МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Астраханский государственный университет имени В. Н. Татищева»
(Астраханский государственный университет им. В. Н. Татищева)

СОГЛАСОВАНО
Руководитель ОПОП
Ю.А. Головко
«05» мая 2025 г.

УТВЕРЖДАЮ
И.о. заведующего кафедрой
информационных технологий
О.Н. Выборнова
«05» мая 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Программирование для анализа данных

наименование

Составитель(-и)	**Мартьянова А.Е., к.т.н., доцент кафедры информационной безопасности**
Направление подготовки	**09.03.02 Информационные системы и технологии**
Направленность (профиль) ОПОП	**«Технологии разработки и администрирования информационных систем»**
Квалификация (степень)	**бакалавр**
Форма обучения	**очно-заочная**
Год приема	**2022**
Курс	**4 курс**
Семестры	7

Астрахань, 2025

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.1. Целями освоения дисциплины (модуля) «Программирование для анализа данных» являются освоение и развитие навыков программирования на языке Python, представления о сборе, обработке и анализе данных, введение в автоматизированные методы работы с данными - машинное обучение и нейронные сети.

1.2. Задачи освоения дисциплины (модуля):

- освоение продвинутых методов исследования взаимосвязей между показателями, характеризующими объекты в социально-экономических исследованиях;
- освоение продвинутых методов распознавания образов и типологизации объектов;
- освоение продвинутых методов оптимизации представления информации об объектах;
- освоение современных пакетов прикладных программ, реализующих алгоритмы многомерного анализа данных;
- приобретение навыков содержательной интерпретации результатов исследования.

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП

2.1. Учебная дисциплина (модуль) Б1.В.Д.04.02 «Программирование для анализа данных» относится к части элективных дисциплин учебного плана направления подготовки 09.03.02 ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ, профиль «Технологии разработки и администрирования информационных систем» 2022 года набора и осваивается в 7 семестре, общая трудоемкость дисциплины — 4 ЗЕ, 144 часа, итоговая форма контроля — дифзачет.

Дисциплина (модуль) «Программирование для анализа данных» встраивается в структуру ОПОП ВО (последовательность в учебном плане) как с точки зрения преемственности содержания, так и с точки зрения непрерывности процесса формирования компетенций выпускника. Обязательными знаниями для успешного освоения курса является основы математической статистики и теории вероятности, основ программирования на языке Python.

2.2. Для изучения данной учебной дисциплины (модуля) необходимы следующие знания, умения и навыки, формируемые предшествующими учебными дисциплинами (модулями):

- Математические основы информационных технологий и вычислительной техники
- Основы программирования.

знания:

- роль и значение информационных ресурсов в современном обществе,
- виды и формы информации,
- современные информационные технологии обработки информации,
- этапы и методы ее обработки информации,

умения:

- применять компьютерную технику и информационные технологии для обработки информации, и решения практических задач,

навыки:

- владения инструментальных средств информационных технологий обработки информации,
- владения инфокоммуникационных технологий.

2.3. Последующие учебные дисциплины (модули) и (или) практики, для которых необходимы знания, умения и навыки, формируемые данной учебной дисциплиной (модулем):

- Машинное обучение.
- Компьютерное зрение.
- Глубокое обучение.
- Математические основы искусственного интеллекта.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Процесс освоения дисциплины (модуля) направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки (специальности):

в) профессиональной (ПК):

ПК-2. Способен разрабатывать программное обеспечение, включая проектирование, отладку, проверку работоспособности и модификацию ПО.

Таблица 1 - Декомпозиция результатов обучения

Код компетенции	Планируемые результаты обучения по дисциплине (модулю)		
	Знать (1)	Уметь (2)	Владеть (3)
ПК-2: способен разрабатывать программное обеспечение, включая проектирование, отладку, проверку работоспособности и модификацию ПО	ИПК 2.1. Знать: современные информационные технологии разработки, отладки, проверки работоспособности, модификации программного обеспечения	ИПК 2.2. Уметь: осуществлять выбор информационных технологий для решения задач по разработке, отладке, проверке работоспособности, модификации программного обеспечения	ИПК 2.3. Владеть: навыками разработки, отладки, проверки работоспособности, модификации программного обеспечения с использованием современных информационных технологий

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Объем дисциплины (модуля) составляет 4 зачетные единицы, 144 час, в том числе 54 часа, выделенных на контактную работу обучающихся с преподавателем выделено (из них 18 часов – лекции, 36 часов – лабораторные работы, и 90 часов – на самостоятельную работу обучающихся).

Таблица 2 - Структура и содержание дисциплины (модуля)

№ п/п	Раздел, тема дисциплины (модуля)	Семестр	Неделя семестра	Контактная работа (в часах)			Самост. работа		Формы текущего контроля успеваемости, форма промежуточной аттестации *(по семестрам)*
				Л	ПЗ	ЛР	КР	СР	
1	Основы программирования на языке Python	5	1-2	2		4		12	Лабораторная работа №1
2	Математический аппарат	5	3-5	3		6		13	Лабораторная работа №2
3	Введение в	5	6-8	3		6		13	Лабораторная работа №3

	модуль NumPy								
4	Основы работы с Pandas	5	9-10	2		4		12	Лабораторная работа №4. Опрос
5	Анализ данных	5	11-13	3		6		13	Лабораторная работа №5
6	Визуализация данных. Представление результатов	5	14-15	2		4		12	Лабораторная работа №6
7	Работа с текстовыми данными. Текстовый анализ	5	16-18	3		6		15	Лабораторная работа №7. Опрос
ИТОГО			**144**	**18**		**36**		**90**	**Дифзачет**

Условные обозначения:

Л – лекция; ПЗ – практическое занятие, ЛР – лабораторная работа; КР – курсовая работа; СР – самостоятельная работа.

Таблица 3 - Матрица соотнесения разделов, тем учебной дисциплины (модуля) и формируемых в них компетенций

Раздел, тема дисциплины (модуля)	Кол-во часов	Код компетенции	Общее количество компетенций
		ПК - 2	
Основы программирования на языке Python	18	+	1
Математический аппарат	22	+	1
Введение в модуль NumPy	22	+	1
Основы работы с Pandas	18	+	1
Анализ данных	22	+	1
Визуализация данных. Представление результатов	18	+	1
Работа с текстовыми данными. Текстовый анализ	24	+	1
Итого:	144		

Краткое содержание каждой темы дисциплины (модуля)

Тема 1. Основы программирования на языке Python

Введение в анализ данных на языке Python. Почему Python становится стандартом для работы с большими данными. Прикладные задачи политологов, для решения которых подходит язык Python. Основы программирования на языке Python: типы данных и методы работы с ними (переменные, листы, словари, кортежи).

Тема 2. Математический аппарат

Математический аппарат для анализа данных: векторы, матрицы, функции и производные. Основы программирования на языке Python: циклы функции, знание синтаксиса языка.

Тема 3. Введение в модуль NumPy

Введение в модуль для работы с числовыми данными NumPy (Numerical Python)

Особенные типы данных в NumPy. Работа с векторами и матрицами. Вычисление главных статистических метрик с помощью NumPy (среднее, медиана, мода, дисперсия).

Тема 4. Основы работы с Pandas
Введение в модуль для работы с табличным представлением данных Pandas. Преобразование словарей в табличный формат Pandas, сбор и загрузка данных из внешних источников. Особенности фильтрации и обращения к данным. Работа с табличными данными в Pandas на примере.

Тема 5. Анализ данных
Анализ сетей. Введение в машинное обучение. Модуль sklearn. Задачи классификации и линейные модели. Деревья решений. Случайный лес. Ансамбли моделей.

Тема 6. Визуализация данных. Представление результатов
Введение в визуализацию данных. Нюансы визуализации данных и принципы человеческого восприятия. Правила создания хороших визуализаций. Создание различных видов визуализаций на синтетических данных и тренировочных наборах данных.

Тема 7. Работа с текстовыми данными. Текстовый анализ
Введение в анализ текста. Применение. Особенности подготовки данных

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРЕПОДОАВАНИЮ И ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1. Указания для преподавателей по организации и проведению учебных занятий по дисциплине (модулю)

При подготовке к лекционным занятиям необходимо воспользоваться учебно-методической литературой (основной) из п.8.

При подготовке к лабораторным занятиям необходимо воспользоваться учебно-методической литературой (дополнительной) из п.8.

5.2. Указания для обучающихся по освоению дисциплины (модулю)

Во время самостоятельной работы необходимо воспользоваться учебно-методической литературой из п.8 (основной), (дополнительной), Интернет-ресурсами.

Таблица 4 - Содержание самостоятельной работы обучающихся

Номер радела (темы)	Темы/вопросы, выносимые на самостоятельное изучение	Кол-во часов	Формы работы
Основы программирования на языке Python	Подготовка отчета по лабораторной работе 1	12	Внеаудиторная, изучение учебных пособий
Математический аппарат	Подготовка отчета по лабораторной работе 2	13	Внеаудиторная, изучение учебных пособий
Введение в модуль NumPy	Подготовка отчета по лабораторной работе 3	13	Внеаудиторная, изучение учебных пособий
Основы работы с Pandas	Подготовка отчета по лабораторной работе 4. Подготовка к опросу	12	Внеаудиторная, изучение учебных пособий

Номер радела (темы)	Темы/вопросы, выносимые на самостоятельное изучение	Кол-во часов	Формы работы
Анализ данных	Подготовка отчета по лабораторной работе 5	13	Внеаудиторная, изучение учебных пособий
Визуализация данных. Представление результатов	Подготовка отчета по лабораторной работе 6	12	Внеаудиторная, изучение учебных пособий
Работа с текстовыми данными. Текстовый анализ	Подготовка отчета по лабораторной работе 7. Подготовка к опросу	15	Внеаудиторная, изучение учебных пособий

5.3. Виды и формы письменных работ, предусмотренных при освоении дисциплины, выполняемые обучающимися самостоятельно – подготовка реферата.

Не предусмотрено.

6. ОБРАЗОВАТЕЛЬНЫЕ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

При реализации различных видов учебной работы по дисциплине могут использоваться электронное обучение и дистанционные образовательные технологии.

6.1. Образовательные технологии

Учебные занятия по дисциплине могут проводиться с применением информационно-телекоммуникационных сетей при опосредованном (на расстоянии) интерактивном взаимодействии обучающихся и преподавателя в режимах on-line в формах: видеолекций, лекций-презентаций, видеоконференции, собеседования в режиме чат, форума, чата, выполнения виртуальных практических и/или лабораторных работ и др.

Максимальный объем занятий обучающегося с применением электронных образовательных технологий не должен превышать 25%.

Таблица 5 – Образовательные технологии, используемые при реализации учебных занятий

Раздел, тема дисциплины (модуля)	Форма учебного занятия		
	Лекция	Практическое занятие, семинар	Лабораторная работа
Основы программирования на языке Python	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Математический аппарат	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Введение в модуль NumPy	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы

Основы работы с Pandas	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Анализ данных	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Визуализация данных. Представление результатов	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Работа с текстовыми данными. Текстовый анализ	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы

Учебные занятия по дисциплине могут проводиться с применением информационно-телекоммуникационных сетей при опосредованном (на расстоянии) интерактивном взаимодействии обучающихся и преподавателя в режимах on-line в формах: видеолекций, лекций-презентаций, видеоконференции, собеседования в режиме чат, форума, чата, выполнения виртуальных практических и/или лабораторных работ и др.

Максимальный объем занятий обучающегося с применением электронных образовательных технологий не должен превышать 25%.

6.2. Информационные технологии

– использование возможностей интернета в учебном процессе (использование сайта преподавателя (рассылка заданий, предоставление выполненных работ, ответы на вопросы, ознакомление обучающихся с оценками и т.д.));

– использование электронных учебников и различных сайтов (например, электронных библиотек, журналов и т. д.) как источников информации;

– использование возможностей электронной почты преподавателя;

– использование средств представления учебной информации (электронных учебных пособий и практикумов, применение новых технологий для проведения очных (традиционных) лекций и семинаров с использованием презентаций и т. д.);

– использование интегрированных образовательных сред, где главной составляющей являются не только применяемые технологии, но и содержательная часть, т. е. информационные ресурсы (доступ к мировым информационным ресурсам, на базе которых строится учебный процесс);

– использование виртуальной обучающей среды (LMS Moodle «Электронное обучение») или иных информационных систем, сервисов и мессенджеров.

6.3. Программное обеспечение, современные профессиональные базы данных и информационные справочные системы

6.3.1. Программное обеспечение

Наименование программного обеспечения	Назначение
AdobeReader	Программа для просмотра электронных документов

Платформа дистанционного обучения LMS Moodle	Виртуальная обучающая среда
Mozilla FireFox	Браузер
Microsoft Office 2013, Microsoft Office Project 2013 , Microsoft Office Visio 2013	Офисная программа
7-zip	Архиватор
Microsoft Windows 7 Professional	Операционная система
Kaspersky Endpoint Security	Средство антивирусной защиты

Среда разработки Anaconda3 (64 –bit) (в свободном доступе)

6.3.2. Современные профессиональные базы данных и информационные справочные системы

1. Электронный каталог Научной библиотеки АГУ на базе MARKSQL НПО «Информ-систем»: https://library.asu.edu.ru.
2. Электронный каталог «Научные журналы АГУ»: http://journal.asu.edu.ru/.
3. Универсальная справочно-информационная полнотекстовая база данных периодических изданий ООО «ИВИС»: http://dlib.eastview.com/
4. Электронно-библиотечная система elibrary. http://elibrary.ru
5. Справочная правовая система КонсультантПлюс: http://www.consultant.ru
6. Информационно-правовое обеспечение «Система ГАРАНТ»: http://garant-astrakhan.ru

7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

7.1. Паспорт фонда оценочных средств

При проведении текущего контроля и промежуточной аттестации по дисциплине (модулю) «Программирование для анализа данных» проверяется сформированность у обучающихся компетенций, указанных в разделе 3 настоящей программы. Этапность формирования данных компетенций в процессе освоения образовательной программы определяется последовательным освоением дисциплин (модулей) и прохождением практик, а в процессе освоения дисциплины (модуля) – последовательным достижением результатов освоения содержательно связанных между собой разделов, тем.

Таблица 6 - Соответствие разделов, тем дисциплины (модуля), результатов обучения по дисциплине (модулю) и оценочных средств

№ п/п	Контролируемые разделы, темы дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
1	Основы программирования на языке Python	ПК 2	Лабораторная работа №1
2	Математический аппарат.	ПК 2	Лабораторная работа №2
3	Введение в модуль NumPy	ПК 2	Лабораторная работа №3
4	Основы работы с Pandas	ПК 2	Лабораторная работа №4. Вопросы для обсуждения.
5	Анализ данных	ПК 2	Лабораторная работа №5

6	Визуализация данных. Представление результатов	ПК 2	Лабораторная работа №6
7	Работа с текстовыми данными. Текстовый анализ	ПК 2	Лабораторная работа №7. Вопросы для обсуждения

7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

Таблица 7 - Показатели оценивания результатов обучения в виде знаний

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует глубокое знание теоретического материала, умение обоснованно излагать свои мысли по обсуждаемым вопросам, способность полно, правильно и аргументированно отвечать на вопросы, приводить примеры
4 «хорошо»	демонстрирует знание теоретического материала, его последовательное изложение, способность приводить примеры, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует неполное, фрагментарное знание теоретического материала, требующее наводящих вопросов преподавателя, допускает существенные ошибки в его изложении, затрудняется в приведении примеров и формулировке выводов
2 «неудовлетворительно»	демонстрирует существенные пробелы в знании теоретического материала, не способен его изложить и ответить на наводящие вопросы преподавателя, не может привести примеры

Таблица 8 - Показатели оценивания результатов обучения в виде умений и владений

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы
4 «хорошо»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует отдельные, несистематизированные навыки, не способен применить знание теоретического материала при выполнении заданий, испытывает затруднения и допускает ошибки при выполнении заданий, выполняет задание при подсказке преподавателя, затрудняется в формулировке выводов
2 «неудовлетворительно»	не способен правильно выполнить задание

7.3. Контрольные задания и иные материалы, необходимые для оценки результатов обучения по дисциплине (модулю)

Раздел 1. Основы программирования на языке Python

Лабораторная работа 1. Основы программирования на языке Python: типы данных и методы работы с ними (переменные, листы, словари, кортежи).

Цель: ознакомиться с основами языка Python, получить умения для выполнения дальнейших лабораторных работ.

Задачи:

• изучить типизацию данных;

• рассмотреть «ветвление» в Python;

• отработать задачи с использованием конструкции «try-except»;

В ней обучающиеся поэтапно проходят все основные аспекта языка, такие как:

1. Типизация данных.
2. Ветвления.
3. Исключения.
4. Переменные, листы, словари, кортежи

Задания для лабораторной работы № 1:

1. Опишите отличия массивов, кортежей, списков и словарей.
2. Приведите пример кода, который записывает/создает текстовый файл, записывает в него две строчки «`Hello`» и «`123`», а затем считывает его и выводит его содержимое. Код прокомментируйте.
3. Что такое исключения (ошибки) и как их можно обработать? Что лучше выбрать – написать программу так, чтобы не возникало ошибок, или чтобы ошибки были обработанными?

Раздел 2. Математический аппарат

Лабораторная работа 2. Основы программирования на языке Python: циклы, функции, синтаксис языка.

Цель: ознакомиться с основами языка Python: циклы функции, знание синтаксиса языка.

Задачи:

• научиться пользоваться циклами «for» и «while»;

• разобрать функции и пространства имён.

В ней обучающиеся поэтапно проходят все основные аспекта языка, такие как:

1. Пространство имён.
2. Функции.
3. Циклы.

Задания для лабораторной работы № 2:

1. Приведите пример кода, реализующего проверку есть ли словаре запись с каким-то определенным ключом. Например, что контакт есть в телефонной книге.
2. При помощи цикла for выведете таблицу умножения для числа 3. Т.е. число 3 должно умножатся на каждое из [0;9] чисел и результат – выводится пользователю.

Раздел 3. Введение в модуль NumPy

Лабораторная работа 3. Библиотека NumPy.

Цель: получение представления о функциональности, доступных методах и объектах библиотеки NumPy. Изучение основных принципов практической работы с ними.

Задачи:

1. Объект ndarray.
2. Массивы в NumPy.
3. Базовые операции в NumPy.
4. Манипуляции с формой в NumPy.
5. Копии и представления в NumPy.
6. Сохранение массива в файл и чтение из файла в NumPy.

Задания для лабораторной работы № 3:

1. Перечислите наиболее важные атрибуты объектов `ndarray`.
2. Базовая функциональность `NumPy` – опишите основные доступные функции.
3. Создание массива, заполненного нулями – приведите пример кода.
4. Математические операции между массивами. Представьте код, где массив со значениями от 1 до 9 умножается на константу 2. Должна получится таблица умножения для числа 2.
5. Слайсы (обрезка массива) – предоставите код для указанных преподавателем случаев
6. Приведите пример объединения массивов из `NumPy`.
7. Опишите применение унарных операций к массивам, приведите примеры.
8. Копии и представления при работе с массивами.
9. Сохранение массива в файл и чтение из файла.

Раздел 4. Основы работы с Pandas

Лабораторная работа № 4. Библиотека pandas.

Цель: Получение представления о функциональности, доступных методах и объектах библиотеки pandas. Изучение основных принципов практической работы с ними.

Задачи:

1. Структура данных Series.
2. Структура данных DataFrame.
3. Доступ к данным в структурах pandas.
4. Добавление элементов в структуры pandas.
5. Обработка отсутствующих данных в pandas.
6. Сохранение объектов в файл и чтение из файла в pandas.

Задания для лабораторной работы № 4:

1. `Pandas` – основные сведения. Сравнение с функциональностью `NumPy`.
2. Опишите Конструктор класса `Series`.
3. Приведите пример `Series` и попробуйте: выводить на экран, добавлять элементы, изменять значения.
4. Опишите Конструктор класса `DataFrame`.
5. Повторите пункт 3 для объекта `DataFrame`: выводить на экран, добавлять элементы, изменять значения.

6. Открыть при помощи `pandas` созданный в процессе выполнения работы `dataframe.csv` файл с пропусками или некорректными значениями и исправить их.

Вопросы для обсуждения

1. Введение в анализ данных на языке Python. Почему Python становится стандартом для работы с большими данными. Прикладные задачи, для решения которых подходит язык Python. Основы программирования на языке Python: типы данных и методы работы с ними (переменные, листы, словари, кортежи).

2. Математический аппарат для анализа данных: векторы, матрицы, функции и производные

3. Основы программирования на языке Python: циклы функции, знание синтаксиса языка.

4. Введение в модуль для работы с числовыми данными NumPy (Numerical Python). Особенные типы данных в NumPy. Работа с векторами и матрицами. Вычисление главных статистических метрик с помощью NumPy (среднее, медиана, мода, дисперсия).

5. Введение в модуль для работы с табличным представлением данных Pandas. Преобразование словарей в табличный формат Pandas, загрузка данных из внешних источников. Особенности фильтрации и обращения к данным. Работа с табличными данными в Pandas на примере.

Раздел 5. Анализ данных

Лабораторная работа № 5. Библиотека scikit-learn. Обучение с учителем и без учителя.

Цель занятия:

Получение представления о функциональности, доступных методах и объектах библиотеки `scikit-learn`. Изучение основных принципов практической работы с ними для реализации математических алгоритмов. Рассмотрение основных положений машинного обучения с учителем и без учителя: метод k–средних, линейная регрессия и деревья решений.

Обучение с учителем

Обучение с учителем (англ. Supervised learning) – это один из способов машинного обучения, в ходе которого испытуемая система принудительно обучается с помощью примеров «стимул–реакция». К обучению с учителем относятся задачи классификации и регрессии.

Классификация – система группировки объектов исследования или наблюдения в соответствии с их общими признаками. При классификации происходит предсказание признака, множество допустимых значений которого ограничено.

Регрессия – выявление зависимости между случайными переменными и математическим выражением, отражающим связь между зависимой переменной y и независимыми переменными x_i при условии, что это выражение будет иметь статистическую значимость. В отличие от чисто функциональной зависимости $y = f(x_i)$, когда каждому значению независимой переменной x_i соответствует одно определённое значение величины y, при регрессионной связи одному и тому же значению x_i могут соответствовать в зависимости от случая различные значения величины y.

Обучение без учителя

Обучение без учителя (самообучение, спонтанное обучение, англ. Unsupervised learning) — один из способов машинного обучения, при котором испытуемая система спонтанно обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора.

Как правило, оно пригодно только для задач, в которых известны описания множества объектов (т.е. существуют обучающие выборки), и требуется обнаружить внутренние взаимосвязи или закономерности, существующие между объектами.

Примерами обучения без учителя являются:

- кластеризация – задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.
- снижение размерности – представление данных в пространстве меньшей размерности с минимальными потерями полезной информации. Обычно в его основе лежит метод главных компонент).
- выявление аномалий – это опознавание во время интеллектуального анализа данных редких данных, событий или наблюдений, которые вызывают подозрения ввиду существенного отличия от большей части данных.

Функциональность Scikit-learn

Библиотека `scikit-learn` реализует следующие основные методы:

1. **Линейные:** модели, задача которых построить разделяющую или аппроксимирующую гиперплоскость (для классификации и регрессии соответственно).
2. **Метрические:** модели, которые вычисляют расстояние по одной из метрик между объектами выборки, и принимают решения в зависимости от этого расстояния (например, метод K–ближайших соседей).
3. **Деревья решений:** обучение моделей, базирующихся на множестве условий, оптимально выбранных для решения задачи.
4. **Ансамблевые методы:** методы, основанные на деревьях решений, которые комбинируют мощь множества деревьев, и таким образом повышают их качество работы, а также позволяют производить отбор признаков (бустинг, бэггинг, случайный лес, мажоритарное голосование).
5. **Нейронные сети:** комплексный нелинейный метод для задач регрессии и классификации.
6. **Метод опорных векторов** (англ. support vector machine, SVM): нелинейный метод, который обучается определять границы принятия решений.
7. **Наивный Байес:** прямое вероятностное моделирование для задач классификации.
8. **Метод главных компонент** (англ. principal component analysis, PCA): линейный метод понижения размерности и отбора признаков.
9. **Стохастическое вложение соседей с t–распределением** (англ. t-distributed Stochastic Neighbor Embedding, t-SNE): нелинейный метод понижения размерности.
10. **K–средних:** самый распространенный метод для кластеризации, требующий на вход число кластеров, по которым должны быть распределены данные.
11. **Кросс–валидация**: метод, при котором для обучения используется весь датасет (в отличие от разбиения на выборки `train/test`), однако обучение происходит многократно, и в качестве валидационной выборки на каждом шаге выступают разные части датасета. Итоговый результат является усреднением полученных результатов.
12. **Поиск по сетке** (англ. Grid Search): метод для нахождения оптимальных гиперпараметров[2] модели путем построения сетки из значений гиперпараметров и последовательного обучения моделей со всеми возможными комбинациями гиперпараметров из сетки.

Задания для лабораторной работы № 5:

1. Классификация – что такое и где применяется. Задачи классификации и линейные модели.
2. Кластерный анализ – что такое и где применяется, чем отличается от классификации.
3. Линейная регрессия – проведите эксперимент с собственным датасетом (к примеру, подготовьте в `Excel` датасет случайных значений, лежащих возле какого–либо уравнения).
4. Проведите и проанализируйте собственные эксперименты с К–средних и деревом решений.
5. Проведите и проанализируйте собственные эксперименты с деревьями решений, случайным лесом и ансамблями моделей.

Раздел 6. Визуализация данных. Представление результатов

Лабораторная работа № 6. Работа с изображениями и создание различных видов визуализаций

Задание 1. Библиотека TensorFlow Keras. Классификация изображений

Цель: Написание программы для распознавания цифр на базе TensorFlow Keras. Создание нейросети и её обучение по базе данных MNIST. Тестирование MNIST выборкой и пользовательскими картинками.

Создается нейронная сеть, которая классифицирует изображения рукописных цифр. При написании программы используется надстройка «tf.keras», которая является высокоуровневой API для построения и обучения моделей в TensorFlow.

Задачи:

1. Изучение работы с датасетом MNIST.
2. Создание обучающих и тренировочных выборок.
3. Построение модели нейронной сети.
4. Обучение нейронной сети.
5. Проверка точности обученной модели.
6. Предсказание изображений с помощью обученной модели.

Задание 2. Создание различных видов визуализаций на синтетических данных и тренировочных наборах данных

Цель: научить обучающихся основам работы с машинным зрением и показать основные алгоритмы работы с ним.

В последние годы машинное зрение получило большую огласку и вызвало интерес со стороны не только ученых, но и различных инженеров и разработчиков «интеллектуальных» приложений. В лабораторной работе обучающиеся рассмотрят вариант библиотеки OpenCV, написанной на языке «C++» для Python, опробуют некоторый её функционал и протестируют классификатор для распознавания лиц.

Задачи:

1. разобрать импорт и просмотр изображения;
2. разобрать кадрирование;
3. научиться изменять размер изображения;
4. научиться переворачивать изображение;
5. рассмотреть способ преобразование изображения в черно-белое;

6. научиться работать со сглаживанием и размытием;
7. изучить метод распознавания лиц.

Задания для лабораторной работы № 6:

1. Что такое набор данных, или датасет? Для чего может использоваться, как может задаваться (в виде каких типов данных)?
2. Для чего необходимо разбивать датасет на обучающие выборки и валидирующие (тестировочные)? Что дает тестирование?
3. Что такое классы объектов? Зачем им нужны имена? Влияет ли имя класса на обучение, или это дополнительное «удобство» при использовании обученной нейросети?
4. Опишите назначения всех узлов элементов нейросети: входы, синапсы, нейроны, аксоны. Каким узлам соответствуют данные понятия: веса, функции активации?
5. Каким набором вызова функций библиотеки `Keras` можно определять слои и их параметры (функции активации нейронов на слоях, количество нейронов, тип слоя и прочие параметры)?
6. Подготовьте собственную картинку цифры. Протестируйте на ней классификацию при помощи нейросети.

Практическая работа № 1. Создание интерактивных визуализаций и отчетов с помощью инструмента Plotly

Цель: Получение представления о функциональности, доступных методах и объектах библиотеки Plotly.

Введение в визуализацию данных. Нюансы визуализации данных и принципы человеческого восприятия. Правила создания хороших визуализаций.

Задачи:

1. Рассмотреть особенности подготовки данных для визуализации.
2. Изучить нюансы визуализации данных и принципы человеческого восприятия.
3. Научиться создавать интерактивные визуализации и отчеты с помощью инструмента Plotly.

Раздел 7. Работа с текстовыми данными. Текстовый анализ

Лабораторная работа № 7. Работа с текстовыми данными

Задание1. Модель, классифицирующая отзывы покупателей

Цель: Построение прогностической модели, анализирующей отзывы покупателей о продукции и сортирующую их по двум классам: положительные отзывы и отрицательные. Модель учится на уже классифицированных образцах предсказывать класс других образцов и реализована с помощью scikit-learn.

Задачи:

1. Получение набора образцов.
2. Преобразование текста в числовые векторы признаков.
3. Обучение и оценка модели.

Задание2. Обработка естественного языка

Цель: Научиться предварительно обрабатывать и анализировать содержание текстов, используя библиотеку nltk и метод мешка слов, а также осуществлять перевод текста с помощью библиотеки googletrans.

Задачи:

1. Предварительная оценка текста.
2. Разделение на предложения.
3. Разделение на слова.
4. Перевод в нижний регистр, удаление стоп-слов и знаков пунктуации.
5. Лемматизация.
6. Стемминг.
7. Мешок слов.
8. Определение тональности текста.
9. Перевод и величины достоверности для определения языка.

Задания для лабораторной работы № 7:

1. С какой целью используется разделение текста на предложения и слова? Как называется этот процесс? Приведите пример.
2. С какой целью используется перевод в нижний регистр, удаление стоп-слов и знаков пунктуации? Приведите пример.
3. С какой целью используется лемматизация? Приведите пример.
4. С какой целью используется стемминг? Приведите пример.
5. Опишите принцип метода мешка слов. Приведите пример. С какой целью используется этот метод.
6. Как осуществляется определение тональности текста с помощью библиотеки `nltk`? Приведите примеры.
7. Как осуществляется перевод текста? Приведите примеры.
8. Какие коллизии могут возникать при переводе текстов? Приведите примеры.
9. Как определить величину достоверности для определения языка с которого осуществляется перевод? Приведите примеры.
10. Как определить язык исходного текста? Приведите примеры.
11. Выполнить загрузку из указанного файла так, как это указано в строке кода ниже и осуществить обработку текста.

Вопросы для обсуждения

1. Анализ сетей. Введение в машинное обучение. Модуль sklearn. Задачи классификации и линейные модели. Деревья решений. Случайный лес. Ансамбли моделей.

2. Введение в визуализацию данных. Нюансы визуализации данных и принципы человеческого восприятия. Правила создания хороших визуализаций. Создание различных видов визуализаций на синтетических данных и тренировочных наборах данных.

3. Введение в анализ текста. Применение. Особенности подготовки данных

Вопросы к зачету:

1. Язык Python и особенности его стиля программирования. Интерактивный режим Python. Ipython. Jupyter Notebook.

2. Синтаксис и управляющие конструкции языка Python. Переменные, значения и их типы. Типы данных в Python.

3. Встроенные операции и функции. Основные алгоритмические конструкции.

4. Условный оператор. Множественное ветвление.

5. Циклы и счетчики.

6. Определение функций. Параметры и аргументы. Вызовы функций. Оператор возврата. Конструкции *args, **kwargs.

7. Списки, кортежи и словари.

8. Операторы общие для всех типов последовательностей.

9. Специальные операторы и функции для работы со списками. Срезы.

10. Работа со словарями. Методы словарей.

11. Случайные числа. random, randrange, choice.

12. Функции обработки строк. join, replace, split.

13. Стандартная библиотека и pip. Модули и пакеты в Python. Основные стандартные модули.

14. Импортирование модулей. Создание собственных модулей и их импортирование. Специализированные модули и приложения.

15. Файлы и исключения. Работа с внешними источниками данных.

16. Исключения, обработка исключений, вызов исключений (try-except-finally).

17. Утверждения (assert). Открытие, чтение, запись. (open, инструкция with).

18. Работа с текстовыми файлами, xml и csv - файлами.

19. Функциональное программирование. Лямбда-функции.

20. Использование функций map, filter, reduce, zip.

21. Генераторы, декораторы, рекурсия.

22. Модификация функций с помощью декораторов.

23. Итерируемые объекты. Использование генераторов (yield).

24. ООП в Python. Классы, объекты и экземпляры классов. Наследование.

25. Магические методы. Переопределение операторов. Методы классов.

26. Инкапсуляция. Условно частные и строго частные методы.

27. Регулярные выражения. Использование регулярных выражений. Пакет re.

28. Наука о данных и Python. Библиотеки: NumPy, pandas, matplotlib, SciPy.

29. Основы NumPy: массивы и векторные вычисления.

30. Инструменты визуализации данных для Python.

31. Введение в API библиотеки matplotlib.

32. Библиотека pandas. Введение в структуры данных pandas.

33. Объекты Dataframe и Series.

34. Визуализация данных в pandas. Seaborn.

35. Агрегирование данных и групповые операции.

36. Сбор и подготовка данных в Python: извлечение данных с web-страниц (web-scraping). Библиотека beautifulsoup.

37. Работа с динамическими сайтами с помощью Selenium.

38. Массовый скрепинг с помощью scrapy.

39. Работа со структурированными данными: JSON и XML.

40. Открытые API. Telegram API, VK API.

Перечень вопросов и заданий,

выносимых на экзамен / зачёт / дифференцированный зачёт

Таблица 9 – Примеры оценочных средств с ключами правильных ответов

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
ПК-2: способен разрабатывать программное обеспечение, включая проектирование, отладку, проверку работоспособности и модификацию ПО				
1.	Задание закрытого типа	Что такое нормализация данных? 1. Усреднение данных 2. Преобразование категориальных признаков в численные 3. Преобразование численных признаков в категориальные 4. Подгонка под единую шкалу	3	5
2.		Укажите соответствие между типами входных/целевых признаков и диаграммой, которую целесообразно использовать для визуализации 1. Входной признак-категориальный, целевая переменная-категориальная 2. Входной признак-категориальный, Целевая переменная-числовая 3. Входной признак-числовой, Целевая переменная-категориальная 4. Входной признак-числовой, Целевая переменная- числовая a. Диаграмма рассеяния b. Диаграмма размаха c. График плотности d. Мозаичная диаграмма	1-d 2-c 3-b 4-a	5
3.		Для разработки нейросетевых моделей используются библиотеки: 1. Pymorphy2 2. Keras 3. PyTorch 4. OpenCV	2,3	2

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
4.		Для реализации Web-интерфейсов приложений ИИ используются: 1. Flask 2. Git 3. Dash 4. Numba	1,3	3
5.		Какой алгоритм основан на гипотезе «Набор слабых обучающих алгоритмов способен создать сильный обучающий алгоритм»? 1. Бустинг 2. Случайный лес 3. Нейронные сети 4. Наивный Байес	1	3
6.	Задание открытого типа	Предположим, в обучающем множестве у некоторых объектов отсутствуют значения признаков. Какие варианты возможны в данной ситуации?	Можно удалить объекты с пропущенными значениями признаков. Можно преобразовать отсутствующие значения в значимые числа и специально созданную категорию. Можно заполнить значения признаков значением предшествующего экземпляра или медианой столбца.	5
7.		Что такое верность (accuracy) классификации?	Верность-это доля правильно распознанных экземпляров.	5
8.		Какие признаки называются категориальными?	Признаки называются категориальными, если их можно отнести к какой-либо группе, но при этом не важен порядок	5
9.		Какая информация хранится в матрицах сопряженности по результатам тестирования классификатора?	Информация о правильно и неправильно распознанных объектах каждого класса: FP, TP, FN, TN	5
10.		Перечислите основные типы алгоритмов кластеризации	• Иерархический • k-средних • c-средних • Выделение связных компонент • Минимальное покрывающее дерево	5

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			• Послойная кластеризация	

7.4. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине (модулю)

Таблица 10 – Технологическая карта рейтинговых баллов по дисциплине (модулю)

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
Основной блок				
1.	*Выполнение лабораторной работы*	10/7	70	По расписанию
2.	*Ответ во время опроса*	10/2	20	
Всего, дифзачет			**90**	-
Блок бонусов				
3.	*Посещение занятий*	0,25/24	6	
4.	*Своевременное выполнение всех заданий*	0,5/8	4	
Всего			**10**	-
ИТОГО			**100**	-

Таблица 11 – Система штрафов (для одного занятия)

Показатель	Балл
Опоздание на занятие	-0,5
Нарушение учебной дисциплины	-5
Неготовность к занятию	-1
Пропуск занятия без уважительной причины	-2

Таблица 12 – Шкала перевода рейтинговых баллов в итоговую оценку за семестр по дисциплине (модулю)

Сумма баллов	Оценка по 4-балльной шкале
90–100	5 (отлично)
85–89	4 (хорошо)
75–84	
70–74	
65–69	3 (удовлетворительно)
60–64	
Ниже 60	2 (неудовлетворительно)

При реализации дисциплины (модуля) в зависимости от уровня подготовленности обучающихся могут быть использованы иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

8. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

8.1. Основная литература

1. Кузьмич, Р. И. Модификации метода логического анализа данных для задач классификации : монография / Р. И. Кузьмич, И. С. Масич. — Красноярск : Сибирский федеральный университет, 2018. — 181 с. — ISBN 978-5-7638-3698-1. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/84252.html

2. Воронова, Л. И. Machine Learning: регрессионные методы интеллектуального анализа данных : учебное пособие / Л. И. Воронова, В. И. Воронов. — Москва : Московский технический университет связи и информатики, 2018. — 82 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/81325.html

3. Ракитский, А. А. Методы машинного обучения : учебно-методическое пособие / А. А. Ракитский. — Новосибирск : Сибирский государственный университет телекоммуникаций и информатики, 2018. — 32 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/90591.

4. Рашка С., Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения / Рашка С. - М. : ДМК Пресс, 2017. - 418 с. - ISBN 978-5-97060-409-0 - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785970604090.html

8.2. Дополнительная литература

1. Федин, Ф. О. Анализ данных. Часть 1. Подготовка данных к анализу : учебное пособие / Ф. О. Федин, Ф. Ф. Федин. — Москва : Московский городской педагогический университет, 2012. — 204 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/26444.html

2. Билл, Фрэнкс Укрощение больших данных : как извлекать знания из массивов информации с помощью глубокой аналитики / Фрэнкс Билл ; перевод А. Баранов. — Москва : Манн, Иванов и Фербер, 2014. — 340 с. — ISBN 978-5-00057-146-0. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/39433.html

3. Флах П., Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Флах П. - М. : ДМК Пресс, 2015. - 400 с. - ISBN 978-5-97060-273-7 - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785970602737.html

8.3. Интернет-ресурсы, необходимые для освоения дисциплины (модуля)

1. Электронно-библиотечная система (ЭБС) ООО «Политехресурс» «Консультант студента». Многопрофильный образовательный ресурс «Консультант студента» является электронной библиотечной системой, предоставляющей доступ через сеть Интернет к учебной литературе и дополнительным материалам, приобретенным на основании прямых договоров с правообладателями. Каталог в настоящее время содержит около 15000 наименований. www.studentlibrary.ru.

2. Kaggle. Система организации конкурсов по исследованию данных, а также социальная сеть специалистов по обработке данных и машинному обучению. www.kaggle.com

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Для проведения лекционных занятий необходима мультимедийная аудитория, оснащенная компьютерной презентационной техникой.

Для проведения лабораторных занятий необходима аудитория, оснащенная компьютерами.

Рабочая программа дисциплины (модуля) при необходимости может быть адаптирована для обучения (в том числе с применением дистанционных образовательных технологий) лиц с ограниченными возможностями здоровья, инвалидов. Для этого требуется заявление обучающихся, являющихся лицами с ограниченными возможностями здоровья, инвалидами, или их законных представителей и рекомендации психолого-медико-педагогической комиссии. Для инвалидов содержание рабочей программы дисциплины (модуля) может определяться также в соответствии с индивидуальной программой реабилитации инвалида (при наличии).