризации. Разбор учебных примеров.

Тема 12. Подбор оптимального числа кластеров, методы локтя и силуэта

Метод наименьших квадратов. Алгоритмы методов подбора числа кластеров и их применение. Разбор учебных примеров.

Тема 13. Кластеризация, метод DBSCAN

Алгоритм построения и порядок применения модели кластеризации. Разбор учебных примеров.

Тема 14. Кластеризация, метод агломеративной кластеризации

Дендрограмма. Алгоритм построения и порядок применения модели кластеризации. Разбор учебных примеров.

Тема 15. Кластеризация, сеть Кохонена

Алгоритм построения и порядок применения модели кластеризации. Разбор учебных примеров.

Тема 16. Кластеризация, карта Кохонена

Алгоритм построения и порядок применения модели кластеризации. Разбор учебных примеров.

Тема 17. Удаление выбросов и выявление аномалий в многомерных данных

Обзор основных приемов.

3.4 Тематика семинарских/практических и лабораторных занятий

3.4.1 Семинарские/практические занятия

Не предусмотрены программой.

3.4.2 Лабораторные занятия

Тема работы 1. Сегментация одномерных данных

Цель работы: изучение теоретических основ исследования одномерных данных на однородность; выработка практических навыков написания программ с применением сегментации однородных данных. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 2. Предобработка многомерных данных

Цель работы: изучение теоретических основ предобработки многомерных данных для решения задач машинного обучения; выработка практических навыков программной подготовки обучающих данных для алгоритмов машинного обучения. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 3. Классификация данных методом k ближайших соседей

Цель работы: изучение теоретических основ применения метода k ближайших соседей для классификации многомерных данных; выработка практических навыков программной разработки классификаторов на основе метода k ближайших соседей. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 4. Классификация данных методом дерева решений

Цель работы: изучение теоретических основ применения метода дерева решений для классификации многомерных данных; выработка практических навыков программной разработки классификаторов на основе метода дерева решений. Содержание работы:

решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 5. Классификация данных наивным методом Байеса

Цель работы: изучение теоретических основ применения наивного метода Байеса для классификации многомерных данных; выработка практических навыков программной разработки классификаторов на основе наивного метода Байеса. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 6. Классификация данных методом опорных векторов

Цель работы: изучение теоретических основ применения метода опорных векторов для классификации многомерных данных; выработка практических навыков программной разработки классификаторов на основе метода опорных векторов. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 7. Кросс-валидация модели классификации

Цель работы: изучение теоретических основ применения кросс-валидации для оценки качества модели классификации; выработка практических навыков программной разработки процедуры применения кросс-валидации для оценки качества модели классификации; изучение теоретических основ применения метода главных компонент для улучшения качества модели классификации; выработка практических навыков программной разработки процедуры применения метода главных компонент для улучшения качества модели классификации. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 8. Понижение размерности признакового пространства

Цель работы: изучение теоретических основ применения метода главных компонент для понижения размерности признакового пространства; выработка практических навыков программной разработки процедуры применения метода главных компонент для понижения размерности признакового пространства. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 9. Кластеризация данных методом k средних

Цель работы: изучение теоретических основ применения метода k средних для кластеризации многомерных данных; выработка практических навыков программной разработки процедуры кластеризации многомерных данных на основе метода k средних. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию

программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

Тема работы 10. Кластеризация документов по содержанию

Цель работы: изучение теоретических основ применения метода главных компонент для кластеризации текстовых документов по их содержанию; выработка практических навыков программной разработки процедуры применения метода главных компонент для кластеризации текстовых документов по их содержанию. Содержание работы: решение разноуровневых задач по теме работы. Пример задания представлен в разделе 7.3. Результат работы студента: разработанные согласно заданию программы. Подготовка студента к работе, выполнение студентом работы и защита студентом работы направлены на формирование у студента компетенции ОПК-8.

3.5 Тематика курсовых проектов (курсовых работ)

Не предусмотрены программой.

4 Учебно-методическое и информационное обеспечение

4.1 Нормативные документы и ГОСТы

1. Федеральный закон от 29 декабря 2012 года № 273-ФЗ «Об образовании в Российской Федерации» (с изменениями и дополнениями).

2. Федеральный государственный образовательный стандарт высшего образования – специалитет по специальности 10.05.03 «Информационная безопасность автоматизированных систем», утвержденный приказом Министерства науки и высшего образования Российской Федерации от 26.11.2020 № 1457.

3. Профессиональный стандарт 06.032 «Специалист по безопасности компьютерных систем и сетей», утвержденный приказом Минтруда России от 01 ноября 2016 г. №598н.

4. Профессиональный стандарт 06.033 «Специалист по защите информации в автоматизированных системах», утвержденный приказом Минтруда России от 15 сентября 2016 г. № 522н.

5. Приказ Министерства образования и науки РФ от 05 апреля 2017 г. № 301 «Об утверждении Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры».

6. «Порядок проведения государственной итоговой аттестации по образовательным программам высшего образования – программам бакалавриата, программам специалитета и программам магистратуры», утвержденный приказом Минобрнауки России от 29 июня 2015 г. № 636.

7. Положение «О практической подготовке обучающихся», утвержденное приказом Министерства науки и высшего образования Российской Федерации и Министерства просвещения Российской Федерации от 5 августа 2020 г. № 885/390.

8. Устав и локальные нормативные акты Московского политеха.

4.2 Основная литература

1. Шелухин, О.И. Искусственный интеллект и машинное обучение в кибербезопасности: учебно-методическое пособие / О. И. Шелухин, А.В. Осин, Д.И. Раковский. – Москва: МТУСИ, 2022. – 52 с. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/333755> (дата обращения: 01.02.2024). – Режим доступа: для авториз. пользователей.

2. Фомичева, С.Г. Методы машинного обучения в задачах обеспечения информационной безопасности: учебное пособие / С. Г. Фомичева. – Санкт-Петербург:

ГУАП, 2023. – 136 с. – ISBN 978-5-8088-1822-4. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/341024> (дата обращения: 01.02.2024). – Режим доступа: для авториз. пользователей.

4.3 Дополнительная литература

1. Дайзенрот М.П., Альдо Ф.А., Чен С.О. Математика в машинном обучении. – СПб.: Питер, 2024. – 512 с.
2. Роганов Е. А., Тихомиров Н. Б., Шелехов А. М. Математика и информатика для юристов. – М.: МГИУ, 2005. – 364 с.
4. Вентцель Е.С. Теория вероятностей. – М.: Высшая школа, 2006. – 575 с.

4.4 Электронные образовательные ресурсы

1. Jupyter Notebook. – URL: <https://jupyter.org/tryjupyter/notebooks/?path=notebooks/Intro.ipynb>
2. Machine Learning in Python. – URL: <https://scikit-learn.org/stable/index.html>
3. Учебник по машинному обучению. – URL: <https://education.yandex.ru/handbook/ml>

Соответствующий настоящей рабочей программе дисциплины электронный образовательный ресурс находится в разработке.

4.5 Лицензионное и свободно распространяемое программное обеспечение

1. Браузер, Atom: <https://browser.ru>
2. Интегрированная среда разработки, PyCharm Community Edition: <https://www.jetbrains.com/pycharm>

4.6 Современные профессиональные базы данных и информационные справочные системы

1. Java Documentation. – URL: <https://docs.oracle.com/en/java>
2. PyCharm. – URL: <https://www.jetbrains.com/help/pycharm>

5 Материально-техническое обеспечение

Требования к оборудованию в аудиториях для проведения лекций: доска, проектор. Требования к аудиториям для проведения лабораторных занятий: компьютерный класс с доступом в интернет. Необходимое для проведения лекций и лабораторных занятий программное обеспечение приведено в разделе 4.5.

6 Методические рекомендации

6.1 Методические рекомендации для преподавателя по организации обучения

Методические рекомендации по организации лабораторного занятия: 1) обзор теоретического минимума для выполнения лабораторной работы; 2) разбор алгоритмов и примеров по теме работы; 3) консультация по тематике работы; 4) предварительные проверка работы и доведение до студента ошибок или недочетов работы; 5) прием работы с вопросами по решению и заданиями на модификацию решения.

6.2 Методические указания для обучающихся по освоению дисциплины

Методические указания по освоению дисциплины: 1) предварительное изучение теоретического материала лабораторных работ с обращением к источникам из списка литературы; 2) последовательное и своевременное выполнение и защита лабораторных работ.

7 Фонд оценочных средств

7.1 Методы контроля и оценивания результатов обучения

Виды текущего контроля успеваемости студентов: проверка и прием лабораторных работ. Вид промежуточной аттестации – экзамен.

В рамках каждой лабораторной работы студенту предлагается для решения несколько практических разноуровневых задач на определенную тему. На выполнение одной лабораторной работы отводится одна неделя. Все работы подлежат защите.

В рамках экзамена студенту предлагается выполнить за ограниченное время (60 минут) практическое задание по билету, а также ответить на уточняющие и дополнительные вопросы по решению и по тематике курса. При подготовке к ответу и непосредственно при ответе студенту разрешается использовать только среду разработки приложений на языке Python и официальную документацию по специализированным библиотекам языка Python.

7.2 Шкала и критерии оценивания результатов обучения

Для оценки результатов обучения используется балльно-рейтинговая система. Экзаменационная оценка определяется суммой баллов, набранных студентом за выполнение и защиту лабораторных работ в течение семестра и сдачу экзамена. Вклад лабораторных работ в итоговую оценку – до 81 баллов, экзаменационных заданий – до 19 баллов. Правила перевода совокупных баллов в экзаменационную оценку следующие: сумма баллов от 0 до 40 – оценка «неудовлетворительно»; сумма баллов от 41 до 60 баллов – оценка «удовлетворительно»; сумма баллов от 61 до 80 баллов – оценка «хорошо»; сумма баллов от 81 до 100 баллов – оценка «отлично».

Первично результаты выполнения и защиты каждой лабораторной работы оцениваются в баллах по следующей шкале: 3 балла – компетенции считаются освоенными на продвинутом уровне; 2 балла – компетенции считаются освоенными на базовом уровне; 1 балл – компетенции считаются не освоенными. Полученные баллы умножаются на коэффициент сложности: для работ 2, 4 и 6 коэффициент сложности равен 2; для работ 1, 3, 5, 7-10 коэффициент сложности равен 3.

Результаты выполнения экзаменационного задания с учетом ответов на уточняющие и дополнительные вопросы по курсу оцениваются по шкале от 0 до 19 баллов. Детализация баллов за задание: от 18 до 19 баллов – решение и ответ не содержат ошибок или существенных недочетов; от 15 до 17 баллов – решение и ответ не содержат ошибок, но имеют недочеты; от 11 до 14 баллов – решение или ответ содержат ошибки; от 0 до 10 баллов – задание не выполнено или не даны ответы на вопросы по решению.

При повторной промежуточной аттестации необходимым условием получения студентом положительной экзаменационной оценки является получение им не менее 11 баллов за выполнение экзаменационного задания (с учетом ответов на вопросы).

7.3 Оценочные средства

7.3.1 Примеры заданий лабораторного практикума

Пример задания из лабораторной работы 1.

Разработайте приложение для сегментации по яркости пикселей заданного растрового изображения.

Пример задания из лабораторной работы 2.

Подберите размеченный набор данных, отвечающий определению больших данных и пригодный для решения некоторой задачи информационной безопасности. Осуществите программным способом их типовую предобработку.

Пример задания из лабораторной работы 3.

По набору данных из задания 3 на основе метода k ближайших соседей разработайте классификатор.

Пример задания из лабораторной работы 4.

По набору данных из задания 3 на основе метода дерева решений разработайте классификатор.

Пример задания из лабораторной работы 5.

По набору данных из задания 3 на основе наивного метода Байеса разработайте классификатор.

Пример задания из лабораторной работы 6.

По набору данных из задания 2 на основе метода опорных векторов разработайте классификатор.

Пример задания из лабораторной работы 7.

Выполните задания 4-7, спроецировав данные на оси пространства главных компонент. Выполните задания 4-7, подобрав с помощью кросс-валидации наилучшие параметры классификаторов.

Пример задания из лабораторной работы 8.

Примените метод главных компонент к данным из задания 3. Выявите наименее информативные признаки.

Пример задания из лабораторной работы 9.

На основе метода k средних разработайте приложение для разбиения на кластеры данных из заданного CSV-файла.

Пример задания из лабораторной работы 10.

С помощью метода главных компонент оцените сходство заданных текстовых документов по содержанию.

7.3.2 Вопросы и задания для самопроверки

1. Дайте определение векторного пространства.
2. Дайте определение базиса векторного пространства.
3. Дайте определение скалярного произведения. Приведите примеры.
4. Сформулируйте условие ортогональности двух векторов. Приведите примеры.
5. Дайте определение собственного вектора. Приведите примеры.
6. Дайте определение гиперплоскости. Приведите примеры.
7. Дайте определение метрики. Приведите примеры.
8. Проверьте расстояние Хэмминга на соответствие аксиомам расстояния.
9. Выведите формулу для вычисления расстояния между точкой и гиперплоскостью.
10. Дайте классическое определение вероятности. Приведите примеры.
11. Дайте определение условной вероятности. Приведите примеры.
12. Сформулируйте теорему Байеса. Приведите примеры применения.

13. Дайте определение плотности вероятности. Приведите примеры.
14. Дайте определение нормального распределения случайной величины. Приведите примеры.
15. Опишите кривую Гаусса. Приведите примеры.
16. Дайте определение количества информации. Приведите примеры.
17. Дайте определение количества информации. Приведите примеры.
18. Дайте определение энтропии. Приведите примеры.
19. Дайте определение производной функции в точке. Приведите примеры.
20. Дайте определение градиента функции в точке. Приведите примеры.
21. Дайте определение математического ожидания случайной величины.
22. Дайте определение дисперсии случайной величины.
23. Дайте определение среднего квадратического отклонения случайной величины.
24. Сформулируйте правило «трех сигм».
25. Опишите порядок построения гистограммы.
26. Хронологические данные о загрузке процессора компьютера (%CPU) за некоторый период следующие: 1.14, 1.10, 1.09, 1.08, 1.08, 1.08, 1.15, 1.13, 1.09, 1.06, 1.04, 1.03, 1.01, 1.00, 0.99, 0.98, 0.97, 0.95, 0.93, 0.92, 0.90, 0.88, 0.88, 0.87, 0.85, 0.89, 0.86, 0.84, 0.81, 0.79, 0.78, 0.76, 0.74, 0.72, 0.71, 0.71, 0.69, 0.68, 0.67, 0.65, 0.65, 0.64, 0.64, 0.64, 0.67, 0.65, 0.70, 0.66, 0.63, 0.61, 0.61, 0.64, 0.62, 0.62, 0.60, 0.57, 0.57, 0.58, 0.58, 0.67, 0.64, 0.62, 0.62, 0.60, 0.60, 0.64, 0.65, 0.65, 0.68, 0.68, 0.76, 0.76, 0.77, 0.78, 0.84, 0.85, 0.96, 0.97, 0.96, 0.99, 1.11, 1.22, 1.17, 1.16, 1.66, 1.60, 1.68, 2.03, 2.02, 1.79, 1.65, 1.52, 1.47, 1.74, 1.64, 1.86, 1.96, 2.11, 2.20, 2.27, 1.93, 1.73, 1.88, 1.75, 1.58, 1.57, 1.79, 1.69, 1.56, 1.47, 1.40, 1.38, 1.45, 1.37, 1.68, 1.54, 1.47, 1.39, 1.32, 1.72, 1.56, 1.42, 1.41, 1.32, 1.27, 1.24, 1.20, 1.18, 1.15, 1.14, 1.17, 1.15, 1.12, 1.11, 1.44, 1.73, 1.53, 1.37, 1.33, 1.23, 1.19, 1.16, 1.14, 1.11, 1.07, 1.05, 1.10, 1.08, 1.05, 1.12, 1.19, 1.13, 1.08, 1.04, 1.09, 1.04, 1.00, 0.97, 0.94, 0.93, 0.93, 0.93, 0.99, 0.96, 0.98, 0.94, 0.91, 0.91, 0.90, 0.86, 0.91, 0.86, 0.84, 0.81, 0.79, 0.79, 0.78, 0.76, 0.74, 0.73, 0.78, 0.74, 0.72, 0.70, 0.68, 0.65, 0.65, 0.64, 0.70, 0.74, 0.78, 0.79, 0.74, 0.72, 0.69, 0.71, 0.66, 0.63, 0.61, 0.61, 0.70, 0.67, 0.65, 0.65, 0.66, 0.65, 0.65, 0.65, 0.74, 0.73, 0.71, 0.69, 0.68, 0.71, 0.73, 0.76, 0.84, 0.82, 0.82, 0.83, 0.86, 0.88, 0.90, 0.92, 1.02, 1.03, 1.32, 1.41, 1.40, 1.36, 1.36, 1.30, 1.29, 1.27, 1.31, 1.32, 1.35, 1.80, 2.28, 2.01, 1.87, 1.72, 2.03, 1.80, 2.05, 2.18, 1.92, 1.78, 1.64, 1.54, 1.53, 1.58, 1.87, 1.74, 1.72, 1.70, 1.59, 1.85, 1.80, 1.65, 1.51, 1.47, 1.44, 1.42, 1.72, 1.59, 1.58, 1.50, 1.42, 1.43, 1.38, 1.32, 1.31, 1.33, 1.34, 1.39, 1.37, 1.31, 1.59, 1.93, 1.79, 1.62, 1.75, 1.60, 1.59, 1.46, 1.40, 1.35, 1.28, 1.23, 1.24, 1.29, 1.25, 1.28, 1.31, 1.27, 1.41, 1.41, 1.67, 1.48, 1.35, 1.27, 1.24, 1.20, 1.23, 1.22, 1.15, 1.14, 1.21, 1.16, 1.11, 1.09, 1.07, 1.04, 1.02, 1.08, 1.04, 0.99, 0.95, 0.92, 0.88, 0.85, 0.84, 0.84, 0.95, 0.88, 0.85, 0.94, 0.87, 0.81, 0.78, 0.75, 0.73, 0.70, 0.75, 0.71, 0.69, 0.68, 0.69, 0.68, 0.68, 0.68, 0.69, 0.75, 0.77, 0.72, 0.70, 0.70, 0.70, 0.79, 0.77, 0.77, 0.77, 0.79, 0.82, 0.84, 0.87, 0.99, 1.02, 1.02, 1.04, 1.06, 1.10, 1.48, 1.51, 1.53, 1.53, 1.53, 1.79, 1.66, 1.86, 2.04, 1.87, 1.84, 2.45, 2.48, 2.20, 1.92, 2.13, 1.87, 2.13, 2.23, 2.03, 2.14, 1.92, 2.15, 1.95, 1.75, 1.93, 2.13, 2.17, 2.24, 2.40, 2.11, 2.18, 1.97, 2.10, 2.53, 2.55, 2.45, 2.20, 2.26, 1.93, 1.77, 2.07, 1.88, 2.34, 2.44, 2.19, 2.04, 1.98, 2.07, 2.48, 2.45, 2.40, 2.41, 2.16, 1.93, 2.14, 1.96, 2.19, 1.97, 2.15, 1.94, 1.81, 1.67, 1.61, 1.58, 1.92, 2.06, 1.87, 1.81, 1.67, 1.56, 1.56, 1.57, 1.57, 1.56, 1.53, 1.56, 1.48, 1.48, 1.49, 1.46, 1.37, 1.32, 1.28, 1.22, 1.26, 1.20, 1.17, 1.23, 1.18, 1.14, 1.10, 1.05, 1.08, 1.03, 1.08, 1.01, 0.95, 0.91, 0.86, 1.00, 1.00, 0.93, 0.88, 0.84, 0.80, 0.88, 0.82, 0.78, 0.83, 0.78, 0.74, 0.73, 0.71, 0.69, 0.69, 0.68, 0.74, 0.72, 0.69, 0.73, 0.71, 0.71, 0.70, 0.69, 0.69, 0.69, 0.70, 0.78, 0.78, 0.76, 0.79, 0.90, 0.89, 0.93, 0.94, 1.02, 1.02, 1.03, 1.06, 1.08, 1.11, 1.15, 1.20, 1.28, 1.26, 1.30, 1.38, 1.39, 1.42, 1.41, 1.39, 1.45, 1.45, 1.65, 1.59, 1.51, 1.47,

1.73, 1.61, 1.54, 1.49, 1.50, 1.42, 1.46, 1.46, 1.49, 1.75, 1.89, 1.82, 2.06, 1.93, 2.14, 1.92, 2.03, 2.08, 2.15, 2.01, 1.83, 1.77, 1.68, 1.72, 1.70, 2.31, 2.03, 1.87, 1.74, 1.62, 1.54, 1.46, 1.46, 1.53, 1.77, 2.01, 1.84, 1.76, 1.62, 1.55, 1.51, 1.57, 1.53, 1.53, 1.55, 1.55, 1.52, 1.48, 1.49, 1.47, 1.42, 1.35, 1.31, 1.36, 1.33, 1.56, 1.56, 1.47, 1.68, 1.65, 1.88, 2.33, 2.16, 1.91, 1.71, 1.59, 1.48, 1.57, 1.59, 1.51, 1.53, 1.47, 1.45, 1.52, 1.45, 1.78, 1.77, 1.67, 1.67, 1.71, 1.58, 1.86, 1.66, 1.61, 1.52, 1.43, 1.34, 1.30, 1.62, 1.53, 0.04, 0.07, 0.03, 0.07, 0.03, 0.04, 0.06, 0.05, 0.07, 0.09, 0.04, 0.07, 0.03, 0.05, 0.05, 0.05, 0.03, 0.07, 0.06, 1.10, 2.13, 2.83, 2.44, 2.15, 2.05, 1.92, 2.17, 2.02, 1.60, 1.29, 1.31, 1.41, 2.22, 1.98, 1.85, 2.00, 2.12, 2.23, 2.50, 2.50, 2.39, 2.49, 2.18, 2.50, 2.55, 2.23, 2.11, 2.17, 2.15. Постройте гистограмму и найдите границы доверительного интервала. Вычислите вероятность того, что загрузка процессора компьютера примет значение в диапазоне от 2.0 до 2.5.

7.3.3 Вопросы и задания для подготовки к защите лабораторных работ

Вопросы к работе 1:

1. Объясните реализованный алгоритм получения датасета и его разметки.
2. Объясните, как подобрано число интервалов (чем руководствовались).
3. Объясните алгоритм вычисления локальных минимумов.
4. Объясните, как с помощью гистограммы данные исследовались на однородность.
5. Категоризируйте каждую область сгущения значений, исходя из знаний о предметной области.
6. Примените разработанное приложение для сегментации другого контрастного изображения со многими элементами (отделите каждый элемент).

Вопросы к работе 2:

1. Приведите примеры задач информационной безопасности, которые можно решить методами машинного обучения на основе выбранного датасета.
2. Выделите качественные признаки. Ответ обоснуйте.
3. Выделите количественные признаки. Ответ обоснуйте.
4. Выделите целевой признак. Ответ объясните.
5. Опишите и обоснуйте процедуру предобработки качественных признаков.
6. Опишите и обоснуйте процедуру предобработки количественных признаков.
7. Опишите и обоснуйте процедуру исключения неинформативных признаков (на основе диаграмм частот и гистограмм).
8. Опишите и обоснуйте процедуру удаления выбросов.
9. Опишите и обоснуйте процедуру замещения или удаления пропусков.
10. Опишите и обоснуйте процедуру выравнивания объемов выборок классов.

Вопросы к работе 3:

1. Опишите и обоснуйте порядок предобработки данных (с учетом специфики алгоритма классификации).
2. Опишите и объясните алгоритм классификации.
3. Перечислите достоинства и недостатки алгоритма.
4. Опишите и объясните программные процедуры создания и обучения классификатора.
5. Объясните целесообразность нормализации данных (с учетом специфики алгоритма классификации).
6. Объясните целесообразность выравнивания объемов выборок классов (с учетом специфики алгоритма классификации).
7. В чем заключается и как проявляется проблема переобучения классификатора?
8. Продемонстрируйте на примере работоспособность программного решения. Интерпретируйте результат.

Вопросы к работе 4:

1. Опишите и обоснуйте порядок предобработки данных (с учетом специфики алгоритма классификации).
2. Опишите и объясните алгоритм классификации. Как представляются решения?
3. Перечислите достоинства и недостатки алгоритма.
4. Опишите и объясните программные процедуры создания и обучения классификатора.
5. Объясните целесообразность выравнивания объемов выборок классов (с учетом специфики алгоритма классификации).
6. В чем заключается и как проявляется проблема переобучения классификатора?
7. Опишите критерии достоверного решения.
8. Интерпретируйте полученные диаграммы.
9. Продемонстрируйте на примере работоспособность программного решения. Интерпретируйте результат.

Вопросы к работе 5:

1. Опишите и обоснуйте порядок предобработки данных (с учетом специфики алгоритма классификации).
2. Опишите и объясните алгоритм классификации. В чем состоит его «наивность»?
3. Перечислите достоинства и недостатки алгоритма.
4. Опишите и объясните программные процедуры создания и обучения классификатора.
5. Объясните целесообразность нормализации данных (с учетом специфики алгоритма классификации).
6. Объясните целесообразность выравнивания объемов выборок классов (с учетом специфики алгоритма классификации).
7. Продемонстрируйте на примере работоспособность программного решения. Интерпретируйте результат.

Вопросы к работе 6:

1. Опишите и обоснуйте порядок предобработки данных (с учетом специфики алгоритма классификации).
2. Опишите и объясните алгоритм классификации.
3. Перечислите достоинства и недостатки алгоритма.
4. Опишите и объясните программные процедуры создания и обучения классификатора.
5. Объясните целесообразность нормализации данных (с учетом специфики алгоритма классификации).
6. Объясните целесообразность выравнивания объемов выборок классов (с учетом специфики алгоритма классификации).
7. Продемонстрируйте на примере работоспособность программного решения. Интерпретируйте результат.

Вопросы к работе 7:

1. Опишите алгоритм метода главных компонент.
2. Опишите и объясните программную процедуру применения к данным метода главных компонент.
3. Перечислите достоинства и недостатки алгоритма.
4. Объясните целесообразность нормализации данных (с учетом специфики алгоритма).
5. Опишите и объясните алгоритм кросс-валидации.

6. Опишите и объясните программную процедуру проведения кросс-валидации.
7. Опишите и объясните порядок использования тестовых и валидационных данных.
8. Интерпретируйте полученные результаты.
9. Приведите примеры целей применения метода главных компонент.

Вопросы к работе 8:

1. Опишите алгоритм метода главных компонент.
2. Опишите и объясните программную процедуру применения к данным метода главных компонент.
3. Интерпретируйте полученные диаграммы.
4. Обоснуйте полученный результат.

Вопросы к работе 9:

1. Опишите и обоснуйте порядок предобработки данных (с учетом специфики алгоритма кластеризации).
2. Опишите и объясните алгоритм кластеризации.
3. Перечислите достоинства и недостатки алгоритма.
4. Опишите и объясните программную процедуру применения к данным метода кластеризации.
5. Объясните целесообразность нормализации данных (с учетом специфики алгоритма кластеризации).
6. Обоснуйте подбор значения числа кластеров.
7. Разъясните сущность центроидов.
8. Продемонстрируйте на примере работоспособность программного решения. Интерпретируйте результат.

Вопросы к работе 10:

1. Опишите алгоритм метода главных компонент.
2. Опишите и объясните программную процедуру применения к данным метода главных компонент.
3. Интерпретируйте полученные диаграммы.
4. Обоснуйте полученный результат.

7.3.4 Вопросы к экзамену

1. Методы машинного обучения для решения задач информационной безопасности.
2. Предварительное изучение одномерных данных.
3. Предварительная обработка многомерных данных.
4. Классификация, метод k ближайших соседей.
5. Классификация, метод дерева решений.
6. Классификация, наивный метод Баеса.
7. Классификация, метод опорных векторов.
8. Классификация, многослойный персептрон (нейронная сеть прямого распространения).
9. Кросс-валидация модели классификации.
10. Метод главных компонент.
11. Кластеризация, метод k средних.
12. Кластеризация, метод DBSCAN.
13. Кластеризация, метод агломеративной кластеризации.
14. Метод локтя.
15. Метод силуэта.
16. Кластеризация, сеть Кохонена.
17. Кластеризация, карта Кохонена.

18. Удаление выбросов и выявление аномалий в многомерных данных.

7.3.5 Пример экзаменационного билета

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(МОСКОВСКИЙ ПОЛИТЕХ)

Факультет «Информационные технологии», кафедра «Информационная безопасность»
Дисциплина «Машинное обучение и анализ данных»
Образовательная программа 10.05.03 «Информационная безопасность автоматизированных систем»
Курс 4, семестр 2

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Имеется размеченный набор данных (payment_fraud.csv) с описанием транзакций, совершенных при онлайн-покупках в интернет-магазине. Атрибуты транзакции: **accountAgeDays** – время существования учетной записи пользователя (в днях), **numItems** – число приобретенных товаров, **localTime** – время покупки (местное для пользователя), **paymentMethod** – способ оплаты, **paymentMethodAgeDays** – число дней от выбора способа оплаты до покупки. Пример записи: (196, 1, 4.962055, creditcard, 5.10625), здесь $\text{localTime} \approx 4:58$. Значения целевого признака **label**: 0 – нормальная транзакция, 1 – мошенническая транзакция. С помощью *метода опорных векторов* и с применением *метода главных компонент* постройте классификатор для распознавания транзакций. Оцените качество классификатора. Приведите примеры использования классификатора.

Утверждено на заседании кафедры « »

2023 г., протокол № .

Зав. кафедрой _____ /

/