

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«СЕВЕРО-КАВКАЗСКАЯ ГОСУДАРСТВЕННАЯ АКАДЕМИЯ»

«УТВЕРЖДАЮ»

Проректор по учебной работе

« 30 » 03

2022 г.

 Г.Ю. Нагорная



РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Машинное обучение и анализ данных

Уровень образовательной программы бакалавриат

Направление подготовки 01.03.02 Прикладная математика и информатика

Направленность (профиль) Прикладная математика и информатика

Форма обучения очная

Срок освоения ОП 4 года

Институт Прикладной математики и информационных технологий

Кафедра разработчик РПД Математика

Выпускающая кафедра Математика

Начальник
учебно-методического управления



Семенова Л.У.

Директор института ПМ и ИТ



Тебுவ Д.Б.

Заведующий выпускающей кафедрой



Кочкаров А.М.

г. Черкесск, 2022 г.

СОДЕРЖАНИЕ

1	Цели освоения дисциплины	4
2	Место дисциплины в структуре образовательной программы	4
3	Планируемые результаты обучения по дисциплине	5
4	Структура и содержание дисциплины	6
	4.1. Объем дисциплины и виды учебной работы	6
	4.2. Содержание дисциплины	7
	4.2.1. Разделы (темы) дисциплины, виды учебной деятельности и формы контроля	7
	4.2.2. Лекционный курс	9
	4.2.3. Лабораторный практикум	11
	4.3. Самостоятельная работа обучающегося	14
5	Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине	16
6	Образовательные технологии	20
7	Учебно-методическое и информационное обеспечение дисциплины	21
	7.1. Перечень основной и дополнительной учебной литературы	22
	7.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»	22
	7.3. Информационные технологии, лицензионное программное обеспечение	22
8	Материально-техническое обеспечение дисциплины	23
	8.1. Требования к аудиториям (помещениям, местам) для проведения занятий	23
	8.2. Требования к оборудованию рабочих мест преподавателя и обучающихся:	24
	8.3. Требования к специализированному оборудованию	24
9	Особенности реализации дисциплины для инвалидов и лиц с ограниченными возможностями здоровья	25
	Приложение 1. Фонд оценочных средств	26
	Приложение 2. Аннотация рабочей программы	52
	Рецензия на рабочую программу	53
	Лист переутверждения рабочей программы дисциплины	54

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цель освоения дисциплины "Машинное обучение и анализ данных" формирование у обучающихся базовых теоретических знаний в области теории вероятностей и математической статистики и способности к применению технологий обработки данных (в том числе bigdata) и машинного обучения к решению прикладных задач

При этом **задачами** дисциплины являются:

- изучение основных методов теории вероятностей, математической и прикладной статистики, используемые для подготовки аналитических решений, экспертных заключений и рекомендаций
- изучение технологий обработки больших данных и машинного обучения;
- Изучение методов использования инструментов описательной статистики и визуализации данных,
- изучение вероятностных и статистических методов для решения профессиональных задач;

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

2.1. Дисциплина «Машинное обучение и анализ данных» относится к обязательной части, формируемой участниками образовательных отношений Блока 1. Дисциплины (модули), и имеет тесную связь с другими дисциплинами.

2.2. В таблице приведены предшествующие и последующие дисциплины, направленные на формирование компетенций дисциплины в соответствии с матрицей компетенций ОП.

Предшествующие и последующие дисциплины, направленные на формирование компетенций

№ п/п	Предшествующие дисциплины	Последующие дисциплины
1.	Математический анализ	Преддипломная практика
2.	Линейная алгебра и аналитическая геометрия	Технологическая (проектно – технологическая) практика
3.	Теория вероятностей и математическая статистика	

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Планируемые результаты освоения образовательной программы (ОП) – компетенции обучающихся определяются требованиями стандарта по направлению подготовки 01.03.02 Прикладная математика и информатика и формируются в соответствии с матрицей компетенций ОП

№ п/п	Номер/ индекс компетенции	Наименование компетенции (или ее части)	В результате изучения дисциплины, обучающиеся должны:
1	2	3	4
1.	ОПК-4	Способен решать задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	<p>ОПК-4.1. Обладает базовыми знаниями в области информатики и информационно-коммуникационных технологий</p> <p>ОПК-4.2. Выбирает современные информационно-коммуникационные технологии, необходимые для решения задач профессиональной деятельности</p> <p>ОПК-4.3. Решает задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий</p>

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Вид учебной работы		Семестр	
		Всего часов	№ 7 часов
1		2	3
Аудиторная контактная работа (всего)		56	56
В том числе:			
Лекции (Л)		28	28
Практические занятия (ПЗ)			
Лабораторные работы (ЛР)		28	28
Контактная внеаудиторная работа, в том числе:			
Индивидуальные и групповые консультации		2	2
Самостоятельная работа обучающегося (СРО) (всего)		95	95
<i>Подготовка к занятиям (ЛР)</i>		25	25
<i>Работа с книжными электронными источниками</i>		30	30
<i>Работа с электронным портфолио</i>		30	30
<i>Самоподготовка</i>		10	10
Промежуточная аттестация	Экзамен (Э) в том числе:	Э (27)	Э (27)
	Прием экз., час.	0,5	0,5
	Консультация, час.	2	2
	СРО, час.	24,5	24,5
ИТОГО: Общая трудоемкость	часов	180	180
	зач. ед.	5	5

4.2. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.2.1. Разделы (темы) дисциплины, виды учебной деятельности и формы контроля

№ п/ п	Наименование раздела (темы) дисциплины	Виды учебной деятельности, включая самостоятельную работу обучающихся (в часах)					Формы текущей и промежуточной аттестации
		Л	ЛР	ПЗ	СРО	всего	
1	2	3	4	5	6	7	8
Семестр 7							
1.	Раздел 1. Введение в предметную область.	2	2	-	6	10	тестовый контроль, контрольные вопросы, доклад
2.	Раздел 2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	2	2	-	10	14	тестовый контроль, контрольные вопросы, индивидуальные задания к лабораторным занятиям
3.	Раздел 3. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	4	4	-	10	18	тестовый контроль, контрольные вопросы, индивидуальные задания к лабораторным занятиям
4.	Раздел 4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	4	4	-	10	14	тестовый контроль, контрольные вопросы, индивидуальные задания к лабораторным занятиям
5.	Раздел 5. Обучение на неразмеченных данных. Кластеризация. Иерархическая кластеризация.	2	2	-	10	14	тестовый контроль, контрольные вопросы, индивидуальные задания к лабораторным занятиям

6	Раздел 6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые.	4	4	-	10	18	тестовый контроль, контрольные вопросы, индивидуальные задания к лабораторным занятиям
7	Раздел 7. Постановка задачи регрессии. Линейный регрессионный анализ.	2	2	-	10	14	тестовый контроль, контрольные вопросы
8	Раздел 8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация.	4	4	-	14	22	тестовый контроль, контрольные вопросы
9	Раздел 9. Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев.	2	2	-	15	19	тестовый контроль, контрольные вопросы, индивидуальные задания к лабораторным занятиям
10	Контактная внеаудиторная работа					2	Индивидуальные и групповые консультации
11	Промежуточная аттестация					27	Экзамен
ИТОГО часов в семестре		28	28		95	180	

4.2.2. Лекционный курс

№ п/п	Наименование раздела дисциплины	Наименование темы лекции	Содержание лекции	Всего часов
1	2	3	4	5
Семестр 7				
1.	Раздел 1. Введение в предметную область машинного обучения. Повторение основ программирования на языке Python.	1.1 Введение в предметную область машинного обучения	Примеры использования методов машинного обучения для практических задач. Краткий обзор синтаксиса языка Python. Встроенные операции и функции, типы и структуры данных.	2
		1.2 Примеры использования методов машинного обучения для решения прикладных задач.	Краткий обзор синтаксиса языка Python. Встроенные операции и функции, типы и структуры данных. Повторение основ программирования на языке Python	
2.	Раздел 2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas..	2.1 Библиотеки NumPy и SciPy. Матрицы. Разреженные матрицы. Индексирование, срезы.	Библиотеки NumPy и SciPy. Матрицы. Разреженные матрицы. Индексирование, срезы. Объединение массивов. Библиотека pandas. Запросы к таблицам: выборка строк/столбцов по заданным критериям. Модификация элементов таблицы. Добавление строк/столбцов. Группировка и агрегирование.	2
3	Раздел 3. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	3.1 Важность нормировки данных. Предобработка данных	Описательные статистики. Обзор библиотек matplotlib, seaborn, bokeh. Базовые типы визуализации данных	2
		3.2. Предобработка данных. Работа с пропущенными значениями.	Знакомство с библиотекой scikit-learn (sklearn). Предобработка данных. Метод главных компонент. Работа с пропущенными значениями	2
4	Раздел 4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	4.1. Задачи машинного обучения.	Обучение с учителем и обучение без учителя. Классы задач машинного обучения: регрессия, классификация, кластерный анализ, поиск аномалий.	2
5	Раздел 5.	5.1. Кластеризация.	Использование методов	2

	Обучение на неразмеченных данных. Кластеризация. Иерархическая кластеризация.	Иерархическая кластеризация	снижения размерности и методов кластеризации в задаче распознавания рукописных цифр (MNIST).	
		5.2. Метод Ксредних, DBSCAN и др.	Обзор методов кластеризации, реализованных в библиотеке sklearn Работа с синтетическими данными	
6	Раздел 6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые.	6.1. Разделение данных на обучающие и тестовые. Определение переобученности модели	Примеры задач обучения с учителем. Важность определения целевой метрики качества.	2
		6.2. Критерии оценки качества полученных моделей	Сравнение различных метрик качества моделей. Работа с несбалансированными наборами.	2
7	Раздел 7. Постановка задачи регрессии. Линейный регрессионный анализ.	7.1. Линейный регрессионный анализ. Отбор признаков, коллинеарность, влиятельные наблюдения, анализ остатков	Объединение алгоритмов, реализованных в sklearn, в цепочки и конвейеры с помощью класса Pipeline. Реализация регрессионных и классификационных моделей с помощью sklearn	2
		7.2. Непараметрическая регрессия (ядерное сглаживание). L1 и L2 регуляризация. Метрики качества.	Работа с синтетическими данными. Самостоятельная реализация метода градиентного спуска	
8	Раздел 8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация.	8.1. Постановка задачи классификации, обзор основных методов ее решения.	Реализация классификационных моделей с помощью sklearn. Реализация моделей на основе метода k-ближайших соседей	2
		8.2. Логистическая регрессия. Решающие деревья. Метрики качества классификации (точность/специфичность, ROC-кривая, площадь под кривой)	Метод логистической регрессии. Самостоятельная реализация метода градиентного спуска. Реализация решающего дерева.	2

9	Раздел 9. Ансамбли алгоритмов машинного обучения.	Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев	Реализация моделей с помощью метода градиентного бустинга, метода случайного леса. Блендинг и стеккинг. Методы отбора признаков. Оптимизация гиперпараметров.	2
ИТОГО часов:				28

4.2.3. Лабораторный практикум

№ семестра	Наименование раздела дисциплины	Наименование лабораторной работы	Содержание лабораторной работы	Всего часов
1	2	3	4	5
Семестр 7				
1	Раздел 1. Введение в предметную область.	Лабораторная работа «Основы языка Python».	Лабораторная работа предназначена для ознакомления обучающихся с языком Python и помощи в выполнении лабораторных и проектных заданий. В ней обучающиеся поэтапно проходят все основные аспекта языка	2
2	Раздел 2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	Лабораторная работа «Классификация данных»	Лабораторная работа, в которой начинается изучение принципов машинного обучения. В ней обучающиеся будут учиться работать с графической информацией и производить анализ полученных данных и самостоятельно строить простейший классификатор.	2
		Лабораторная работа «Основы работы с Pandas»	Pandas — это библиотека языка Python, которая позволяет создавать объекты, в которых данные хранятся в табличной форме. Это первая лабораторная работа, связанная с Pandas. В ней обучающиеся разберут	

			структуру и основные функции данной библиотеки.	
3	Раздел 3. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	Лабораторная работа «Классификация методом "К-ближайших соседей"»	В лабораторной работе рассматривается метод машинного обучения «К-ближайших соседей», а также его положительные и отрицательные стороны; рассматривается библиотека для визуализации данных и математического анализа «matplotlib»;	4
4	Раздел 4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	Лабораторная работа «Анализ данных с помощью Pандас»	В данной работе обучающиеся научатся делать срезы, группировки, индексацию данных по разным параметрам, а также строить графики по созданным срезам.	4
5	Раздел 5. Обучение на размеченных данных. Кластеризация. Иерархическая кластеризация.	Лабораторная работа «Дерева решений»	Одним из самых простых для понимания является классификатор «Дерево решений». Эта модель строит «дерево» классов, у каждого из которых есть свои параметры. Попадающие под эти параметры данные причисляются к тому или иному классу. Модель очень чувствительна к переобучению и данным, которые получает на вход.	2
6	Раздел 6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые.	Лабораторная работа «Метод случайного леса»	Случайный лес является модифицированным вариантом «Древа решений» и не имеет всех его отрицательных особенностей, например, он не чувствителен к переобучению. В лабораторной работе будет рассмотрено строение данного метода классификации и	4

			поговорим об его устройстве.	
7	Раздел 7. Постановка задачи регрессии. Линейный регрессионный анализ.	Лабораторная работа «Линейная регрессия»	Линейная регрессия является одним из самых универсальных методов машинного обучения, т.к. может принимать на вход данные различных типов и структур. Также данная модель достаточно устойчива к переобучению и имеет большую скорость обработки данных.	2
8	Раздел 8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация.	Лабораторная работа «Работа с OpenCV»	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,	4
9	Раздел 9. Ансамбли алгоритмов машинного обучения.	Лабораторная работа «Машинное зрение и распознавание лиц в реальном времени»	Лабораторная работа позволит обучающимся больше углубиться в изучение библиотеки OpenCV, а также научит их использовать официальную документацию к программным продуктам	2
ИТОГО:				28

4.2.4. Практические занятия (Не предусмотрены учебным планом)

4.3. САМОСТОЯТЕЛЬНАЯ РАБОТА ОБУЩАЮЩЕГОСЯ

№ п/п	Наименование раздела (темы) дисциплины	№ п/п	Виды СРО	Всего часов
1	2	3	4	5
Семестр 7				
1.	Раздел 1. Введение в предметную область.	1.1.	Самостоятельное изучение материала по теме «Введение в машинное обучение»	6
		1.2.	Работа с основной и дополнительной литературой. Чтение конспекта лекций. Подготовка к лабораторным работам.	
		1.3.	Установка и начало работы с Python	
		1.4.	Изучение конспекта лекций. Выполнения индивидуальных заданий по лабораторной работе.	
		1.5.	Подготовка к лабораторным занятиям. Подготовка к тестированию по первому разделу.	
2.	Раздел 2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	2.1.	Чтение конспекта лекций, работа с литературой. Подготовка к лабораторным работам.	10
		2.2.	Работа с лекционным материалом, поиск и обзор литературы и электронных источников информации по теме практических занятий	
		2.3.	Работа с лекционным материалом, поиск и обзор литературы и электронных источников	
		2.4.	Подготовка к тестированию по второму разделу.	
3.	Раздел 3. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	3.1	Работа с лекционным материалом. Операторы языка программирования.	10
		3.2	Проработка лекций. Работа с электронными источниками.	
		3.3	Подготовка к тестированию по темам третьего раздела	
4.	Раздел 4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения	4.1	Работа с лекционным материалом, поиск и обзор литературы и электронных источников	10
		4.2	Проработка алгоритмов решения задач	
		4.3	Подготовка к тестированию по четвертому разделу.	
5	Раздел 5. Обучение на размеченных данных. Кластеризация. Иерархическая кластеризация.	5.1.	Работа с лекционным материалом. Операторы языка программирования.	10
		5.2	Проработка алгоритмов решения задач логического программирования	
		5.3	Детальное изучение модели «Дерево решений»	

6	Раздел 6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые.	6.1	Работа с лекционным материалом, поиск и обзор литературы и электронных источников	10
		6.2	Чтение конспекта лекций, работа с литературой. Подготовка к лабораторным работам.	
		6.3	Подготовка к тестированию по пятому разделу.	
7	Раздел 7. Постановка задачи регрессии. Линейный регрессионный анализ.	7.1.	Чтение конспекта лекций, работа с литературой. Подготовка к лабораторным работам.	10
		7.2	Работа с лекционным материалом, поиск и обзор литературы и электронных источников информации по теме практических занятий	
		7.3	Подготовка к тестированию по седьмому разделу.	
8	Раздел 8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация.	8.1	Чтение конспекта лекций, работа с литературой. Подготовка к лабораторным работам.	14
		8.2	Чтение конспекта лекций, работа с литературой. Подготовка к лабораторным работам.	
		8.3	Подготовка к тестированию по восьмому разделу.	
9	Раздел 9. Ансамбли алгоритмов машинного обучения.	9.1	Чтение конспекта лекций, работа с литературой. Подготовка к лабораторным работам.	15
		9.2	Чтение конспекта лекций, работа с литературой. Подготовка к лабораторным работам.	
		9.3	Подготовка к итоговому тестированию	
ИТОГО часов:				95

5. ПЕРЕЧЕНЬ УЧЕБНО-МЕТОДИЧЕСКОГО ОБЕСПЕЧЕНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

5.1. Методические указания для подготовки обучающихся к лекционным занятиям

Ведущим методом в лекции выступает устное изложение учебного материала, сопровождающееся демонстрацией видеофильмов, схем, плакатов, показом моделей, приборов, макетов, использование мультимедиа аппаратуры.

Лекция является исходной формой всего учебного процесса, играет направляющую и организующую роль в самостоятельном изучении предмета. Важнейшая роль лекции заключается в личном воздействии лектора на аудиторию.

На лекциях раскрываются основные теоретические аспекты, приводятся примеры реализации на практике, освещается достигнутый уровень формализации деятельности по автоматизации экономических процессов.

Освоение дисциплины предполагает следующие направления работы:

- изучение понятийного аппарата дисциплины;
- изучение тем самостоятельной подготовки по учебно-тематическому плану;
- работу над основной и дополнительной литературой;
- изучение вопросов для самоконтроля (самопроверки);
- самоподготовка к практическим и другим видам занятий;
- самостоятельная работа обучающегося при подготовке к экзамену;
- самостоятельная работа обучающегося в библиотеке;
- изучение сайтов по темам дисциплины в сети «Интернет».

Требуется творческое отношение и к самой программе учебного курса. Вопросы, составляющие ее содержание, обладают разной степенью важности. Есть вопросы, выполняющие функцию логической связки содержания темы и всего курса, имеются вопросы описательного или разъяснительного характера. Все эти вопросы не составляют сути, понятийного, концептуального содержания темы, но необходимы для целостного восприятия изучаемых проблем. Проработка лекционного курса является одной из важных активных форм самостоятельной работы. Лекция преподавателя не является озвученным учебником, а представляет плод его индивидуального творчества. Он читает свой авторский курс со своей логикой со своими теоретическими и методическими подходами. Это делает лекционный курс конкретного преподавателя индивидуально-личностным событием, которым вряд ли обучающемуся стоит пренебрегать. Кроме того, в своих лекциях преподаватель стремится преодолеть многие недостатки, присущие опубликованным учебникам, учебным пособиям, лекционным курсам. Количество часов, отведенных для лекционного курса, не позволяет реализовать в лекциях весь объем содержания дисциплины. Исходя из этого, каждый лектор создает свою тематику лекций, которую в устной или письменной форме представляет обучающимся при первой встрече. Важно обучающемуся понять, что лекция есть своеобразная творческая форма самостоятельной работы. Надо пытаться стать активным соучастником лекции: думать, сравнивать известное с вновь получаемыми знаниями, войти в логику изложения материала лектором, по возможности вступать с ним в мысленную полемику. ВПО время лекции можно задать лектору вопрос. Вопросы можно задать и ВПО время перерыва (письменно или устно), а также после лекции или перед началом очередной. Лектор найдет формы и способы реагирования на вопросы обучающихся.

Обучающийся, готовясь к лекционному занятию, включает выполнение всех видов заданий размещенных в каждой лекции, т.е. задания выполняются еще до лекционного занятия по соответствующей теме. В ходе лекционных занятий, обучающийся должен:

- вести конспектирование лекционного материала;
- обращать внимание на термины и определения, а также формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации.
- обучающийся должен оставить в рабочих конспектах поля, на которых делает

пометки из рекомендованной литературы, дополняющие материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений;

- обучающийся должен задавать преподавателю уточняющие вопросы с целью уяснения теоретических положений, а так же разрешения спорных ситуаций;

- дорабатывать свой конспект лекций, делая в нем соответствующие записи из литературы, рекомендованной преподавателем и предусмотренной учебной программой. Обучающийся должен уметь проводить параллель текущей лекции с предшествующей.

5.2. Методические указания для подготовки обучающихся к лабораторным занятиям

Ведущей дидактической целью лабораторных занятий является систематизация и обобщение знаний по изучаемой теме, приобретение практических навыков по тому или другому разделу курса, закрепление полученных теоретических знаний. Лабораторные работы сопровождают и поддерживают лекционный курс.

Обучающийся должен подготовиться к выполнению лабораторных работ строго в соответствии с содержанием курса.

В начале каждого лабораторного занятия обучающийся должен ознакомиться с теоретическим материалом, необходимым для выполнения текущей лабораторной работы.

Подготовить ответы на контрольные вопросы, которые представляются в конце каждой лабораторной работы.

Каждая лабораторная работа содержит список индивидуальных заданий, на выполнение которых обучающийся должен быть готовым.

Полученные результаты выполнения лабораторной работы обучающийся должен уметь оформить и быть готовым к устной защите.

5.3. Методические указания для подготовки обучающихся к практическим занятиям (Не предусмотрены учебным планом)

5.4. Методические указания по самостоятельной работе обучающихся

Самостоятельная работа обучающихся включает в себя выполнение различного рода заданий, которые ориентированы на более глубокое усвоение материала изучаемой дисциплины. По каждой теме дисциплины обучающимся предлагается перечень заданий для самостоятельной работы. К выполнению заданий для самостоятельной работы предъявляются следующие требования: задания должны исполняться самостоятельно и представляться в установленный срок, а также соответствовать установленным требованиям по оформлению. Обучающимся следует:

- руководствоваться графиком самостоятельной работы, определенным на кафедре;
- выполнять все плановые задания, выдаваемые преподавателем для самостоятельного выполнения, и разбирать на лабораторных и консультациях неясные вопросы;

- при подготовке к экзамену параллельно прорабатывать соответствующие теоретические и практические разделы дисциплины, фиксируя неясные моменты для их обсуждения на плановых консультациях.

В процессе подготовки к практическим занятиям, студентам необходимо обратить особое внимание на самостоятельное изучение рекомендованной учебно-методической (а также научной и популярной) литературы.

Самостоятельная работа с учебниками, учебными пособиями, научной, справочной и популярной литературой, материалами периодических изданий и Интернета, статистическими данными является наиболее эффективным методом получения знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала, формирует у обучающихся свое отношение к

конкретной проблеме.

Более глубокому раскрытию вопросов способствует знакомство с дополнительной литературой, рекомендованной преподавателем по каждой теме семинарского или практического занятия, что позволяет студентам проявить свою индивидуальность в рамках выступления на данных занятиях, выявить широкий спектр мнений по изучаемой проблеме.

Подготовка презентации и доклада

Для подготовки презентации рекомендуется использовать: PowerPoint, MS Word, AcrobatReader, LaTeX-овский пакет beamer. Самая простая программа для создания презентаций – MicrosoftPowerPoint. Для подготовки презентации необходимо собрать и обработать начальную информацию.

Последовательность подготовки презентации:

1. Четко сформулировать цель презентации: вы хотите свою аудиторию мотивировать, убедить, заразить какой-то идеей или просто формально отчитаться.
2. Определить каков будет формат презентации: живое выступление (тогда, сколько будет его продолжительность) или электронная рассылка (каков будет контекст презентации).
3. Отобрать всю содержательную часть для презентации и выстроить логическую цепочку представления.
4. Определить ключевые моменты в содержании текста и выделить их.
5. Определить виды визуализации (картинки) для отображения их на слайдах в соответствии с логикой, целью и спецификой материала.
6. Подобрать дизайн и форматировать слайды (количество картинок и текста, их расположение, цвет и размер).
7. Проверить визуальное восприятие презентации.

К видам визуализации относятся иллюстрации, образы, диаграммы, таблицы. Иллюстрация - представление реально существующего зрительного ряда. Образы – в отличие от иллюстраций - метафора. Их назначение - вызвать эмоцию и создать отношение к ней, воздействовать на аудиторию. С помощью хорошо продуманных и представляемых образов, информация может надолго остаться в памяти человека. Диаграмма - визуализация количественных и качественных связей. Их используют для убедительной демонстрации данных, для пространственного мышления в дополнение к логическому. Таблица - конкретный, наглядный и точный показ данных. Ее основное назначение - структурировать информацию, что порой облегчает восприятие данных аудиторией.

Практические советы по подготовке презентации готовьте отдельно:

- печатный текст + слайды + раздаточный материал;
- слайды - визуальная подача информации, которая должна содержать минимум текста, максимум изображений, несущих смысловую нагрузку, выглядеть наглядно и просто;
- текстовое содержание презентации – устная речь или чтение, которая должна включать аргументы, факты, доказательства и эмоции;
- рекомендуемое число слайдов 17-22;
- обязательная информация для презентации: тема, фамилия и инициалы выступающего; план сообщения; краткие выводы из всего сказанного; список использованных источников;
- раздаточный материал – должен обеспечивать ту же глубину и охват, что и живое выступление: люди больше доверяют тому, что они могут унести с собой, чем исчезающим изображениям, слова и слайды забываются, а раздаточный материал остается постоянным осязаемым напоминанием; раздаточный материал важно раздавать в конце презентации; раздаточный материалы должны отличаться от слайдов, должны быть более информативными.

Тема доклада должна быть согласованна с преподавателем и соответствовать теме учебного занятия. Материалы при его подготовке, должны соответствовать научно-

методическим требованиям вуза и быть указаны в докладе. Необходимо соблюдать регламент, оговоренный при получении задания. Иллюстрации должны быть достаточными, но не чрезмерными.

Работа обучающегося над докладом-презентацией включает отработку умения самостоятельно обобщать материал и делать выводы в заключении, умения ориентироваться в материале и отвечать на дополнительные вопросы слушателей, отработку навыков ораторства, умения проводить диспут.

Докладчики должны знать и уметь: сообщать новую информацию; использовать технические средства; хорошо ориентироваться в теме всего семинарского занятия; дискутировать и быстро отвечать на заданные вопросы; четко выполнять установленный регламент (не более 10 минут); иметь представление о композиционной структуре доклада и др.

Структура выступления

Вступление помогает обеспечить успех выступления по любой тематике. Вступление должно содержать: название, сообщение основной идеи, современную оценку предмета изложения, краткое перечисление рассматриваемых вопросов, живую интересную форму изложения, акцентирование внимания на важных моментах, оригинальность подхода.

Основная часть, в которой выступающий должен глубоко раскрыть суть затронутой темы, обычно строится по принципу отчета. Задача основной части – представить достаточно данных для того, чтобы слушатели заинтересовались темой и захотели ознакомиться с материалами. При этом логическая структура теоретического блока не должны даваться без наглядных пособий, аудио-визуальных и визуальных материалов.

Заключение – ясное, четкое обобщение и краткие выводы, которых всегда ждут слушатели

Методические указания по подготовке к тестированию

Успешное выполнение тестовых заданий является необходимым условием для закрепления изученного материала. Тестовые задания подготовлены на основе лекционного материала, учебников и учебных пособий по дисциплине, изданных за последние 5 лет. Форма изложения тестовых заданий позволяет закрепить и восстановить в памяти пройденный материал. Предлагаемые тестовые задания охватывают узловые вопросы теоретических и практических основ по дисциплине. Для формирования заданий использована закрытая и открытая формы вопросов. У обучающегося есть возможность выбора правильного ответа или нескольких правильных ответов из числа предложенных вариантов. А в вопросах открытой формы дополнить самостоятельно. Для выполнения тестовых заданий обучающиеся должны изучить лекционный материал по теме, соответствующие разделы учебников, учебных пособий и других литературных источников. Репетиционные тестовые задания содержатся в рабочей учебной программе дисциплины. С ними целесообразно ознакомиться при подготовке к контрольному тестированию.

Промежуточная аттестация

По итогам 7 семестра проводится экзамен. При подготовке к сдаче экзамена рекомендуется пользоваться материалами лекционных занятий и материалами, изученными в ходе текущей самостоятельной работы.

Экзамен проводится в устной форме, включает подготовку и ответы обучающегося на теоретические вопросы. По итогам экзамена выставляется оценка.

По итогам обучения проводится экзамен, к которому допускаются студенты, имеющие положительные результаты по защите лабораторных работ.

6. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

№ п/п	Виды учебной работы	Образовательные технологии	Всего часов
1	2	3	4
Семестр 7			
1	<i>Лекция.</i> Знакомство с различными методами предобработки данных, описательными статистиками и основными способами визуализации данных, методами снижения размерности.	Обзорная лекция.	2
2	<i>Лекция.</i> Этапы развития искусственного интеллекта	Технология развития критического мышления	2
3	<i>Лекция.</i> Анализ данных с помощью Pandas	Лекция, презентация с использованием PowerPoint.	2
4	<i>Лекция.</i> Анализ данных, визуализация и классификация	Технология развития критического мышления	2
5	<i>Лекция.</i> Машинное зрение и распознавание лиц в реальном времени	Лекция, презентация с использованием PowerPoint.	2
6	<i>Лекция.</i> Постановка задачи регрессии. Линейный регрессионный анализ.	Технология развития критического мышления	2
7	<i>Лекция.</i> Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных.	Игровые технологии	2
Итого часов			14

УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

7.1. Перечень основной и дополнительной учебной литературы

Основная литература

1. Синева, И. С. Анализ данных в среде R. Ч. 1 : учебное пособие / И. С. Синева. — Москва : Московский технический университет связи и информатики, 2018. — 32 с. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/92422.html> .
2. Любимцева, О. Л. Блочное планирование эксперимента и анализ данных : учебное пособие / О. Л. Любимцева. — Нижний Новгород : Нижегородский государственный архитектурно-строительный университет, ЭБС АСВ, 2018. — 30 с. — ISBN 978-5-528-00276-7. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/80885.html> .
3. Мельниченко, А. С. Математическая статистика и анализ данных : учебное пособие / А. С. Мельниченко. — Москва : Издательский Дом МИСиС, 2018. — 45 с. — ISBN 978-5-906953-62-9. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/78563.html> .

Дополнительная литература

1. Анализ данных качественных исследований : лабораторный практикум / составители А. П. Истомина. — Ставрополь : Северо-Кавказский федеральный университет, 2018. — 108 с. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/92674.html> .
2. Шнарева, Г. В. Анализ данных : учебно-методическое пособие / Г. В. Шнарева, Ж. Г. Пономарева. — Симферополь : Университет экономики и управления, 2019. — 129 с. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/89482.html>
3. . Уэс, Маккинли Python и анализ данных / Маккинли Уэс ; перевод А. А. Слинкин. — Саратов : Профобразование, 2017. — 482 с. — ISBN 978-5-4488-0046-7. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <http://www.iprbookshop.ru/64058.html> .

7.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. <http://fcior.edu.ru> - Федеральный центр информационно-образовательных ресурсов;
2. <http://elibrary.ru>- Научная электронная библиотека.

3.

7.3. Информационные технологии, лицензионное программное обеспечение

Лицензионное программное обеспечение	Реквизиты лицензий/ договоров
Microsoft Azure Dev Tools for Teaching 1. Windows 7, 8, 8.1, 10	Идентификатор подписчика: 1203743421 Срок действия: 30.06.2022
MS Office 2003, 2007, 2010, 2013	(продление подписки) Сведения об Open Office: 63143487, 63321452, 64026734, 6416302, 64344172, 64394739, 64468661, 64489816, 64537893, 64563149, 64990070, 65615073
Антивирус Dr.Web Desktop Security Suite	Лицензия бессрочная Лицензионный сертификат Серийный № 8DVG-V96F-H8S7-NRBC Срок действия: с 20.10.2022 до 22.10.2023
Цифровой образовательный ресурс IPRsmart	Лицензионный договор № 10423/23П от 30.06.2023 г. Срок действия: с 01.07.2023 г. до 01.07.2024г.

Бесплатное ПО: Python, Lazarus, Firebird, IBE Expert, Pascal ABC.

8. Материально-техническое обеспечение дисциплины

8.1. Требования к аудиториям (помещениям, местам) для проведения занятий

Учебная аудитория для проведения занятий лекционного

Специализированная мебель:

Доска меловая – 1 шт., парты – 10 шт., стулья – 30 шт., стул мягкий – 1 шт., стол одностумбовый преподавательский – 1 шт., компьютерные столы – 10 шт.

Технические средства обучения, служащие для предоставления учебной информации большой аудитории:

ПК – 8 шт. Экран настенный рулонный – 1 шт.

Проектор – 1 шт.

Учебная аудитория для проведения занятий семинарского типа, курсового проектирования (выполнение курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации

Специализированная мебель:

Доска меловая – 1 шт., парты – 10 шт., стулья – 30 шт., стул мягкий – 1 шт., стол одностумбовый преподавательский – 1 шт., компьютерные столы – 10 шт.

Технические средства обучения, служащие для предоставления учебной информации большой аудитории:

ПК-8 шт.

Экран настенный рулонный – 1 шт.

Проектор – 1 шт.

Лаборатория метрологии, стандартизации и подтверждения соответствия

Специализированная мебель:

Доска меловая – 1 шт., парты – 10 шт., стулья – 30 шт., стул мягкий – 1 шт., стол одностумбовый преподавательский – 1 шт., компьютерные столы – 10 шт.

Технические средства обучения, служащие для предоставления учебной информации большой аудитории:

ПК – 8 шт.

Экран настенный рулонный – 1 шт.

Проектор – 1 шт.

Помещение для самостоятельной работы.

Библиотечно-издательский центр. Отдел обслуживания печатными изданиями

Специализированная мебель: Рабочие столы на 1 место – 21 шт. Стулья – 55 шт. Набор демонстрационного оборудования и учебно-наглядных пособий, обеспечивающих тематические иллюстрации: экран настенный – 1 шт. Проектор – 1 шт. Ноутбук – 1 шт.

Информационно-библиографический отдел. Специализированная мебель:

Рабочие столы на 1 место - 6 шт. Стулья - 6 шт. Компьютерная техника с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду ФГБОУ ВО «СевКавГА»: персональный компьютер – 1 шт, сканер – 1 шт, МФУ – 1 шт.

Отдел обслуживания электронными изданиями.

Специализированная мебель: рабочие столы на 1 место – 24 шт, стулья – 24 шт.

Набор демонстрационного оборудования и учебно-наглядных пособий, обеспечивающих тематические иллюстрации: интерактивная система - 1 шт., монитор – 21 шт. Сетевой терминал – 18 шт. Персональный компьютер – 3 шт. МФУ – 2 шт. Принтер – 1 шт.

Помещение для хранения и профилактического обслуживания учебного оборудования.

Специализированная мебель: Шкаф – 1 шт., стул – 2 шт., кресло компьютерное – 2 шт., стол угловой компьютерный – 2 шт., тумбочки с ключом – 2 шт. Учебное пособие (персональный

компьютер в комплекте) – 2 шт.

8.2. Требования к оборудованию рабочих мест преподавателя и обучающихся

1. Рабочее место преподавателя, оснащенное компьютером.
2. Рабочие места обучающихся оснащенные компьютерами с доступом в Интернет, предназначенные для работы в электронной образовательной среде, *и т.п.*

8.3. Требования к специализированному оборудованию –нет

9. ОСОБЕННОСТИ РЕАЛИЗАЦИИ ДИСЦИПЛИНЫ ДЛЯ ИНВАЛИДОВ И ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

Для обеспечения образования инвалидов и обучающихся с ограниченными возможностями здоровья разрабатывается (в случае необходимости) адаптированная образовательная программа, индивидуальный учебный план с учетом особенностей их психофизического развития и состояния здоровья, в частности применяется индивидуальный подход к освоению дисциплины, индивидуальные задания: рефераты, письменные работы и, наоборот, только устные ответы и диалоги, индивидуальные консультации, использование диктофона и других записывающих средств для воспроизведения лекционного и семинарского материала.

В целях обеспечения обучающихся инвалидов и лиц с ограниченными возможностями здоровья комплектуется фонд основной учебной литературой, адаптированной к ограничению электронных образовательных ресурсов, доступ к которым организован в БИЦ Академии. В библиотеке проводятся индивидуальные консультации для данной категории пользователей, оказывается помощь в регистрации и использовании сетевых и локальных электронных образовательных ресурсов, предоставляются места в читальном зале.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

ПО ДИСЦИПЛИНЕ Машинное обучение и анализ данных

1. ПАСПОРТ ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ

Машинное обучение и анализ данных.

Компетенции, формируемые в процессе изучения дисциплины

Индекс	Формулировка компетенции
ОПК-4	Способен решать задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий и с учетом основных требований информационной безопасности

2. Этапы формирования компетенции в процессе освоения дисциплины

Основными этапами формирования указанных компетенций при изучении обучающимися дисциплины являются последовательное изучение содержательно связанных между собой разделов (тем) учебных занятий. Изучение каждого раздела (темы) предполагает овладение обучающимися необходимыми компетенциями. Результат аттестации обучающихся на различных этапах формирования компетенций показывает уровень освоения компетенций обучающимися.

Этапность формирования компетенций прямо связана с местом дисциплины в образовательной программе.

Разделы (темы) дисциплины	Формируемые компетенции (коды)
	ОПК-4
Раздел 1. Введение в предметную область.	+
Раздел 2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	+
Раздел 3. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	+
Раздел 4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	+
Раздел 5. Обучение на неразмеченных данных. Кластеризация. Иерархическая кластеризация.	+

<p>Раздел 6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые.</p>	+
<p>Раздел 7. Постановка задачи регрессии. Линейный регрессионный анализ.</p>	+
<p>Раздел 8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и много классовая классификация.</p>	+
<p>Раздел 9. Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев.</p>	+

3. Показатели, критерии и средства оценивания компетенций, формируемых в процессе изучения дисциплины

ОПК-4 - Способен решать задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий и с учетом основных требований информационной безопасности

Индикаторы достижения компетенции	Критерии оценивания результатов обучения				Средства оценивания результатов обучения	
	неудовлетворительно	удовлетворительно	хорошо	отлично	Текущий контроль	Промежуточная аттестация
1	2	3	4	5	6	7
ОПК-4.1 Обладает базовыми знаниями в области информатики и информационно-коммуникационных технологий	Не обладает базовыми знаниями в области информатики и информационно-коммуникационных технологий	Допускает ошибки в базовых знаниях в области информатики и информационно-коммуникационных технологий	Обладает базовыми знаниями в области информатики и информационно-коммуникационных технологий	Демонстрирует глубокие знания в области информатики и информационно-коммуникационных технологий	контрольные вопросы, тестирование.	Экзамен
ОПК-4.2 Выбирает современные информационно-коммуникационные технологии, необходимые для решения задач профессиональной деятельности	Не выбирает современные информационно-коммуникационные технологии, необходимые для решения задач профессиональной деятельности	Демонстрирует в целом удовлетворительные современные информационно-коммуникационные технологии, необходимые для решения задач профессиональной деятельности	Выбирает современные информационно-коммуникационные технологии, необходимые для решения задач профессиональной деятельности	Демонстрирует глубокие знания в использовании современных информационно-коммуникационных технологий, необходимых для решения задач профессиональной деятельности	контрольные вопросы, тестирование, индивидуальные задания к лабораторным занятиям	Экзамен
ОПК-4.3 Решает задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий	Не решает задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий	Решает задачи с неточностями в профессиональной деятельности с использованием существующих информационно-коммуникационных технологий	Решает задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий	Уверенно решает задачи профессиональной деятельности с использованием существующих информационно-коммуникационных технологий	контрольные вопросы, тестирование, индивидуальные задания к лабораторным занятиям.	Экзамен

4. Комплект контрольно-оценочных средств по дисциплине.

Вопросы к экзамену по дисциплине по дисциплине Машинное обучение и анализ данных

1. Общая постановка задачи машинного обучения.
2. Обучение с учителем и без учителя, обучение с подкреплением.
3. Задачи классификации, восстановления регрессии, предсказания.
4. Модели алгоритмов. Признаки. Типы признаков. Понятие функционала качества.
5. Вероятностная постановка задачи. Оценка обобщающей способности.
6. Проблема переобучения. Критерии оценки качества работы алгоритмов машинного обучения. ROC-кривые.
7. Примеры практических задач машинного обучения
8. Инструменты описательной статистики. Измерение центра распределения. Измерение разброса данных.
9. Визуализация качественных признаков. Сводные таблицы и сводные диаграммы. Таблицы сопряженности и парадокс Симпсона. Иерархия признаков.
10. Предварительная обработка данных. Выбросы и их обработка. Пропущенные значения и их обработка. Повторяющиеся строки и их обработка.
11. Условные вероятности. Условная вероятность. Независимость событий. Формула полной вероятности. Формула Байеса.
12. Последовательности испытаний. Биномиальная схема. Формула Бернулли. Формула Пуассона.
13. Определение случайной величины. Понятие случайной величины. Функция распределения случайной величины. Свойства функции распределения. Индикатор события как простейшая случайная величина. Функция распределения индикатора события.
14. Дискретные случайные величины и их важнейшие числовые характеристики. Дискретная случайная величина. Ряд распределения и функция распределения дискретной случайной величины.
15. Первичный анализ данных с библиотекой Pandas
16. Визуальный анализ данных с библиотеками Seaborn и Matplotlib
17. Обобщенный метрический классификатор. Виды метрик.
18. Метод ближайшего соседа. Алгоритм k-ближайших соседей.
19. Взвешенная версия алгоритма k-ближайших соседей.
20. Метод окна Парзена. Метод потенциальных функций.
21. Понятие эталона. Отступы и классификация объектов качества.
22. Классификация, деревья решений и метод ближайших соседей
23. Постановка задач линейной регрессии и линейной классификации.
24. Метод наименьших квадратов в матричной форме. Аналитическое решение.
25. Регуляризация в задачах регрессии.

26. Мультиколлинеарность и плохая обусловленность ковариационной матрицы.
27. Гребневая регрессия. Метод лассо.
28. Линейные классификаторы.
29. Метод стохастического градиента. Улучшение сходимости метода SGD.
30. Логистическая регрессия.
31. Метод опорных векторов.
32. Линейно разделимые выборки.
33. Двойственная задача. Нелинейные обобщения. Возможные виды ядер
34. Линейные модели классификации и регрессии
35. Логистическая регрессия и случайный лес.
36. Линейная регрессия, Lasso и RF-регрессия. Построение и отбор признаков.
37. Методы кластеризации. Типы кластерных структур.
38. Функционал качества кластеризации. EM-алгоритм.
39. Метод k-средних. Иерархическая кластеризация.
40. Формула Ланса-Уильямса.
41. Быстрая агломеративная кластеризация.
42. Основы анализа временных рядов
43. Обучение без учителя. Метод главных компонент. Кластеризация

СЕВЕРО-КАВКАЗСКАЯ ГОСУДАРСТВЕННАЯ
АКАДЕМИЯ

Кафедра Математика
20__ - 20__ учебный год

Экзаменационный билет № 1

по дисциплине Машинное обучение и анализ данных

для обучающихся направления подготовки 01.03.02 Прикладная математика и информатика

1. Общая постановка задачи машинного обучения.
2. Линейная регрессия, Lasso и RF-регрессия. Построение и отбор признаков.

Зав. кафедрой

Кочкаров А.М

Контрольные вопросы

по дисциплине Машинное обучение и анализ данных.

Вопросы к разделу 1.

1. Общая постановка задачи машинного обучения.
2. Обучение с учителем и без учителя, обучение с подкреплением.
3. Задачи классификации, восстановления регрессии, предсказания.
4. Модели алгоритмов. Признаки. Типы признаков.
5. Понятие функционала качества.
6. Вероятностная постановка задачи.
7. Оценка обобщающей способности. Проблема переобучения.
8. Критерии оценки качества работы алгоритмов машинного обучения. ROC-кривые.
9. Примеры практических задач машинного обучения

Вопросы к разделу 2.

1. Библиотеки NumPy и SciPy.
2. Матрицы. Разреженные матрицы.
3. Индексирование, срезы. Объединение массивов.
4. Библиотека pandas.
5. Запросы к таблицам: выборка строк/столбцов по заданным критериям. Модификация элементов таблицы.
6. Добавление строк/столбцов. Группировка и агрегирование. Объединение таблиц (различные виды join).
7. Многомерные данные: мультииндексы. Операции stack-unstack. Построение сводных таблиц (pivottables).

Вопросы к разделу 3.

1. Описательные статистики.
2. Обзор библиотек matplotlib, seaborn, bokeh.
3. Базовые типы визуализации данных.
4. Знакомство с библиотекой scikit-learn (sklearn).
5. Предобработка данных.

Вопросы к разделу 4.

1. Пакет sklearn.
2. Основные функции.
3. Работа с данными из набора MNIST (рукописные цифры).
4. Работа с синтетическими данными

Вопросы к разделу 5.

1. Обучение на неразмеченных данных.
2. Нормировка данных.
3. Кластеризация.
4. Иерархическая кластеризация.
5. Метод K-средних, DBSCAN и др.
6. Обзор методов кластеризации, реализованных в библиотеке sklearn.

Вопросы к разделу 6

1. Обучение с учителем и без учителя, обучение с подкреплением.
2. Задачи классификации, восстановления регрессии, предсказания.
3. Модели алгоритмов. Признаки. Типы признаков.
4. Понятие функционала качества. Вероятностная постановка задачи.
5. Оценка обобщающей способности. Проблема переобучения.
6. Критерии оценки качества работы алгоритмов машинного обучения. ROC-кривые.
7. Примеры практических задач машинного обучения

Вопросы к разделу 7

1. Постановка задач линейной регрессии и линейной классификации.
2. Метод наименьших квадратов в матричной форме.
3. Аналитическое решение. Регуляризация в задач регрессии.
4. Мультиколлинеарность и плохая обусловленность ковариационной матрицы.
5. Гребневая регрессия.
6. Метод лассо.
7. Линейные классификаторы.
8. Метод стохастического градиента.
9. Улучшение сходимости метода SGD.
10. Логистическая регрессия.
11. Метод опорных векторов.
12. Линейно разделимые выборки. Двойственная задача.
13. Нелинейные обобщения. Возможные виды ядер

Вопросы к разделу 8

1. Постановка задачи классификации, обзор основных методов ее решения.
2. Бинарная и многоклассовая классификация.
3. Логистическая регрессия.
4. Решающие деревья.
5. Метрики качества
6. Классификации(точность/специфичность, ROC-кривая, площадь под кривой).
7. Метод логистической регрессии.
8. Самостоятельная реализация метода градиентного спуска

Вопросы к разделу 9

1. Ансамбли алгоритмов машинного обучения.
2. Агрегирование моделей.
3. Ансамбли решающих деревьев. Метод случайного леса. Градиентный бустинг.

Комплект заданий для лабораторных занятий по дисциплине «Машинное обучение и анализ данных»

Лабораторная работа №1

Описание лабораторной работы «Основы языка Python»

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

Задачи:

- изучить типизацию данных;
- научиться пользоваться циклами «for» и «while»;
- рассмотреть «ветвление» в Python;
- отработать задачи с использованием конструкции «try-except»;
- разобрать функции и пространства имён.

Данная лабораторная работа предназначена для ознакомления обучающихся с языком Python и помощи в выполнении лабораторных и проектных заданий. В ней обучающиеся поэтапно проходят все основные аспекта языка, такие как:

1. Типизация данных.
2. Пространство имён.
3. Функции.
4. Циклы.
5. Ветвления.
6. Исключения

Лабораторная работа №2

Описание лабораторной работы «Классификация данных»

Цель: научиться работать с данными при помощи визуальных инструментов и разобрать азы классификации при помощи построения простейшего классификатора со статичными параметрами.

Задачи:

- научиться анализировать данные;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;
- построить классификатор на основе данных полученных при анализе;
- научиться калибровать нейросеть для получения более точных ответов.

Первая лабораторная работа, в которой начинается изучение принципов машинного обучения. В ней обучающиеся будут учиться работать с графической информацией и производить анализ полученных данных и самостоятельно строить простейший классификатор. Данная работа поможет понять простейшие понятия и принципы работы более сложных классификаторов

Лабораторная работа №3

Описание лабораторной работы «Классификация методом "К-ближайших соседей"»

Цель: изучить метод простейший метод классификации данных «К-ближайших соседей» и научиться производить оценку данных с помощью визуальных инструментов Python.

Задачи:

- детально разобрать метод машинного обучения «К-ближайших соседей»;
- научиться работать с информацией;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;

- научиться калибровать нейросеть для получения более точных ответов.

В этой лабораторной работе будет рассматриваться метод машинного обучения «К-ближайших соседей», а также его положительные и отрицательные стороны; рассмотрим библиотеку для визуализации данных и математического анализа «matplotlib»; будем производить калибровку нейронной сети для получения максимального процента правильных ответов, а также обсудим проблему переобучения в нейросетях.

Лабораторная работа №4

Описание лабораторной работы «Основы работы с Pandas»

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для визуализации данных в дата сетах.

Задачи:

- получить умения по использованию библиотеки Pandas;
- сформировать понятия о DataFrame и Series;
- научиться строить графики с помощью scattermatrix (матрица рассеивания) и matplotlib.

Pandas — это библиотека языка Python, которая позволяет создавать объекты, в которых данные хранятся в табличной форме. Это основной структурный объект необходимый для анализа и очистки данных перед их использованием в обучении модели. Это первая лабораторная работа, связанная с Pandas. В ней обучающиеся разберут структуру и основные функции данной библиотеки.

Лабораторная работа №4

Описание лабораторной работы «Анализ данных с помощью Pandas»

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для анализа данных в дата сетах.

Задачи:

- получить умения по использованию библиотеки Pandas;
- научиться анализировать и обрабатывать данные с помощью Pandas;
- закрепить умения визуализации в Pandas.

В продолжении к лабораторной работе № 4 обучающиеся переходят к анализу данных с помощью библиотеки Pandas. В данной работе обучающиеся научатся делать срезы, группировки, индексацию данных по разным параметрам, а также строить графики по созданным срезам. Знания и умения полученные в ходе выполнения работы помогут понять, некоторые механизмы машинного обучения.

Лабораторная работа №5

Описание лабораторной работы «Линейная регрессия»

Цель: понять и научиться применять метод линейной регрессии в машинном обучении.

Задачи:

- изучить модель линейного регрессора;
- произвести обучение модели;
- рассмотреть особенности данного метода машинного обучения;
- произвести предсказание на основе созданной модели.

Линейная регрессия является одним из самых универсальных методов машинного

обучения, т.к. может принимать на вход данные различных типов и структур. Также данная модель достаточно устойчива к переобучению и имеет большую скорость обработки данных. В этой лабораторной будут рассматриваться все нюансы и принципы работы этого метода классификации.

Лабораторная работа №6

Описание лабораторной работы «Деревья решений»

Цель: познакомить обучающихся с методом машинного обучения, построенном на деревьях решений, а также научить строить сами деревья.

Задачи:

- рассмотреть понятие дерева решений;
- рассмотреть варианты применения данной классификации;
- обучить модель на основе классов;
- отобразить дополнительный класс на модели и посмотреть результат;
- рассмотреть плюсы и минусы данной модели.

Одним из самых простых для понимания является классификатор «Дерево решений». Эта модель строит «дерево» классов, у каждого из которых есть свои параметры. Попадающие под эти параметры данные причисляются к тому или иному классу. Модель очень чувствительна к переобучению и данным, которые получает на вход. Лабораторная работа поможет нам разобраться во всех тонкостях данного метода и покажет все его нюансы.

Лабораторная работа №7

Описание лабораторной работы «Метод случайного леса»

Цель: сформировать понятие случайного леса, а также научить обучающихся использовать данную модель для решения задач.

Задачи:

- рассмотреть понятие случайного леса;
- рассмотреть пример кода для решения простых задач;
- научить подбирать параметры модели для улучшения качества прогнозов модели.

Случайный лес является модифицированным вариантом «Дерева решений» и не имеет всех его отрицательных особенностей, например, он не чувствителен к переобучению. В лабораторной работе будет рассмотрено строение данного метода классификации и поговорим об его устройстве

Лабораторная работа №8.

Описание лабораторной работы «Работа с OpenCV»

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

Задачи:

- разобрать импорт и просмотр изображения;
- разобрать кадрирование;
- научиться изменять размер изображения;
- научиться переворачивать изображение;
- рассмотреть способ преобразование изображения в черно-белое;
- научиться работать со сглаживанием и размытием;
- изучить метод распознавания лиц.

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 «С++» для Python, опробуем некоторый её функционал и протестируем классификатор для распознавания лиц

ЗАДАНИЕ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Описание проектных заданий

1 Описание задания «Машинное зрение и распознавание лиц в реальном времени»

Цель: сформировать умения по использованию библиотеки OpenCV самостоятельному поиску информации, связанной с решением поставленной задачи.

Задачи:

- закрепить умения работы с машинным зрением;
- сформировать умение поиска информации;
- сформировать умения по разработке полноценного приложения;
- научиться отрабатывать ошибки при разработке приложения.

Самостоятельная работа неотъемлемая задача каждого обучающегося и именно поэтому обучающимся предоставляется возможность самим развиваться в данной области. Проектное задание «Машинное зрение и распознавание лиц в реальном времени» позволит обучающимся больше углубиться в изучение библиотеки OpenCV, а также научит их использовать официальную документацию к программным продуктам.

2 Описание задания «Оптимизация и переобучение на примере метода "K-соседей"»

Цель: сформировать у обучающихся понятие о переобучении и научить оптимизации модели машинного обучения.

Задачи:

- закрепить умения по использованию метода «K-соседей»;
- сформировать понятие переобучения;
- научиться оптимизировать модель;

Любая модель обучения имеет свои особенности и черты, но есть вещь, которая объединяет их — это переобучение. В данной работе обучающиеся будут тестировать свою модель на устойчивость и рассматривать на графиках последствия переобучения, пытаться избегать подобных эффектов и правильно настраивать свою модель.

3 Описание задания «Анализ данных, визуализация и классификация»

Цель: научиться использовать различные методы классификации и визуализировать данные разными средствами.

Задачи:

- научиться использовать различные методы классификации;
- научиться работать с имеющимися данными;
- закрепить умения по использованию инструментов визуализации.

Работа с данными — это один из самых сложных этапов машинного обучения и данная работа поможет обучающимся еще глубже погрузиться в неё. Обучающиеся должны

рассмотреть, как будут вести себя различные модели машинного обучения при работе с различными наборами данных. Так как все модели по-разному работают с разными типами данных, а некоторые и вовсе поддерживают только определенные типы данных, то и результат может очень сильно отличаться.

Комплект тестовых вопросов

по дисциплине Машинное обучение и анализ данных

1. Суммарная оперативная память IBM Watson составляет порядка:
 - А) 100 гигабайт
 - Б) 5000 терабайт
 - В) 10 зетабайт
 - Г) 15 терабайт
2. Термин Большие данные определил - _____
3. Данные, которые занимают больше мировой памяти относительно остальных – это _____
4. BigData – это _____
5. Компания, первым создала технологию MapReduce – это
 - А) Google
 - Б) Yahoo
 - В) EMC
 - Г) Oracle
6. Данные текстовых файлов с определенными паттернами для их обработки (например, XML) являются:
 - А) Структурированными
 - Б) Полуструктурированными
 - В) Квазиструктурированными
 - Г) Неструктурированными
7. Термин «BigData» в информационных технологиях определяет
 - А) Комплексный набор методов для создания файлов большого объема
 - Б) Комплексный набор методов обработки структурированных и неструктурированных данных колоссальных объемов.
 - В) Файлы с большим количеством данных.
 - Г) Представление времени, дня, месяца и года в качестве значения количеств миллисекунд, прошедших с начала нашей эры.
8. Данные имеющие определенный тип, формат и структуру (например, транзакционные данные) являются:
 - А) Структурированными
 - Б) Полуструктурированными
 - В) Квазиструктурированными
 - Г) Неструктурированными
9. Примерный объем всей существующей на земле информации (в байтах) ...
 - А) 10^{11}
 - Б) 10^{21}
 - В) $10^{1010101}$
 - Г) 10^{171}
10. Средством анализа в BI (Business Intelligence) является _____

11. Основное умение исследователя данных характеризуется ...
- А) Умение находить наиболее важные элементы в хранимой информации
 - Б) Уметь прогнозировать исход работы системы
 - В) Находить скрытые логические связи в системе собранной информации
 - Г) Отличать неструктурированные данные от структурированных
12. Определите, какой язык программирования из перечисленных является наиболее важным для аналитика
- А) C++
 - Б) PHP
 - В) F#
 - Г) R
13. Термин «BusinessIntelligence» в информационных технологиях – это _____
14. Языком, на котором был разработан RabbitMQ, является:
- А) Java
 - Б) Python
 - В) C++
 - Г) Erlang
15. Главным результатом процесса Business Intelligence является _____
16. Не является средством анализа ...
- А) Продвинутая визуализация
 - Б) Reporting
 - В) Predictive Modelling
 - Г) Data Mining
17. К средствам предоставления информации в «Business Intelligence» относится ...
- А) Генератор нерегламентированных запросов
 - Б) Совместная работа и управление рабочими процессами
 - В) Предиктивное моделирование и DataMining
 - Г) Карты показателей
18. Процессом создания и выбора модели для предсказания вероятности наступления некоторого события является:
- А) OLAP
 - Б) Data Mining
 - В) Predictive Modelling
 - Г) Data Science
19. Целью процесса не является BusinessIntelligence
- А) Интерпретация большого количества данных;
 - Б) Моделирование исходов различных вариантов действий;
 - В) Модификация существующего программного обеспечения;
 - Г) Отслеживание результатов решений.
20. Определите что из этого не является реализацией Hadoop
- А) Google MapReduce
 - Б) Phoenix

- В) GreenMint
 - Г) Qizmt
21. Определите, какие из перечисленных пунктов являются достоинствами MapReduce:
- А) Оптимальная производительность
 - Б) Эффективное применение в маленьких кластерах с небольшим объемом данных
 - В) Масштабируемость
 - Г) Отказоустойчивость
22. Oozie – это система рабочих процессов, определяемая как _____
23. Какие компоненты являются частями MapReduce
- А) Task Tracker
 - Б) Name Node и Data Node
 - В) Job Tracker и Task Tracker
 - Г) Job Tracker, Task Tracker, Name Node и Data Node
24. Spark как программная платформа, определяющая структуру программной системы, является ...
- А) Инструментом для кластерных вычислений
 - Б) Графическим движком
 - В) Библиотекой для работы с графами
 - Г) Технологией распределенных вычислений
25. Дайте определение MapReduce _____
26. Что из перечисленного является недостатком MapReduce
- А) Фиксированный алгоритм обработки данных
 - Б) Масштабируемость
 - В) Отказоустойчивость
 - Г) Возможность автоматического распараллеливания
27. Какое API было добавлено в Hadoop v2.0
- А) YAWN
 - Б) YARN
 - В) SARN
 - Г) DARN
28. Целью NameNode в HDFS является хранение _____
29. Вертикальное масштабирование...
- А) Требуется изменений в прикладных программах, работающих на таких системах
 - Б) Не требует никаких изменений в прикладных программах, работающих на таких системах
 - В) Уменьшает производительность каждого компонента БД
 - Г) Увеличивает скорость загрузки данных
30. Термин NoSQL означает _____

Темы докладов

по дисциплине «Машинное обучение и анализ данных»

1. Современные тенденции развития BigData в мировой компьютерной индустрии.
2. Современные тенденции развития BigData в Российской Федерации.
3. Предпосылки развития BigData в КЧР.
4. Современное состояние операционной системы Hadoop.
5. Отрасли народного хозяйства КЧР, в которых может быть задействована операционная система Hadoop.
6. Перечень народно-хозяйственных задач в КЧР, которые целесообразно решать с помощью технологии MapReduce.
7. Применение технологий NoSQL в современных бизнес проектах.
8. Перспективы применения среды R для анализа статистических данных, необходимых для принятия решений в управлении г. Черкесск.
9. Отрасли народного хозяйства КЧР, в которых могут быть задействованы системы класса DataMining.
10. Перспективы применения кластерного анализа и его результатов для управления экономикой КЧР.

Методические материалы, определяющие процедуры оценивания компетенции

5.1 Критерии оценивания качества выполнения лабораторного практикума

Оценка **«зачтено»** выставляется обучающемуся, если лабораторная работа выполнена правильно и обучающийся ответил на все вопросы, поставленные преподавателем на защите. Оценка **«не зачтено»** выставляется обучающемуся, если лабораторная работа выполнена не правильно или обучающийся не проявил глубоких теоретических знаний при защите работы

5.2 Критерии оценивания качества устного ответа на контрольные вопросы

Оценка **«отлично»** выставляется за глубокое знание предусмотренного программой материала, за умение четко, лаконично и логически последовательно отвечать на поставленные вопросы.

Оценка **«хорошо»** – за твердое знание основного (программного) материала, за грамотные, без существенных неточностей ответы на поставленные вопросы.

Оценка **«удовлетворительно»** – за общее знание только основного материала, за ответы, содержащие неточности или слабо аргументированные, с нарушением последовательности изложения материала.

Оценка **«неудовлетворительно»** – за незнание значительной части программного материала, за существенные ошибки в ответах на вопросы, за неумение ориентироваться в материале, за незнание основных понятий дисциплины.

5.3 Критерии оценивания тестирования

При тестировании все верные ответы берутся за 100%.

90%-100% отлично

75%-90% хорошо

60%-75% удовлетворительно

менее 60% неудовлетворительно

5.4 Критерии оценивания результатов освоения дисциплины на экзамене:

Оценка **«отлично»** выставляется за глубокое знание предусмотренного программой материала, содержащегося в основных и дополнительных рекомендованных литературных источниках, за умение четко, лаконично и логически последовательно отвечать на поставленные вопросы, за умение анализировать изучаемые явления в их взаимосвязи и диалектическом развитии, применять теоретические положения при решении практических задач.

Оценка **«хорошо»** – за твердое знание основного (программного) материала, включая расчеты (при необходимости), за грамотные, без существенных неточностей ответы на поставленные вопросы, за умение применять теоретические положения для решения практических задач.

Оценка **«удовлетворительно»** – за общее знание только основного материала, за ответы, содержащие неточности или слабо аргументированные, с нарушением последовательности изложения материала, за слабое применение теоретических положений при решении практических задач.

Оценка **«неудовлетворительно»** – за незнание значительной части программного материала, за существенные ошибки в ответах на вопросы, за неумение ориентироваться в расчетах, за незнание основных понятий дисциплины

5.5 Критерии оценивания результатов освоения дисциплины

Оценка **«отлично»** выставляется за глубокое знание предусмотренного программой материала, содержащегося в основных и дополнительных рекомендованных литературных источниках, за умение четко, лаконично и логически последовательно отвечать на поставленные вопросы, за умение анализировать изучаемые явления в их взаимосвязи и диалектическом развитии, применять теоретические положения при решении практических задач.

Оценка **«хорошо»** – за твердое знание основного (программного) материала, включая расчеты (при необходимости), за грамотные, без существенных неточностей ответы на поставленные вопросы, за умение применять теоретические положения для решения практических задач.

Оценка **«удовлетворительно»** – за общее знание только основного материала, за ответы, содержащие неточности или слабо аргументированные, с нарушением последовательности изложения материала, за слабое применение теоретических положений при решении практических задач.

Оценка **«неудовлетворительно»** – за незнание значительной части программного материала, за существенные ошибки в ответах на вопросы, за неумение ориентироваться в расчетах, за незнание основных понятий дисциплины.

5.6 Критерии оценивания доклада

Оценка **«отлично»** выставляется за глубокое знание предусмотренного программой материала, за умение четко, лаконично и логически последовательно отвечать на поставленные вопросы.

Оценка **«хорошо»** – за твердое знание основного (программного) материала, за грамотные, без существенных неточностей ответы на поставленные вопросы.

Оценка **«удовлетворительно»** – за общее знание только основного материала, за ответы, содержащие неточности или слабо аргументированные, с нарушением последовательности изложения материала.

Оценка **«неудовлетворительно»** – за незнание значительной части программного материала, за существенные ошибки в ответах на вопросы, за неумение ориентироваться в материале, за незнание основных понятий дисциплины.