МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ
имени И.Т. ТРУБИЛИНА»

Факультет прикладной информатики
Системного анализа и обработки информации

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ) «СОВРЕМЕННЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ И МАШИННОГО ОБУЧЕНИЯ»

Уровень высшего образования: бакалавриат

Направление подготовки: 38.03.05 Бизнес-информатика

Направленность (профиль)подготовки: Анализ, моделирование и формирование интегрального представления стратегий и целей, бизнес-процессов и информационно-логической инфраструктуры предпри

Квалификация (степень) выпускника: бакалавр

Форма обучения: очная

Год набора: 2024

Срок получения образования: 4 года

Объем: в зачетных единицах: 2 з.е.
в академических часах: 72 ак.ч.

2024

Разработчики:
Доцент, кафедра системного анализа и обработки информации Павлов Д.А.

Рабочая программа дисциплины (модуля) составлена в соответствии с требованиями ФГОС ВО по направлени ю подготовки 38.03.05 Бизнес-информатика, утвержденного приказом Минобрнауки от 29.07.2020 № 838, с учетом трудовых функций профессиональных стандартов: "Менеджер по информационным технологиям", утвержден приказом Минтруда России от 30.08.2021 № 588н; "Специалист по информационным системам", утвержден приказом Минтруда России от 13.07.2023 № 586н; "Системный аналитик", утвержден приказом Минтруда России от 27.04.2023 № 367н.

Согласование и утверждение

№	Подразделение или коллегиальный орган	Ответственное лицо	ФИО	Виза	Дата, протокол (при наличии)
1	Системного анализа и обработки информации	Заведующий кафедрой, руководитель подразделения, реализующего ОП	Барановская Т.П.	Согласовано	08.04.2024, № 8

1. Цель и задачи освоения дисциплины (модуля)

Цель освоения дисциплины - развитие навыков анализа данных и применения машинного обучения для извлечения бизнес-выводов и решения практических задач

Задачи изучения дисциплины:

- Изучение языка программирования Python.Уверенное владение языком на базовом уровне для решения задач аналитики. Умение применять специальные библиотеки Numpy, Pandas для анализа бизнес-данных. Выработка навыков осуществления первичного анализа данных, построения визуализаций данных и их интерпретация. Ознакомление с принципами извлечения данных с внешних ресурсов и обработки полученных данных. Проработка математической основы использования машинного обучения, применение линейной алгебры и математического анализа для осуществления алгоритмов. Знакомство с машинным обучением в контексте задач классического машинного обучения без учителя. Знакомство с машинным обучением в контексте задач классического машинного обучения с учителем..

2. Планируемые результаты обучения по дисциплине (модулю), соотнесенные с планируемыми результатами освоения образовательной программы

Компетенции, индикаторы и результаты обучения

ОПК-5 Способен организовывать взаимодействие с клиентами и партнерами в процессе решения задач управления жизненным циклом информационных систем и информационно-коммуникационных технологий

ОПК-5.1 Использует современные средства электронной коммуникации при взаимодействии с клиентами и партнерами

Знать:

ОПК-5.1/Зн1 Знает современные средства электронной коммуникации при взаимодействии с клиентами и партнерами

Уметь:

ОПК-5.1/Ум1 Умеет использовать знания о современных средствах электронной коммуникации при взаимодействии с клиентами и партнерами

Владеть:

ОПК-5.1/Нв1 Владеет знаниями использования современных средств электронной коммуникации при взаимодействии с клиентами и партнерами

ОПК-5.2 Выявляет основные потребности клиентов и партнеров в области управления ИТ-сервисами

Знать:

ОПК-5.2/Зн1 Знает методы выявления основных потребностей клиентов и партнеров в области управления ИТ-сервисами

Уметь:

ОПК-5.2/Ум1 Умеет выявлять основные потребности клиентов и партнеров в области управления ИТ-сервисами

Владеть:

ОПК-5.2/Нв1 Владеет знаниями определения основных потребностей клиентов и партнеров в области управления ИТ-сервисами

ОПК-5.3 Демонстрирует умение взаимодействовать с клиентами и партнерами при заключении договоров на предоставление услуг ИТ-сервисов

Знать:

ОПК-5.3/Зн1 Знает методы взаимодействия с клиентами и партнерами при заключении договоров на предоставление услуг ИТ-сервисов

Уметь:

ОПК-5.3/Ум1 Умеет применять методы взаимодействия с клиентами и партнерами при заключении договоров на предоставление услуг ИТ-сервисов

Владеть:

ОПК-5.3/Нв1 Владеет знаниями взаимодействия с клиентами и партнерами при заключении договоров на предоставление услуг ИТ-сервисов

ОПК-5.4 Разрабатывает индикаторы качества оказания услуг ИТ-сервисов, проводит их мониторинг и оценку

Знать:

ОПК-5.4/Зн1 Знает методы управления качеством оказания услуг ИТ-сервисов, проведения их мониторинга и оценки

Уметь:

ОПК-5.4/Ум1 Умеет разрабатывать индикаторы качества оказания услуг ИТ-сервисов, проводить их мониторинг и оценку

Владеть:

ОПК-5.4/Нв1 владеет знаниями разработки индикаторов качества оказания услуг ИТ-сервисов, проведения их мониторинга и оценки

ОПК-5.5 Организует процесс оказания услуг ИТ-сервиса, проводит эскалацию и закрытие инцидентов

Знать:

ОПК-5.5/Зн1 Знает способы организации процесса оказания услуг ИТ-сервиса, проведения эскалации и закрытия инцидентов

Уметь:

ОПК-5.5/Ум1 Умеет организовывать процесс оказания услуг ИТ-сервиса, проводить эскалацию и закрытие инцидентов

Владеть:

ОПК-5.5/Нв1 Владеет знаниями организации процесса оказания услуг ИТ-сервиса, проведения эскалации и закрытия инцидентов

ПК-П12 Способность использовать знание основных методов искусственного интеллекта в последующей профессиональной деятельности в качестве научных сотрудников, преподавателей образовательных организаций высшего образования, инженеров, технологов

ПК-П12.1 Знает методы разработки оригинальных алгоритмов и программных решений с использованием современных технологий

Знать:

ПК-П12.1/Зн1 Предметная область автоматизации

ПК-П12.1/Зн2 Системы хранения и анализа баз данных

ПК-П12.1/Зн3 Основы программирования

ПК-П12.1/Зн4 Современные объектно-ориентированные языки программирования

ПК-П12.1/Зн5 Современные структурные языки программирования

ПК-П12.1/Зн6 Современные стандарты информационного взаимодействия систем

ПК-П12.1/Зн7 Современный отечественный и зарубежный опыт в профессиональной деятельности

Уметь:

ПК-П12.1/Ум1 Кодировать на языках программирования

ПК-П12.1/Ум2 Тестировать результаты прототипирования

Владеть:

ПК-П12.1/Нв1 Владеть методами кодирования на языках программирования

ПК-П12.2 Умеет использовать методы разработки оригинальных алгоритмов и программных решений с использованием современных технологий

Знать:

ПК-П12.2/Зн1 Предметная область автоматизации

ПК-П12.2/Зн2 Возможности ИС

ПК-П12.2/Зн3 Источники информации, необходимой для профессиональной деятельности

Уметь:

ПК-П12.2/Ум1 Анализировать входные данные

ПК-П12.2/Ум2 Планировать работы

Владеть:

ПК-П12.2/Нв1 Владеет знаниями разработки оригинальных алгоритмов и программных решений с использованием современных технологий

ПК-П12.3 Владеть навыками применения методов разработки оригинальных алгоритмов и программных решений с использованием современных технологий

Знать:

ПК-П12.3/Зн1 Языки программирования и работы с базами данных

ПК-П12.3/Зн2 Инструменты и методы верификации структуры программного кода

ПК-П12.3/Зн3 Основы программирования

ПК-П12.3/Зн4 Современные объектно-ориентированные языки программирования

ПК-П12.3/Зн5 Современные структурные языки программирования

ПК-П12.3/Зн6 Современный отечественный и зарубежный опыт в профессиональной деятельности

Уметь:

ПК-П12.3/Ум1 Кодировать на языках программирования

ПК-П12.3/Ум2 Верифицировать структуру программного кода

Владеть:

ПК-П12.3/Нв1 Владеет навыками применения методов разработки оригинальных алгоритмов и программных решений с использованием современных технологий

3. Место дисциплины в структуре ОП

Дисциплина (модуль) «Современные методы анализа данных и машинного обучения» относится к обязательной части образовательной программы и изучается в семестре(ах): 7.

В процессе изучения дисциплины студент готовится к решению типов задач профессиональной деятельности, предусмотренных ФГОС ВО и образовательной программой.

4. Объем дисциплины и виды учебной работы

Период обучения	Общая трудоемкость (часы)	Общая трудоемкость (ЗЕТ)	Контактная работа (часы, всего)	Внеаудиторная контактная работа (часы)	Зачет (часы)	Лабораторные занятия (часы)	Лекционные занятия (часы)	Самостоятельная работа (часы)	Промежуточная аттестация (часы)

Седьмой семестр	72	2	47	1		30	16	25	Зачет
Всего	72	2	47	1		30	16	25	

5. Содержание дисциплины

5.1. Разделы, темы дисциплины и виды занятий

(часы промежуточной аттестации не указываются)

Наименование раздела, темы	Всего	Внеаудиторная контактная работа	Лабораторные занятия	Лекционные занятия	Самостоятельная работа	Планируемые результаты обучения, соотнесенные с результатами освоения программы
Раздел 1. Введение в машинное обучение	**14**		**6**	**4**	**4**	ОПК-5.1 ОПК-5.2
Тема 1.1. Основные понятия и задачи в машинном обучении	6		2	2	2	
Тема 1.2. Работа с данными и признаками	8		4	2	2	
Раздел 2. Линейные модели	**26**		**12**	**6**	**8**	ОПК-5.1 ОПК-5.2 ОПК-5.3 ОПК-5.4 ОПК-5.5 ПК-П12.1 ПК-П12.2 ПК-П12.3
Тема 2.1. Линейная регрессия	14		6	4	4	
Тема 2.2. Линейная классификация	12		6	2	4	
Раздел 3. Ансамблевые методы	**22**	**1**	**8**	**4**	**9**	ПК-П12.1 ПК-П12.2 ПК-П12.3
Тема 3.1. Решающие деревья	10		4	2	4	
Тема 3.2. Бэггинг, случайные леса и разложение ошибки на смещение и разброс	12	1	4	2	5	
Раздел 4. Снижение размерности	**10**		**4**	**2**	**4**	ПК-П12.1 ПК-П12.2 ПК-П12.3
Тема 4.1. Снижение размерности	10		4	2	4	
Итого	**72**	**1**	**30**	**16**	**25**	

5.2. Содержание разделов, тем дисциплин

Раздел 1. Введение в машинное обучение
(Лабораторные занятия - 6ч.; Лекционные занятия - 4ч.; Самостоятельная работа - 4ч.)

Тема 1.1. Основные понятия и задачи в машинном обучении
(Лабораторные занятия - 2ч.; Лекционные занятия - 2ч.; Самостоятельная работа - 2ч.)

Введение в машинное обучение.
Виды задач и методов машинного обучения.
Основные термины, постановки задач и примеры применения.

Тема 1.2. Работа с данными и признаками
(Лабораторные занятия - 4ч.; Лекционные занятия - 2ч.; Самостоятельная работа - 2ч.)
Работа с данными и признаками.
Метод k ближайших соседей.

Раздел 2. Линейные модели
(Лабораторные занятия - 12ч.; Лекционные занятия - 6ч.; Самостоятельная работа - 8ч.)

Тема 2.1. Линейная регрессия
(Лабораторные занятия - 6ч.; Лекционные занятия - 4ч.; Самостоятельная работа - 4ч.)
Область применимости линейных моделей
Измерение ошибки в задачах регрессии
Переобучение
Оценивание качества моделей
Обучение линейной регрессии
Градиентный спуск и оценивание градиента
Модификация градиентного спуска
Регуляризация
Преобразования признаков

Тема 2.2. Линейная классификация
(Лабораторные занятия - 6ч.; Лекционные занятия - 2ч.; Самостоятельная работа - 4ч.)
Линейные модели классификации
Метрики качества классификации
Логистическая регрессия
Метод опорных векторов
Многоклассовая классификация

Раздел 3. Ансамблевые методы
(Внеаудиторная контактная работа - 1ч.; Лабораторные занятия - 8ч.; Лекционные занятия - 4ч.; Самостоятельная работа - 9ч.)

Тема 3.1. Решающие деревья
(Лабораторные занятия - 4ч.; Лекционные занятия - 2ч.; Самостоятельная работа - 4ч.)
Построение деревьев
Критерии информативности
Критерии останова
Методы стрижки деревьев

Тема 3.2. Бэггинг, случайные леса и разложение ошибки
на смещение и разброс
(Внеаудиторная контактная работа - 1ч.; Лабораторные занятия - 4ч.; Лекционные занятия - 2ч.; Самостоятельная работа - 5ч.)

Бутстрап
Смещение и разброс
Бэггинг
Случайные леса
Бустинг
Современные методы градиентного бустинга
Стекинг

Раздел 4. Снижение размерности
(Лабораторные занятия - 4ч.; Лекционные занятия - 2ч.; Самостоятельная работа - 4ч.)

Тема 4.1. Снижение размерности
(Лабораторные занятия - 4ч.; Лекционные занятия - 2ч.; Самостоятельная работа - 4ч.)
PCA
tSNE
Кластеризация

6. Оценочные материалы текущего контроля

Раздел 1. Введение в машинное обучение

Форма контроля/оценочное средство: Расчетно-графическая работа

Вопросы/Задания:

1. Задание состоит из двух разделов, посвященных работе с табличными данными с помощью библиотеки pandas и визуализации с помощью matplotlib. В первом разделе вам предстоит выполнить базовые задания с помощью вышеуказанных библиотек, а во втором распределить студентов по курсам. Баллы даются за выполнение отдельных пунктов. Задачи в рамках одного раздела рекомендуется решать в том порядке, в котором они даны в задании.

Задание направлено на освоение jupyter notebook (будет использоваться в дальнейших заданиях), библиотек pandas и matplotlib.

Задание состоит из двух разделов, посвященных работе с табличными данными с помощью библиотеки pandas и визуализации с помощью matplotlib. В первом разделе вам предстоит выполнить базовые задания с помощью вышеуказанных библиотек, а во втором распределить студентов по курсам. Баллы даются за выполнение отдельных пунктов. Задачи в рамках одного раздела рекомендуется решать в том порядке, в котором они даны в задании.

Задание направлено на освоение jupyter notebook (будет использоваться в дальнейших заданиях), библиотек pandas и matplotlib.

Раздел 2. Линейные модели

Форма контроля/оценочное средство: Кейс-задание

Вопросы/Задания:

1. В этом задании мы попытаемся научиться анализировать данные и выделять из них полезные признаки. Мы также научимся пользоваться seaborn и sklearn, а заодно привыкнем к основным понятиям машинного обучения.

В этом задании мы попытаемся научиться анализировать данные и выделять из них полезные признаки. Мы также научимся пользоваться seaborn и sklearn, а заодно привыкнем к основным понятиям машинного обучения.

Задание 1. Мы будем работать с данными из соревнования New York City Taxi Trip Duration, в котором нужно было предсказать длительность поездки на такси. Скачайте обучающую выборку из этого соревнования и загрузите ее.

Задание 2. Для начала давайте посмотрим, сколько всего было поездок в каждый из дней.

Постройте график зависимости количества поездок от дня в году (например, можно воспользоваться sns.countplot).

Задание 3. Нарисуйте на одном графике зависимости количества поездок от часа в сутках для разных месяцев (разные кривые, соответствующие разным месяцам, окрашивайте в разные цвета, воспользуйтесь hue в sns.relplot). Аналогично нарисуйте зависимости количества поездок от часа в сутках для разных дней недели.

Задание 4. Разбейте выборку на обучающую и тестовую в отношении 7:3 (используйте train_test_split из sklearn). По обучающей выборке нарисуйте график зависимости среднего логарифма времени поездки от дня недели. Затем сделайте то же самое, но для часа в сутках и дня в году.

Задание 5. Обучите Ridge-регрессию с параметрами по умолчанию, закодировав все категориальные признаки с помощью OneHotEncoder. Численные признаки отмасштабируйте с помощью StandardScaler. Используйте только признаки, которые мы выделили в этой части задания.

Задание 6. Как мы все прекрасно помним, ср
, поэтому очевидно, что самым сильным признаком будет расстояние, которое необходимо проехать. Мы не можем посчитать точное расстояние, которое необходимо преодолеть такси, но мы можем его оценить, посчитав кратчайшее расстояние между точками начала и конца поездки. Чтобы корректно посчитать расстояние между двумя точками на Земле, можно использовать функцию haversine. Также можно воспользоваться кодом с первого семинара. Посчитайте кратчайшее расстояние для объектов и запишите его в колонку haversine

Задание 7. Давайте изучим среднюю скорость движения такси. Посчитайте среднюю скорость для каждого объекта обучающей выборки, разделив haversine на trip_duration, и нарисуйте гистограмму ее распределения

Задание 8. Для каждого из замеченных вами выше 2-3 пунктов добавьте в выборку по два признака:

началась ли поездка в этом пункте
закончилась ли поездка в этом пункте
Как вы думаете, почему эти признаки могут быть полезны?

Задание 9. Сейчас мы почти что не используем сами значения координат. На это есть несколько причин: по отдельности рассматривать широту и долготу не имеет особого смысла, стоит рассматривать их вместе. Во-вторых, понятно, что зависимость между нашим таргетом и координатами не линейная. Чтобы как-то использовать координаты, можно прибегнуть к следующему трюку: обрамим область с наибольшим количеством поездок прямоугольником (как на рисунке). Разобьем этот прямоугольник на ячейки. Каждой точке сопоставим номер ее ячейки, а тем точкам, что не попали ни в одну из ячеек, сопоставим значение -1.

Напишите трансформер, который сначала разбивает показанную на рисунке область на ячейки, а затем создает два признака: номер ячейки, в которой началась поездка, и номер ячейки, в которой закончилась поездка. Количество строк и столбцов выберите самостоятельно.

Обратите внимание, что все вычисления должны быть векторизованными, трансформер не должен модифицировать передаваемую ему выборку inplace, а все необходимые статистики (если они вдруг нужны) нужно считать только по обучающей выборке в методе fit.

Задание 10. Обучите Ridge-регрессию со стандартными параметрами на признаках, которые

мы выделили к текущему моменту. Категориальные признаки закодируйте через one-hot-кодирование, числовые признаки отмасштабируйте.

Форма контроля/оценочное средство: Расчетно-графическая работа

Вопросы/Задания:

1. В данном задании необходимо реализовать обучение линейной регрессии с помощью различных вариантов градиентного спуска.

Задание 1. Реализация градиентного спуска
В этом задании вам предстоит написать собственные реализации различных подходов к градиентному спуску с опорой на подготовленные шаблоны в файле descents.py.

Все реализуемые методы должны быть векторизованы!

Задание 2. Реализация линейной регресии
В этом задании вам предстоит написать свою реализацию линейной регресии, обучаемой с использованием градиентного спуска, с опорой на подготовленные шаблоны в файле linear_regression.py - LinearRegression. По сути линейная регрессия будет оберткой, которая запускает обучение

Необходимо соблюдать следующие условия:

Все вычисления должны быть векторизованы;
Циклы средствами python допускаются только для итераций градиентного спуска;
В качестве критерия останова необходимо использовать (одновременно):
Квадрат евклидовой нормы разности весов на двух соседних итерациях меньше tolerance;
Разность весов содержит наны;
Достижение максимального числа итераций max_iter.
Будем считать, что все данные, которые поступают на вход имеют столбец единичек последним столбцом;
Чтобы проследить за сходимостью оптимизационного процесса будем использовать loss_history, в нём будем хранить значения функции потерь до каждого шага, начиная с нулевого (до первого шага по антиградиенту) и значение функции потерь после оптимизации.

Задание 3. Работа с данными
Мы будем использовать датасет объявлений по продаже машин на немецком Ebay. В задаче предсказания целевой переменной для нас будет являться цена.

Постройте график распределения целевой переменной в данных, подумайте, нужно ли заменить её на логарифм. Присутствуют ли выбросы в данных с аномальной ценой? Если да, то удалите их из данных.

Проведите исследование данных:

Проанализируйте тип столбцов, постройте графики зависимости целевой переменной от признака, распределения значений признака;
Подумайте, какие признаки могут быть полезными на основе этих графиков, обработайте выбросы;
Подумайте, какие трансформации признаков из известных вам будет уместно применить;
Разделите полезные признаки на категориальные, вещественные и те, которые не надо предобрабатывать.
Разделите данные на обучающую, валидационную и тестовую выборки в отношении 8:1:1.

Задание 4. Сравнение методов градиентного спуска (2 балла)
В этом задании вам предстоит сравнить методы градиентного спуска на подготовленных вами данных из предыдущего задания.

Задание 5. Стохастический градиентный спуск и размер батча (1 балл)
В этом задании вам предстоит исследовать влияние размера батча на работу стохастического градиентного спуска.

Сделайте по несколько запусков (например, k = 10) стохастического градиентного спуска на обучающей выборке для каждого размера батча из перебираемого списка. Замерьте время в секундах и количество итераций до сходимости. Посчитайте среднее этих значений для каждого размера батча.
Постройте график зависимости количества шагов до сходимости от размера батча.
Постройте график зависимости времени до сходимости от размера батча.
Посмотрите на получившиеся результаты. Какие выводы можно сделать про подбор размера батча для стохастического градиентного спуска?

Задание 6. Регуляризация
В этом задании вам предстоит исследовать влияние регуляризации на работу различных методов градиентного спуска. Напомним, регуляризация - это добавка к функции потерь, которая штрафует за норму весов. Мы будем использовать l2 регуляризацию

Допишите класс BaseDescentReg в файле descents.py.

Протестируйте ваше решение в контесте.

Вставьте ссылку на успешную посылку:

BaseDescentReg:
Найдите лучшие параметры обучения с регуляризацией аналогично 5 заданию. Будем подбирать длину шага
(lambda_) и коэффициент регуляризации
(mu).

Сравните для каждого метода результаты с регуляризацией и без регуляризации (нужно опять сохранить ошибку и качество по метрике
на обучающей и тестовой выборках и количество итераций до сходимости).

Постройте для каждого метода график со значениями функции потерь MSE с регуляризацией и без регуляризации (всего должно получиться 4 графика).

Посмотрите на получившиеся результаты. Какие можно сделать выводы, как регуляризация влияет на сходимость? Как изменилось качество на обучающей выборке? На тестовой? Чем вы можете объяснить это?

2. Работа с интерактивным блокнотом

В этом задании вы:

ознакомитесь с тем, что происходит "внутри" метода опорных векторов и логистической регрессии
познакомитесь с калибровкой вероятности
изучите методы трансформации переменных и методы отбора признаков
попробуете оценить экономический эффект модели

Скачайте интерактивный блокнот с заданиями. В пустые ячейки впишите свой программный код

Для сдачи задания переименуйте получившийся файл *.ipynb в соответствии со следующим форматом: homework-practice-04-linclass-Username.ipynb, где Username — ваша фамилия и имя на латинице именно в таком порядке (например, homework-practice-04-linclass-IvanovIvan.ipynb).

Раздел 3. Ансамблевые методы

Форма контроля/оценочное средство: Расчетно-графическая работа

Вопросы/Задания:

1. Необходимо воспользоваться возможностями бутстрапа для оценки смещения и разброса алгоритмов машинного обучения.

В этом задании вам предстоит воспользоваться возможностями бутстрапа для оценки смещения и разброса алгоритмов машинного обучения. Делать мы это будем на данных boston:

from sklearn.datasets import load_boston
boston = load_boston()

Сгенерировать
выборок
методом бутстрапа.
На каждой выборке
обучить алгоритм
.
Для каждой выборки
определить множество объектов
, не вошедших в нее (out-of-bag). Вычислить предсказания алгоритма
на объектах
.
Поскольку у нас есть только один ответ для каждого объекта, мы будем считать шум равным 0, а
равным имеющемуся правильному ответу для объекта
.

Итоговые оценки:

Смещение: для одного объекта - квадрат разности среднего предсказания и правильного ответа. Среднее предсказание берется только по тем алгоритмам
, для которых этот объект входил в out-of-bag выборку
. Для получения общего смещения выполнить усреденение смещений по объектам.
Разброс: для одного объекта - выборочная дисперсия предсказаний алгоритмов
, для которых этот объект входил в out-of-bag выборку
. Для получения общего разброса выполнить усреденение разбросов по объектам.
Ошибка
: усреднить квадраты разностей предсказания и правильного ответа по всем выполненным предсказаниям для всех объектов.
В результате должно получиться, что ошибка приблизительно равна сумме смещения и разброса!

Раздел 4. Снижение размерности

Форма контроля/оценочное средство: Задача

Вопросы/Задания:

1. В задании необходимо решить предложенные задачи по программированию – вписать свой код в ячейки после условий задач вместо комментария ### YOUR CODE HERE ### и сохранить изменения. Для решения использовать приложенный файл.

В задании необходимо решить предложенные задачи по программированию – вписать свой код в ячейки после условий задач вместо комментария ### YOUR CODE HERE ### и сохранить изменения. Для решения использовать приложенный файл.

Задание начального уровня
Задание просто уровня Дана матрица A , вычислите SVD разложение с помощью функции numpy.linalg.svd. Найдите определитель матрицы U с помощью функции numpy.linalg.det. Чему равен определитель?

Задание высокого уровня
Необходимо поработать с изображениями (похожий трюк можно увидеть в этой статье на хабре). Для этого нужно установить модуль Pillow (если у вас Анаконда, то https://anaconda.org/anaconda/pillow)

7. Оценочные материалы промежуточной аттестации

Седьмой семестр, Зачет
Контролируемые ИДК: ОПК-5.1 ОПК-5.2 ОПК-5.3 ОПК-5.4 ОПК-5.5 ПК-П12.1 ПК-П12.2 ПК-П12.3
Вопросы/Задания:

1. Основные этапы решения задач машинного обучения

Основные этапы решения задач машинного обучения

2. Бинаризация числовых признаков

Бинаризация числовых признаков

3. Измерение ошибки в задачах регрессии

Измерение ошибки в задачах регрессии

4. Переобучение линейной регрессии

Переобучение линейной регрессии

5. Оценивание качества линейной регрессии

Оценивание качества линейной регрессии

6. Обучение линейной регрессии

Обучение линейной регрессии

7. Градиентный спуск и оценивание градиента

Градиентный спуск и оценивание градиента

8. Стохастический градиентный спуск

Стохастический градиентный спуск

9. Модификации градиентного спуска

Модификации градиентного спуска

10. Регуляризация

Регуляризация

11. Преобразование признаков

Преобразование признаков

12. Линейные модели классификации

Линейные модели классификации

13. Метрики качества классифик

Метрики качества классифик

14. Логистическая регрессия

Логистическая регрессия

15. Метод опорных векторов

Метод опорных векторов

16. Многоклассовая классификация

Многоклассовая классификация

17. Решающие деревья

Решающие деревья

18. Случайные леса

Случайные леса

19. Бэггинг

Бэггинг

20. Бустинг

Бустинг

21. Стекинг

Стекинг

22. Градиентный бустинг

Градиентный бустинг

23. Понижение размерности

Понижение размерности

24. Кластеризация

Кластеризация

8. Материально-техническое и учебно-методическое обеспечение дисциплины

8.1. Перечень основной и дополнительной учебной литературы

Основная литература

1. Сапрыкин О. Н. Интеллектуальный анализ данных / Сапрыкин О. Н.. - Самара: Самарский университет, 2020. - 80 с. - 978-5-7883-1563-8. - Текст: электронный. // RuSpLAN: [сайт]. - URL: https://e.lanbook.com/img/cover/book/188906.jpg (дата обращения: 21.02.2024). - Режим доступа: по подписке

Дополнительная литература

1. Цзэн, М. Как Alibaba использует искусственный интеллект в бизнесе: Сетевое взаимодействие и анализ данных: Практическое пособие / М. Цзэн; Школа предпринимательства Хуань. - Москва: ООО "Альпина Паблишер", 2022. - 360 с. - 978-5-9614-3322-7. - Текст: электронный. // Общество с ограниченной ответственностью «ЗНАНИУМ»: [сайт]. - URL: https://znanium.com/cover/1905/1905832.jpg (дата обращения: 20.02.2024). - Режим доступа: по подписке

8.2. Профессиональные базы данных и ресурсы «Интернет», к которым обеспечивается доступ обучающихся

Профессиональные базы данных

Не используются.

Ресурсы «Интернет»

1. http://www.iprbookshop.ru/ - IPRbook
2. https://edu.kubsau.ru/ - Образовательный портал КубГАУ
3. https://znanium.com/ - Znanium.com

8.3. Программное обеспечение и информационно-справочные системы, используемые при осуществлении образовательного процесса по дисциплине

Информационные технологии, используемые при осуществлении образовательного процесса по дисциплине позволяют:

– обеспечить взаимодействие между участниками образовательного процесса, в том числе синхронное и (или) асинхронное взаимодействие посредством сети «Интернет»;
– фиксировать ход образовательного процесса, результатов промежуточной аттестации по дисциплине и результатов освоения образовательной программы;
– организовать процесс образования путем визуализации изучаемой информации посредством использования презентаций, учебных фильмов;
– контролировать результаты обучения на основе компьютерного тестирования.
Перечень лицензионного программного обеспечения:
1 Microsoft Windows - операционная система.
2 Microsoft Office (включает Word, Excel, Power Point) - пакет офисных приложений.
Перечень профессиональных баз данных и информационных справочных систем:
1 Гарант - правовая, https://www.garant.ru/
2 Консультант - правовая, https://www.consultant.ru/
3 Научная электронная библиотека eLibrary - универсальная, https://elibrary.ru/
Доступ к сети Интернет, доступ в электронную информационно-образовательную среду университета.

Перечень программного обеспечения
(обновление производится по мере появления новых версий программы)
Не используется.

Перечень информационно-справочных систем
(обновление выполняется еженедельно)
Не используется.

8.4. Специальные помещения, лаборатории и лабораторное оборудование

Компьютерный класс
224гл
Интерактивная панель Samsung - 1 шт.
Компьютер персональный DELL 3050 i3/4Gb/500Gb/21.5" - 1 шт.
Компьютер персональный iRU Corp 312 MT - 1 шт.
Сплит-система LS-H12KPA2/LU-H12KPA2 - 1 шт.

9. Методические указания по освоению дисциплины (модуля)

Учебная работа по направлению подготовки осуществляется в форме контактной работы с преподавателем, самостоятельной работы обучающегося, текущей и промежуточной аттестаций, иных формах, предлагаемых университетом. Учебный материал дисциплины структурирован и его изучение производится в тематической последовательности. Содержание методических указаний должно соответствовать требованиям Федерального государственного образовательного стандарта и учебных программ по дисциплине. Самостоятельная работа студентов может быть выполнена с помощью материалов, размещенных на портале поддержки Moodl.

Методические указания по формам работы

Лекционные занятия
Передача значительного объема систематизированной информации в устной форме достаточно большой аудитории. Дает возможность экономно и систематично излагать учебный материал. Обучающиеся изучают лекционный материал, размещенный на портале поддержки обучения Moodl.

Лабораторные занятия

Практическое освоение студентами научно-теоретических положений изучаемого предмета, овладение ими техникой экспериментирования в соответствующей отрасли науки. Лабораторные занятия проводятся с использованием методических указаний, размещенных на образовательном портале университета.

Описание возможностей изучения дисциплины лицами с ОВЗ и инвалидами

Для инвалидов и лиц с ОВЗ может изменяться объём дисциплины (модуля) в часах, выделенных на контактную работу обучающегося с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающегося (при этом не увеличивается количество зачётных единиц, выделенных на освоение дисциплины).

Фонды оценочных средств адаптируются к ограничениям здоровья и восприятия информации обучающимися.

Основные формы представления оценочных средств – в печатной форме или в форме электронного документа.

Формы контроля и оценки результатов обучения инвалидов и лиц с ОВЗ с нарушением зрения:

– устная проверка: дискуссии, тренинги, круглые столы, собеседования, устные коллоквиумы и др.;

– с использованием компьютера и специального ПО: работа с электронными образовательными ресурсами, тестирование, рефераты, курсовые проекты, дистанционные формы, если позволяет острота зрения - графические работы и др.;

– при возможности письменная проверка с использованием рельефно-точечной системы Брайля, увеличенного шрифта, использование специальных технических средств (тифлотехнических средств): контрольные, графические работы, тестирование, домашние задания, эссе, отчеты и др.

Формы контроля и оценки результатов обучения инвалидов и лиц с ОВЗ с нарушением слуха:

– письменная проверка: контрольные, графические работы, тестирование, домашние задания, эссе, письменные коллоквиумы, отчеты и др.;

– с использованием компьютера: работа с электронными образовательными ресурсами, тестирование, рефераты, курсовые проекты, графические работы, дистанционные формы и др.;

– при возможности устная проверка с использованием специальных технических средств (аудиосредств, средств коммуникации, звукоусиливающей аппаратуры и др.): дискуссии, тренинги, круглые столы, собеседования, устные коллоквиумы и др.

Формы контроля и оценки результатов обучения инвалидов и лиц с ОВЗ с нарушением опорно-двигательного аппарата:

– письменная проверка с использованием специальных технических средств (альтернативных средств ввода, управления компьютером и др.): контрольные, графические работы, тестирование, домашние задания, эссе, письменные коллоквиумы, отчеты и др.;

– устная проверка, с использованием специальных технических средств (средств коммуникаций): дискуссии, тренинги, круглые столы, собеседования, устные коллоквиумы и др.;

– с использованием компьютера и специального ПО (альтернативных средств ввода и управления компьютером и др.): работа с электронными образовательными ресурсами, тестирование, рефераты, курсовые проекты, графические работы, дистанционные формы предпочтительнее обучающимся, ограниченным в передвижении и др.

Адаптация процедуры проведения промежуточной аттестации для инвалидов и лиц с ОВЗ.

В ходе проведения промежуточной аттестации предусмотрено:

– предъявление обучающимся печатных и (или) электронных материалов в формах, адаптированных к ограничениям их здоровья;

– возможность пользоваться индивидуальными устройствами и средствами, позволяющими адаптировать материалы, осуществлять приём и передачу информации с учетом их

индивидуальных особенностей;
– увеличение продолжительности проведения аттестации;
– возможность присутствия ассистента и оказания им необходимой помощи (занять рабочее место, передвигаться, прочитать и оформить задание, общаться с преподавателем).
Формы промежуточной аттестации для инвалидов и лиц с ОВЗ должны учитывать индивидуальные и психофизические особенности обучающегося/обучающихся по АОПОП ВО (устно, письменно на бумаге, письменно на компьютере, в форме тестирования и т.п.).
Специальные условия, обеспечиваемые в процессе преподавания дисциплины студентам с нарушениями зрения:
– предоставление образовательного контента в текстовом электронном формате, позволяющем переводить плоскопечатную информацию в аудиальную или тактильную форму;
– возможность использовать индивидуальные устройства и средства, позволяющие адаптировать материалы, осуществлять приём и передачу информации с учетом индивидуальных особенностей и состояния здоровья студента;
– предоставление возможности предкурсового ознакомления с содержанием учебной дисциплины и материалом по курсу за счёт размещения информации на корпоративном образовательном портале;
– использование чёткого и увеличенного по размеру шрифта и графических объектов в мультимедийных презентациях;
– использование инструментов «лупа», «прожектор» при работе с интерактивной доской;
– озвучивание визуальной информации, представленной обучающимся в ходе занятий;
– обеспечение раздаточным материалом, дублирующим информацию, выводимую на экран;
– наличие подписей и описания у всех используемых в процессе обучения рисунков и иных графических объектов, что даёт возможность перевести письменный текст в аудиальный;
– обеспечение особого речевого режима преподавания: лекции читаются громко, разборчиво, отчётливо, с паузами между смысловыми блоками информации, обеспечивается интонирование, повторение, акцентирование, профилактика рассеивания внимания;
– минимизация внешнего шума и обеспечение спокойной аудиальной обстановки;
– возможность вести запись учебной информации студентами в удобной для них форме (аудиально, аудиовизуально, на ноутбуке, в виде пометок в заранее подготовленном тексте);
– увеличение доли методов социальной стимуляции (обращение внимания, апелляция к ограничениям по времени, контактные виды работ, групповые задания и др.) на практических и лабораторных занятиях;
– минимизирование заданий, требующих активного использования зрительной памяти и зрительного внимания;
– применение поэтапной системы контроля, более частый контроль выполнения заданий для самостоятельной работы.
Специальные условия, обеспечиваемые в процессе преподавания дисциплины студентам с нарушениями опорно-двигательного аппарата (маломобильные студенты, студенты, имеющие трудности передвижения и патологию верхних конечностей):
– возможность использовать специальное программное обеспечение и специальное оборудование и позволяющее компенсировать двигательное нарушение (коляски, ходунки, трости и др.);
– предоставление возможности предкурсового ознакомления с содержанием учебной дисциплины и материалом по курсу за счёт размещения информации на корпоративном образовательном портале;
– применение дополнительных средств активизации процессов запоминания и повторения;
– опора на определенные и точные понятия;
– использование для иллюстрации конкретных примеров;
– применение вопросов для мониторинга понимания;
– разделение изучаемого материала на небольшие логические блоки;
– увеличение доли конкретного материала и соблюдение принципа от простого к сложному при объяснении материала;
– наличие чёткой системы и алгоритма организации самостоятельных работ и проверки заданий с обязательной корректировкой и комментариями;

– увеличение доли методов социальной стимуляции (обращение внимания, аппеляция к ограничениям по времени, контактные виды работ, групповые задания др.);
– обеспечение беспрепятственного доступа в помещения, а также пребывания них;
– наличие возможности использовать индивидуальные устройства и средства, позволяющие обеспечить реализацию эргономических принципов и комфортное пребывание на месте в течение всего периода учёбы (подставки, специальные подушки и др.).
Специальные условия, обеспечиваемые в процессе преподавания дисциплины студентам с нарушениями слуха (глухие, слабослышащие, позднооглохшие):
– предоставление образовательного контента в текстовом электронном формате, позволяющем переводить аудиальную форму лекции в плоскопечатную информацию;
– наличие возможности использовать индивидуальные звукоусиливающие устройства и сурдотехнические средства, позволяющие осуществлять приём и передачу информации; осуществлять взаимообратный перевод текстовых и аудиофайлов (блокнот для речевого ввода), а также запись и воспроизведение зрительной информации;
– наличие системы заданий, обеспечивающих систематизацию вербального материала, его схематизацию, перевод в таблицы, схемы, опорные тексты, глоссарий;
– наличие наглядного сопровождения изучаемого материала (структурно-логические схемы, таблицы, графики, концентрирующие и обобщающие информацию, опорные конспекты, раздаточный материал);
– наличие чёткой системы и алгоритма организации самостоятельных работ и проверки заданий с обязательной корректировкой и комментариями;
– обеспечение практики опережающего чтения, когда студенты заранее знакомятся с материалом и выделяют незнакомые и непонятные слова и фрагменты;
– особый речевой режим работы (отказ от длинных фраз и сложных предложений, хорошая артикуляция; четкость изложения, отсутствие лишних слов; повторение фраз без изменения слов и порядка их следования; обеспечение зрительного контакта во время говорения и чуть более медленного темпа речи, использование естественных жестов и мимики);
– чёткое соблюдение алгоритма занятия и заданий для самостоятельной работы (называние темы, постановка цели, сообщение и запись плана, выделение основных понятий и методов их изучения, указание видов деятельности студентов и способов проверки усвоения материала, словарная работа);
– соблюдение требований к предъявляемым учебным текстам (разбивка текста на час¬ти; выделение опорных смысловых пунктов; использование наглядных средств);
– минимизация внешних шумов;
– предоставление возможности соотносить вербальный и графический материал; комплексное использование письменных и устных средств коммуникации при работе в группе;
– сочетание на занятиях всех видов речевой деятельности (говорения, слушания, чтения, письма, зрительного восприятия с лица говорящего).
Специальные условия, обеспечиваемые в процессе преподавания дисциплины студентам с прочими видами нарушений (ДЦП с нарушениями речи, заболевания эндокринной, центральной нервной и сердечно-сосудистой систем, онкологические заболевания):
– наличие возможности использовать индивидуальные устройства и средства, позволяющие осуществлять приём и передачу информации;
– наличие системы заданий, обеспечивающих систематизацию вербального материала, его схематизацию, перевод в таблицы, схемы, опорные тексты, глоссарий;
– наличие наглядного сопровождения изучаемого материала;
– наличие чёткой системы и алгоритма организации самостоятельных работ и проверки заданий с обязательной корректировкой и комментариями;
– обеспечение практики опережающего чтения, когда студенты заранее знакомятся с материалом и выделяют незнакомые и непонятные слова и фрагменты;
– предоставление возможности соотносить вербальный и графический материал; комплексное использование письменных и устных средств коммуникации при работе в группе;
– сочетание на занятиях всех видов речевой деятельности (говорения, слушания, чтения, письма, зрительного восприятия с лица говорящего);
– предоставление образовательного контента в текстовом электронном формате;

– предоставление возможности предкурсового ознакомления с содержанием учебной дисциплины и материалом по курсу за счёт размещения информации на корпоративном образовательном портале;
– возможность вести запись учебной информации студентами в удобной для них форме (аудиально, аудиовизуально, в виде пометок в заранее подготовленном тексте);
– применение поэтапной системы контроля, более частый контроль выполнения заданий для самостоятельной работы;
– стимулирование выработки у студентов навыков самоорганизации и самоконтроля;
– наличие пауз для отдыха и смены видов деятельности по ходу занятия.

10. Методические рекомендации по освоению дисциплины (модуля)