Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»

УТВЕРЖДАЮ

Директор по образовательной деятельности

_____ С.Т. Князев

«7» сентября 2023 г.

Машинное обучение

Учебно-методические материалы по направлению подготовки
09.03.03 Прикладная информатика
Образовательная программа «Прикладной искусственный интеллект»

Екатеринбург

РАЗРАБОТЧИКИ УЧЕБНО-МЕТОДИЧЕСКИХ МАТЕРИАЛОВ

Доцент кафедры радиоэлектроники и телекоммуникаций		Долганов Антон Юрьевич

СОДЕРЖАНИЕ

Раздел 1. Материалы лекций

1. Данные

Список тем, которые должны быть обсуждены на лекции:

1. Типы данных
2. Базы данных

Ключевые моменты по темам:

1. Обсуждение подходов к классификации типов данных

– Категориальные данные (номинальные и порядковые)

– Числовые данные (дискретные и непрерывные)

– Табличные данные

– Изображения

– Временные ряды

– Естественный язык

Рассмотрение примера табличных данных и определение типа данных (дискуссия).

Обсуждение возможности сведения «других» типов данных к табличным, а также недостатков этого подхода для каждого типа данных.

2. Рассмотрение открытых баз данных для задач машинного обучения. «Классические» базы данных (UCR, UCI) и современные платформы (kaggle, openml и др)

Подведение итогов лекции

Ссылки на справочные материалы

1. Платформа используется для обмена открытыми наборами данных, участия в соревнованиях по машинному обучению или обмена кодом в среде Data Science https://www.kaggle.com/
2. Платформа открытого машинного обучения https://www.openml.org
3. Классический репозиторий данных для машинного обучения https://archive.ics.uci.edu/ml/index.php

Примерные вопросы для контроля

1. Приведите несколько примеров непрерывных и дискретных данных (вопрос-дискуссия)
2. К какому типу данных можно отнести диагноз, поставленный врачом? (категориальные данные)
3. Опишите, какие данные содержатся в наборе данных Iris (https://archive.ics.uci.edu/ml/datasets/Iris)

4. Что делает изображения, тексты на естественном языке и временные ряды особым типом данных? (вопрос-дискуссия)

2. Линейная алгебра

Список тем, которые должны быть обсуждены на лекции:

1. Объекты
2. Операции

Ключевые моменты по темам:

1. Обсуждение основных объектов линейной алгебры

– Скаляр
– Вектор
– Матрица
– Тензор

Обозначения основных объектов. Размерности объектов. Рассмотрение примеров объектов линейной алгебры (дискуссия)

2. Основные операции линейной алгебры

Сложение матриц и умножение матриц. Акцент внимания на отслеживании размерностей объектов при выполнении операций.

Транспонирование матриц

Поиск обратной матрицы

Геометрический смысл матриц. Собственные значения и собственные вектора матриц. Определитель матрицы

Подведение итогов лекции

Ссылки на справочные материалы

1. Плейлист с хорошей визуализацией понятий и концепций Линейной Алгебры (на английском) https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab
2. Сайт с интерактивной визуализацией матричного умножения http://matrixmultiplication.xyz/

Примерные вопросы для контроля

1. Приведите несколько примеров данных в виде тензоров (вопрос-дискуссия)
2. У вас есть три матрицы A, B, C: A имеет размеры 5 × 4, B имеет размеры 4 × 6, C имеет размеры 3 × 5. Напишите все возможные матрицы, которые можно перемножить между собой, и укажите размеры результирующих матриц

A*B = AB (5x6)

C*A = CA (3x4)

CA*B = CAB (3x6)

3. Предварительная обработка

Список тем, которые должны быть осуждены на лекции:

1. Начальные Шаги Предварительной Обработки
2. Типы Предварительной Обработки
3. Базовая Генерация Признаков

Ключевые моменты по темам:

1. Рассмотрение возможностей библиотеки Pandas для анализа данных. Объект датафрейм (dataframe).

Методы библиотеки Pandas для поиска пропусков (.isna()), заполнения пропусков (.fillna(value)), поиска дубликатов (.duplicated()), удаления дубликатов (.drop_duplicates()).

Применение агрегации для первичного анализа данных (.groupby())

Особенности применения библиотеки seaborn для визуализации датафреймов Pandas

2. Основные типы предварительной обработки данных

Мотивация для предварительной обработки данных.

Стандартизация и нормализация для предварительной обработки данных (линейные преобразования, которые не изменяют распределение данных)

Анализ выбросов при предварительной обработке

Степенное преобразование (нелинейное преобразование для получения нормального распределения)

3. Простые подходы к генерации дополнительных признаков.

One-hot encoding для использования категориальных признаков в линейных моделях. Метод .get_dummies библиотеки Pandas для реализации One-hot encoding.

Анализ распределения категориальных признаков. Редкие категории.

Категориальное сопоставление. Особенности методов .cut() и .qcut() библиотеки Pandas для разбивки значений на дискретные интервалы и дискретизации на основе квантилей.

Комбинации категориальных признаков.

Обсуждение возможных источников для новых признаков:

Знания, специфичные для предметной области (чтение статей, взаимодействие с научным руководителем и/или экспертом в данной предметной области и т.д.)

Исследовательский анализ данных (статистический анализ данных, анализ распределений данных)

Подведение итогов лекции.

Ссылки на справочные материалы:

1. Документация библиотеки Pandas для работы с данными https://pandas.pydata.org/
2. Документация библиотеки Seaborn для визуализации данных https://seaborn.pydata.org/

Примерные вопросы для контроля:

1. Опишите разные ситуации, в которых вы будете использовать разные типы предварительной обработки данных (вопрос-дискуссия)
2. Как правильно использовать категориальные признаки в линейных моделях? (One-hot encoding)

4. Основы математического анализа

Список тем, которые должны быть осуждены на лекции:

1. Функции
2. Производные

Ключевые моменты по темам:

1. Определение функции в математике. Определение функции в программировании
   Способы представления функции: табличная форма, график, формула
   Примеры функций: линейные функции, нелинейные функции.
   Рассмотрение нейронных сетей как совокупность сложных функций
2. Определение производной
   Примеры производных ряда функций

Таблицы производных: производные степенных функций, экспоненты, логарифма. Правила взятия производных сложных функций: производная произведения, производная линейной комбинации функции, цепное правило

Градиент – вектор производных. Рассмотрение примеров градиентов разных функций

Рассмотрение примеров взятия производных сложной функции

Подведение итогов лекции.

Ссылки на справочные материалы:

1. Плейлист с хорошей визуализацией понятий и концепций Математического Анализа (на английском) https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr
2. Сводная информация о производной функции

https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%B8%D0%B7%D0%B2%D0%BE%D0%B4%D0%BD%D0%B0%D1%8F_%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D0%B8

5. История и базовые понятия

Список тем, которые должны быть обсуждены на лекции:

1. Что такое машинное обучение
2. История машинного обучения
3. Основные термины в машинном обучении
4. Классификация задач машинном обучении

Ключевые моменты по темам:

1. Обсуждение того, где можно встретить результаты машинного обучения (задать студентам вопросы вида «когда в последний раз вы сталкивались с результатами машинного обучения»)

Обсуждение определений «машинное обучение», «искусственный интеллект», «интеллект»

Сравнение подхода, основанного на традиционном программировании с машинным обучением

2. Историческая ретроспектива об искусственном интеллекте и машинном обучении.

   Обсуждение «взлетов и падений» подходов, основанных на машинном обучении.

   Текущее состояние искусственного интеллекта и машинного обучения
3. Обсуждение основных терминов, связанных с машинным обучением
   - Объект, Цель, Признаки
   - Тренировочная и Тестовая выборка
   - Модель, параметры и гиперпараметры модели
   - Функция потерь и цель обучения
   - Разложение ошибок модели на смещение и дисперсию. Переобучение
   - Валидация моделей. Отложенная выборка и кросс-валидация
4. Обсуждение типовых задач машинного обучения и ключевых различий между ними.

   Обучение с учителем: задачи регрессии и классификации

   Обучение без учителя: задачи кластеризации и уменьшения размерности

   Обучение с подкреплением

   Обсуждение примеров различных задач машинного обучения

   Подведение итогов лекции.

Ссылки на справочные материалы:

1. Блог с описанием ключевых терминов и понятий машинного обучения «простыми словами» https://vas3k.ru/blog/machine_learning/
2. Хронология ключевых событий, связанных с машинным обучением https://en.wikipedia.org/wiki/Timeline_of_machine_learning
3. Блог, в котором освещаются некоторые важные прикладные аспекты машинного обучения, с пониманием которых у практиков часто возникают трудности https://dyakonov.org/

Примерные вопросы для контроля:

1. Опишите, как вы поняли, что такое машинное обучение (вопрос-дискуссия)
2. Назовите ученого, который считается родоначальником машинного обучения и теории искусственного интеллекта (Алан Тьюринг)
3. Напишите название шахматного суперкомпьютера, который впервые смог обыграть Гарри Каспарова в матче из 6 партий (Deep Blue)
4. Опишите разницу между подходом машинного обучения и традиционным программированием (вопрос-дискуссия)
5. Опишите разницу между обучением с учителем и обучением без учителя (вопрос-дискуссия)
6. Опишите разницу между задачами классификации и задачами регрессии (вопрос-дискуссия)
7. Вас попросят создать программу, которая будет определять кошку или собаку по изображению. К какому типу задач машинного обучения относится эта просьба? (классификация)

6. Регрессия

Список тем, которые должны быть осуждены на лекции:

1. Регрессия
2. Метод Наименьших Квадратов
3. Градиентный Спуск
4. Регуляризация
5. Метрики Регрессии

Ключевые моменты по темам:

1. Задача регрессии. Обсуждение примеров задач из реальной жизни, которые можно свести к регрессии.
   Области применения задач регрессии (дискуссия)
2. Модель линейной регрессии

Метод наименьших квадратов для поиска коэффициентов линейной регрессии

Ограничение метода наименьших квадратов

3. Модель линейной регрессии

   Градиентный спуск для поиска коэффициентов линейной регрессии

4. Регуляризация

   Взаимосвязь величины коэффициентов линейной регрессии и высокой дисперсией

   L1 и L2 регуляризация

   SWOT-анализ линейной регрессии

5. Обсуждение метрик регрессии

   Подведение итогов лекции

Ссылки на справочные материалы

1. Блог с описанием ключевых терминов и понятий машинного обучения «простыми словами». *Блок про регрессию* https://vas3k.ru/blog/machine_learning/
2. Описание линейных моделей https://ml-handbook.ru/chapters/linear_models/intro#%D0%BB%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%B0%D1%8F-%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F-%D0%B8-%D0%BC%D0%B5%D1%82%D0%BE%D0%B4-%D0%BD%D0%B0%D0%B8%D0%BC%D0%B5%D0%BD%D1%8C%D1%88%D0%B8%D1%85-%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%BE%D0%B2-%D0%BC%D0%BD%D0%BA
3. Описание метрик регрессии https://ml-handbook.ru/chapters/model_evaluation/intro#%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F

Примерные вопросы для контроля:

1. Каковы ключевые части метода наименьших квадратов для поиска коэффициентов линейной регрессии? (вопрос-дискуссия)
2. В чем заключаются основные различия между методом наименьших квадратов и градиентным спуском для нахождения коэффициентов регрессии? (вопрос-дискуссия)

3. Может ли коэффициент детерминации быть отрицательным числом? Если «да» - в каких случаях, если «нет» - почему? (Может. Если модель предсказывает хуже, чем среднее значение по выборке)
4. Почему L1-регуляризация может привести к отбору признаков (в отличие от L2-регуляризации)? (вопрос-дискуссия)
5. Почему регуляризация помогает уменьшить переобучение? (вопрос-дискуссия)

7. Классификация

Список тем, которые должны быть осуждены на лекции:

1. Классификация
2. Логистическая регрессия
3. Метрики Классификации

Ключевые моменты по темам:

1. Задача классификации. Обсуждение примеров задач из реальной жизни, которые можно свести к классификации.

   Типы классов (бинарная и мультиклассовая классификация, пересекающиеся и непересекающиеся классы, нечеткие классы)

   Области применения задач классификации (дискуссия)
2. Логистическая регрессия

   Переход от модели линейной регрессии для решения задачи классификации

   Функция потерь логистической регрессии

   Градиентный спуск и регуляризация модели логистической регрессии

   Стратегии один против всех и один против одного для мультиклассовой классификации

   SWOT-анализ логистической регрессии
3. Обсуждение метрик классификации

   Подведение итогов лекции.

Ссылки на справочные материалы

1. Блог с описанием ключевых терминов и понятий машинного обучения «простыми словами». Блок про классификацию https://vas3k.ru/blog/machine_learning/
2. Описание линейных моделей https://ml-handbook.ru/chapters/linear_models/intro#%D0%BB%D0%BE%D0%B3%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F-%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F

3. Описание метрик классификации
   https://ml-handbook.ru/chapters/model_evaluation/intro#%D0%B1%D0%B8%D0%BD%D0%B0%D1%80%D0%BD%D0%B0%D1%8F-%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F-%D0%BC%D0%B5%D1%82%D0%BA%D0%B8-%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%BE%D0%B2

Примерные вопросы для контроля

1. В чем основное различие между задачами классификации и задачами регрессии? (вопрос-дискуссия)
2. В чем разница между применением градиентного спуска в линейной регрессии и логистической регрессии? (вопрос-дискуссия)
3. Приведите пару примеров задач бинарной классификации и мультиклассовой классификации. (вопрос-дискуссия)

8. Методы разложения матриц

Список тем, которые должны быть осуждены на лекции:

1. Уменьшение Размерности
2. Ковариационная Матрица
3. Метод Главных Компонент
4. Сингулярное разложение

Ключевые моменты по темам:

1. Задача уменьшения размерности. Обсуждение примеров, в которых пригодится уменьшение размерности.
   Подход к уменьшению размерности основанный на распределении данных. Необходимость стандартизации данных
2. Определение ковариации, матрицы ковариации.
   Расчет матрицы ковариации для стандартизованных данных
3. Метод главных компонент. Поиск собственных значений и собственных векторов матрицы ковариации
   Объясненная дисперсия, кумулятивная объясненная дисперсия – анализ собственных значений матрицы ковариации
   Главные компоненты – линейная комбинация исходных признаков, коэффициенты в этой комбинации – собственные вектора

Подходы к выбору уменьшенного пространства признаков (для визуализации, исходя из значений кумулятивной объясненной дисперсии, порога объясненной дисперсии)

Ключевые шаги метода главных компонент

4. Представление метода главных компонент как разложения квадратных матриц.

   Сингулярное разложение матриц, как обобщенный случай разложения неквадратных матриц.

   Матрицы U, S, V

   Варианты сингулярного разложения матриц: базовый вариант, экономный вариант, компактный вариант, усеченный вариант.

   Сжатие данных с использованием сингулярного разложения матриц

   Получение псеводообратной матрицы с использованием сингулярного разложения матриц

   SWOT-анализ метода главных компонент и сингулярного разложения матриц

   Подведение итогов лекции.

Ссылки на справочные материалы:

1. Блог с описанием ключевых терминов и понятий машинного обучения «простыми словами». Блок про уменьшение размерности (обобщение) https://vas3k.ru/blog/machine_learning/
2. Плейлист с хорошей визуализацией понятий и концепций Линейной Алгебры (на английском). Обратить внимание на видео про собственные значения и собственные вектора https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab
3. Описание метода главных компонент https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D1%8B%D1%85_%D0%BA%D0%BE%D0%BC%D0%BF%D0%BE%D0%BD%D0%B5%D0%BD%D1%82
4. Описание сингулярного разложение матриц https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D0%B7%D0%BB%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5

Примерные вопросы для контроля:

1. Напишите, как вы поняли, из-за чего происходит уменьшение размерности в методе главных компонент (вопрос-дискуссия)

2. Объясните, почему перед применением метода главных компонентов необходимо провести стандартизацию данных (вопрос-дискуссия)
3. Что означают собственные значения и собственные векторы ковариационной матрицы в методе главных компонент? (вопрос-дискуссия)
4. Как связаны главные компоненты с исходными данными?
5. Что означают матрицы U, S и V в сингулярном разложении?
6. В чем отличие между вариантами сингулярного разложения матриц: компактное, экономичное и усеченное?

9. Кластеризация

Список тем, которые должны быть осуждены на лекции:

1. Кластеризация
2. Расстояние
3. Метод к-Средних
4. Метрики Кластеризации

Ключевые моменты по темам:

1. Задача кластеризации. Обсуждение примеров задач из реальной жизни, которые можно свести к кластеризации.
   Области применения задач кластеризации (дискуссия)
2. Определение расстояния. Особенности различных метрик расстояния
   - Евклидово расстояние
   - Манхэттенское расстояние
   - Расстояние Чебышева
   - Расстояние Минковского

   Эквидистантные области при использовании разных метрик расстояния
3. Метод кластеризации к-Средних (k-Means)
   Описание ключевых шагов алгоритма кластеризации к-Средних
   Демонстрация работы алгоритма кластеризации к-Средних
   Метод локтя для определения оптимального числа к
   SWOT-анализ кластеризации методом к-Средних
   Обсуждение ситуаций, когда алгоритм к-Средних будет выдавать заведомо неправильные ответы
4. Обсуждение метрик кластеризации.
   Описание метода силуэтов для оценки кластеризации
   Подведение итогов лекции

Ссылки на справочные материалы:

1. Блог с описанием ключевых терминов и понятий машинного обучения «простыми словами». Блок про кластеризацию https://vas3k.ru/blog/machine_learning/
2. Описание кластеризации методом к-средних https://en.wikipedia.org/wiki/K-means_clustering
3. Описание метрик кластеризации https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B0_%D0%B2_%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B5_%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8

Примерные вопросы для контроля:

1. Какая связь между евклидовым расстоянием и расстоянием Минковского? (вопрос-дискуссия)
2. Что является основным гиперпараметром алгоритма кластеризации k-средних? (к – число кластеров)
3. Может ли коэффициент силуэта быть равным отрицательному числу? Если «да» - в каких случаях, если «нет» - почему? (Может. Если «свой кластер» занимает большое место, а ближайший чужой – компактный и близкий)
4. Как можно использовать метод локтя для определения оптимального количества кластеров? (вопрос-дискуссия)

10. Библиотека scikit-learn

Список тем, которые должны быть осуждены на лекции:

1. Предварительная обработка
2. Разложение матриц
3. Кластеризация
4. Линейная регрессия
5. Логистическая регрессия

Ключевые моменты по темам:

1. Пакет sklearn.preprocessing для предварительной обработки данных
   Общая структура использования методов .fit(), .transform(), .inverse_transfrorm()
   Стандартизация (StandardScaler), нормализация (MinMaxScaler), Степенное преобразование (PowerTransformer)
2. Пакет sklearn.decomposition для разложения матриц

Метод главных компонент (PCA)

3. Пакет sklearn.cluster для кластеризации

   Кластеризация к-Средних (KMeans). Общая структура использования методов .fit(), .predict()

   Метрики кластеризации

4. Пакет sklearn.linear_model для линейных моделей

   Линейная регрессия (LinearRegression). Объект полиномиальные признаки (PolynomialFeatures)

   Пакет sklearn.pipeline

   Регуляризация линейной регрессии (Lasso и Ridge)

   Метрики регрессии

   Модуль sklearn.model_selection для выбора моделей.

   Перекрестная проверка (cross_validate, ShuffleSplit)

5. Логистическая регрессия (LogisticRegression)

   Метрики классификации

   Перекрестная проверка (cross_validate, StratifiedKFold)

   Подведение итогов лекции

Ссылки на справочные материалы:

1. Документация библиотеки scikit-learn для предварительной обработки данных https://scikit-learn.org/stable/modules/preprocessing.html
2. Документация библиотеки scikit-learn для метода главных компонент https://scikit-learn.org/stable/modules/decomposition.html#pca
3. Документация библиотеки scikit-learn для кластеризации к-средних https://scikit-learn.org/stable/modules/clustering.html#k-means
4. Документация библиотеки scikit-learn для линейных моделей https://scikit-learn.org/stable/modules/linear_model.html
5. Документация библиотеки scikit-learn для перекрестной проверки https://scikit-learn.org/stable/modules/cross_validation.html
6. Документация библиотеки scikit-learn для составных моделей https://scikit-learn.org/stable/modules/compose.html

Раздел 2. Домашние работы

1. Практика по теме «Данные»

Демонстрационные блокноты:

Ссылка на блокнот по **основам работы с Наборами Данных**

https://colab.research.google.com/drive/1IrjJEnResbR2fZecaCU7HFrfUFcdyDE6?usp=sharing

Ссылка на Блокнот по **предварительной обработке Данных**

https://colab.research.google.com/drive/1q3HsjCPgl9gXwiNxI8OTkpPId8a5QcPD?usp=sharing

Ссылка на Блокнот с **Интерактивной Визуализацией**

https://colab.research.google.com/drive/1OcYvCADVed4PybnreJRjHxZxxXfQ6QHx?usp=sharing

Задание:

1. Ознакомьтесь с содержанием демонстрационных блокнотов

2. Создайте новый блокнот, импортируйте необходимые библиотеки

3. Выполните следующие блоки заданий (каждый блок рекомендуется выполнять в отдельном блокноте)

I. Основы работы с наборами данных

- Найдите и загрузите несколько интересных наборов данных из OpenML (или любой другой сайт с данными. Можете "что-то свое"). Это пригодится на будущее — вам понадобится набор данных для регрессии, классификации, кластеризации и уменьшения размерности.
- Упаковать набор данных в ДатаФрейм pandas с Именованными столбцами
- Выполните Расчет статистик (в т.ч. с использованием аггрегации)
- Нарисуйте как минимум 3 разные графика (по крайней мере, на одном графике вам нужно сделать цвет или размер маркеров на основе целевого класса / значений)

II. Предварительная обработка данных

- Для выбранных наборов данных из блока I выберите параметры, которые необходимо предварительно обработать.
- Выполните адекватную предварительную обработку данных
- Проанализируйте результат

Вы должны загрузить ipynb вашего решения или ссылку (если дана ссылка, вам нужно убедиться, что режим доступа открыт)

2. Практика по теме «Линейная регрессия»

Демонстрационные Блокноты

Ссылка на блокнот с **Линейной Регрессией (генерируемые данные) + Полиномы**

https://colab.research.google.com/drive/1PdSFGsfSkgbKf0820uMldLYQQBjye2FP?usp=sharing

Ссылка на Блокнот по **Линейной Регрессией (генерируемые данные) + Регуляризация**

https://colab.research.google.com/drive/1zK7TVzRtQyuTTqDmQLoMgSzFqYcbIVqq?usp=sharing

Ссылка на Блокнот с **Линейной Регрессией на Реальных данных**

https://colab.research.google.com/drive/1VyzvaOYKeckKdI8ISYNGn3zvzBeV9ilq?usp=sharing

Задание:

1. Ознакомьтесь с содержанием демонстрационных блокнотов

2. Создайте новый блокнот, импортируйте необходимые библиотеки

3. Выполните следующие блоки заданий (каждый блок рекомендуется выполнять в отдельном блокноте)

I. Линейная Регрессия

- Выберите набор данных регрессии из OpenML для анализа
- Выполните регрессию с помощью разных подходов:
  * Вы можете использовать простую линейную модель
  * Вы можете использовать только регуляризацию
  * Вы можете комбинировать регуляризацию и полиномиальные параметры
- Оцените метрики регрессии с помощью перекрестной проверки
- Визуализируйте результаты (веса, предсказания, и т.п.)

Вы должны загрузить ipynb вашего решения или ссылку (если дана ссылка, вам нужно убедиться, что режим доступа открыт)

3. Практика по теме «Логистическая регрессия»

Демонстрационные блокноты:

Ссылка на блокнот **Логистическая Регрессия (генерируемые данные)**

https://colab.research.google.com/drive/1_jXjYT9lranyYP66Fo1VF8y8It2CvMVD?usp=sharing

Ссылка на Блокнот **Логистическая Регрессия (данные Ирисы)**

https://colab.research.google.com/drive/1JcGufcslWwJ6VpYhb8PkauOhG89grQhu?usp=sharing

Задание:

1. Ознакомьтесь с содержанием демонстрационных блокнотов

2. Создайте новый блокнот, импортируйте необходимые библиотеки

3. Выполните следующие блоки заданий (каждый блок рекомендуется выполнять в отдельном блокноте)

I. Логистическая Регрессия

- Выберите набор данных классификации из OpenML для анализа (предпочтительна бинарная классификация)

* Вы можете использовать методы уменьшения размерности

- Выполните классификацию с разными подходами

* Вы можете использовать оригинальные параметры

* Вы можете использовать полиномиальные параметры

* Вы можете использовать параметры после применения уменьшения размерности

- Оцените показатели классификации с помощью перекрестной проверки и матрицы ошибок
- Визуализируйте результаты

Вы должны загрузить ipynb вашего решения или ссылку (если дана ссылка, вам нужно убедиться, что режим доступа открыт)

4. Практика по теме «Матричное разложение»

Демонстрационные блокноты:

Ссылка на блокнот **PCA (генерируемые данные)**

https://colab.research.google.com/drive/1y9IxMm1UMx0pAyGf897xISKfLeqYzvFE?usp=sharing

Ссылка на Блокнот **PCA (данные Ирисы)**

https://colab.research.google.com/drive/10HUyk0RZQCrsEmKYvncu8c9xa_ROUaHp?usp=sharing

Ссылка на Блокнот **PCA (данные MNIST)**

https://colab.research.google.com/drive/15szVgQnYcUJ99mj3aywzlFhSmUu52OFp?usp=sharing

Ссылка на Блокнот **SVD (данные MNIST)**

https://colab.research.google.com/drive/1ee365Zm-13jg09IWRkq9h3jHldPzTDV5?usp=sharing

Задание:

1. Ознакомьтесь с содержанием демонстрационных блокнотов

2. Создайте новый блокнот, импортируйте необходимые библиотеки

3. Выполните следующие блоки заданий (каждый блок рекомендуется выполнять в отдельном блокноте)

I. PCA

- Выберите набор данных с сайта OpenML (он должен иметь> 10 параметров, как минимум 2 класса и не слишком много образцов (менее 10000))
- Примените PCA
- Визуализируйте несколько различных главных компонент (вы можете использовать двухмерные или трехмерные графики и различные комбинации главных компонент, такие как pca-1 pca-2 pca-5; pca-2 pca-3 pca-4; pca-1 pca-5 pca-9)

- Визуализируйте веса, чтобы понять, что означают различные основные компоненты. Сделайте краткий анализ

II. SVD

- Для этой задачи используйте набор данных с изображениями по типу Olivetti_Faces. Это набор данных из 400 изображений лиц (10 изображений для 40 разных людей, изображения 64x64)
- Импортируйте этот набор данных с помощью функции fetch_openml. Визуализируйте несколько примеров
- Примените SVD
- Визуализировать

*матрицу VT

*различные проекции

*реконструкция для разного количества компонентов для разных примеров

Вы должны загрузить ipynb вашего решения или ссылку (если дана ссылка, вам нужно убедиться, что режим доступа открыт)

5. Практика по теме «Кластеризация»

Демонстрационные блокноты:

Ссылка на блокнот **Кластеризация К-Средних (генерируемые данные)**

https://colab.research.google.com/drive/1phWx6FdVRJ35WHY6pHaWIhU8pAnf-04b?usp=sharing

Ссылка на Блокнот **Кластеризация (данные Ирисы)**

https://colab.research.google.com/drive/1OFDiNyAZizh5YR0f3zpLjzG9MDrmsyAz?usp=sharing

Задание:

1. Ознакомьтесь с содержанием демонстрационных блокнотов

2. Создайте новый блокнот, импортируйте необходимые библиотеки

3. Выполните следующие блоки заданий (каждый метод кластеризации рекомендуется выполнять в отдельном блокноте)

I. Кластеризация

- Выберите набор данных для кластеризации или классификации из OpenML для анализа (это может быть 2-х или 3-мерный набор данных, или вы можете использовать только 2/3 оси для визуализации)

* Рекомендуется выполнить стандартизацию данных до применения кластеризации.

* Вы можете использовать уменьшение размерности (PCA) вместо исходных параметров

- Выполните кластеризацию с использованием различных подходов

* Настройте параметры различных подходов к кластеризации для получения лучших результатов

- Визуализируйте результаты

Вы должны загрузить ipynb вашего решения или ссылку (если дана ссылка, вам нужно убедиться, что режим доступа открыт)

6. Практика по теме «Деревья решений»

Демонстрационные блокноты:

Ссылка на блокнот с **Классификацией Деревьями Решений + данные Ирисы**

https://colab.research.google.com/drive/1MfMQ7OB3CVjEoMgguWvLNJ-8Bd9eR4wE?usp=sharing

Ссылка на блокнот с **Регрессией Деревьями Решений + данные Диабет**

https://colab.research.google.com/drive/1uOLpaKdBqXFiG1EkxMxeXP1H5qcNF5N6?usp=sharing

Задание:

1. Ознакомьтесь с содержанием демонстрационных блокнотов

2. Создайте новый блокнот, импортируйте необходимые библиотеки

3. Выполните следующие блоки заданий (каждый блок рекомендуется выполнять в отдельном блокноте)

I. Классификация Деревьями Решений

- Выберите набор данных классификации из OpenML для анализа
- Выполните классификацию Деревьями Решений. Выберите оптимальные гиперпараметры
- Сравните результаты с логистической регрессией (и прочими методами)

II. Регрессия Деревьями Решений

- Выберите набор данных регресии из OpenML для анализа
- Примените Регрессия Деревьями Решений. Выберите оптимальные гиперпараметры
- Сравните результаты с линейной регрессией (и прочими методами)

Вы должны загрузить ipynb вашего решения или ссылки (если дана ссылка, вам нужно убедиться, что режим доступа открыт)

7. Практика по теме «Ансамблевые методы»

Демонстрационные блокноты:

Ссылка на блокнот с **Классификацией Ансамблями + данные Ирисы**

https://colab.research.google.com/drive/1q15-Gfadc63z8YFsUUK8fFlAxTpptVLN?usp=sharing

Ссылка на блокнот с **Регрессией Ансамблями + данные Диабет**

https://colab.research.google.com/drive/10pcCRekGtlwp35GwiDNFN9phK-Gp6kR3?usp=sharing

Задание:

1. Ознакомьтесь с содержанием демонстрационных блокнотов

2. Создайте новый блокнот, импортируйте необходимые библиотеки

3. Выполните следующие блоки заданий (каждый блок рекомендуется выполнять в отдельном блокноте)

I. Классификация Ансамблями

- Выберите набор данных классификации из OpenML для анализа
- Выполните классификацию **Ансамблями** (выбирете как минимум два вида). Выберите оптимальные гиперпараметры
- Сравните результаты с логистической регрессией (и прочими методами)

II. Регрессия Ансамблями

- Выберите набор данных регресии из OpenML для анализа
- Примените Регрессию **Ансамблями** (выбирете как минимум два вида). Выберите оптимальные гиперпараметры
- Сравните результаты с линейной регрессией (и прочими методами)

Вы должны загрузить ipynb вашего решения или ссылки (если дана ссылка, вам нужно убедиться, что режим доступа открыт)