МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

ФАКУЛЬТЕТ ПРОГРАММНОЙ ИНЖЕНЕРИИ И КОМПЬЮТЕРНОЙ ТЕХНИКИ

**Отчёт по модулю №2**

по дисциплине

«Системы искусственного интеллекта»

***Выполнил:*** Студент группы P3334 Баянов Равиль Динарович

***Преподаватель:***

Авдюшина А. Е.

Санкт-Петербург, 2024

**Содержание**

[**Лабораторная работа 1. Метод линейной регрессии** 3](#_Toc180716898)

[**Введение** 3](#_Toc180716899)

[**Описание метода** 3](#_Toc180716900)

[**Псевдокод метода** 3](#_Toc180716901)

[**Результаты выполнения** 4](#_Toc180716902)

[**Примеры использования кода** 4](#_Toc180716903)

[**Лабораторная работа 2. Метод k-ближайших соседей (k-NN)** 6](#_Toc180716904)

[**Введение** 6](#_Toc180716905)

[**Описание метода** 6](#_Toc180716906)

[**Псевдокод метода** 6](#_Toc180716907)

[**Результаты выполнения** 6](#_Toc180716908)

[**Примеры использования кода** 10](#_Toc180716909)

[**Лабораторная работа 3. Деревья решений** 11](#_Toc180716910)

[**Введение** 11](#_Toc180716911)

[**Описание метода** 11](#_Toc180716912)

[**Псевдокод метода** 11](#_Toc180716913)

[**Результаты выполнения** 11](#_Toc180716914)

[**Примеры использования кода** 11](#_Toc180716915)

[**Лабораторная работа 4. Логистическая регрессия** 13](#_Toc180716916)

[**Введение** 13](#_Toc180716917)

[**Описание метода** 13](#_Toc180716918)

[**Псевдокод метода** 13](#_Toc180716919)

[**Результаты выполнения** 13](#_Toc180716920)

[**Примеры использования кода** 13](#_Toc180716921)

[**Сравнение методов** 14](#_Toc180716922)

[**Заключение** 16](#_Toc180716923)

**Лабораторная работа 1. Метод линейной регрессии**

## **Введение**

Данный метод ставит перед собой цель предсказания целевой переменной по одному или нескольким признакам. В нём мы пытаемся построить такую прямую, которая будет имитировать распределение точек (данных) на плоскости.

## **Описание метода**

Линейная регрессия – это метод МО, который мы используем для моделирования зависимости между целевой переменной и признаками. Идея заключается в том, что мы строим прямую, стремящуюся как можно лучше описать связь зависимых переменных и целевой переменной.

**Принцип работы:**

* Линейная регрессия предполагает линейную зависимость между входными переменными и результатом.
* Алгоритм находит коэффициенты модели таким образом, чтобы минимизировать ошибку (разницу между реальными и предсказанными значениями).
* Метрика качества часто измеряется через **MSE (среднеквадратичную ошибку)**.

## **Псевдокод метода**

|  |
| --- |
| # Псевдокод для линейной регрессии  1. Инициализировать коэффициенты (w) случайными значениями.  2. Для каждого наблюдения в данных:  - Вычислить предсказание: y\_pred = w0 + w1 \* X1 + ... + wn \* Xn  - Рассчитать ошибку: error = y\_actual - y\_pred  3. Обновить коэффициенты:  - w\_new = w\_old - learning\_rate \* производная\_ошибки\_по\_w  4. Повторять шаги 2-3 до тех пор, пока ошибка не станет минимальной или не достигнем заданного количества итераций.  5. Выдать коэффициенты модели. |

## **Результаты выполнения**

Мы знаем, что идеальный способ определить качество, построенной модели с помощью метода ЛР нам надо посмотреть на коэффициент детерминации. Коэффициент детерминации – это доля дисперсии ошибок модели. В нашей лабораторной работе мы получили такие результаты:

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Если коэффициент находится недалеко от 0.5, то в целом это неплохой показатель модели, так как данных, на которых мы обучали нашу модель, было не так уж и много. Главное, что коэффициент детерминации не стремится к нулю.

## **Примеры использования кода**

Метод линейной регрессии полезен в следующих ситуациях:

1. **Прогнозирование цен на жилье:** применяется для оценки зависимости цены от площади, количества комнат и других факторов. Линейная регрессия позволяет эффективно моделировать эту зависимость.
2. **Анализ продаж:** метод используется для предсказания объёма продаж на основе рекламных расходов, временных трендов и других факторов.
3. **Медицинские исследования:** врачи могут использовать линейную регрессию для определения влияния различных медицинских показателей на вероятность выздоровления пациента.

Метод выбран из-за его простоты, интерпретируемости и скорости обучения на больших наборах данных.

**Лабораторная работа 2. Метод k-ближайших соседей (k-NN)**

## **Введение**

Наша цель всё та же – нужно построить модель, которая будет связывать зависимую переменную с целевой переменной. Но теперь мы будем делать это методом k-ближайших соседей (k-NN), который также применяется не только для регрессии датасета, но и для классификации данных.

## **Описание метода**

Данный метод интуитивно прост. Здесь мы каждый новый объект определяем к каком-либо классу на основе принадлежности соседних объектов к своим классам. В зависимости от параметра k мы рассматриваем разное код-во соседей для каждого нового объекта.

**Принцип работы:**

* Рассчитываются расстояния между объектом и всеми объектами обучающей выборки (чаще всего используется Евклидово расстояние).
* Выбираются k ближайших объектов.
* Определяется класс нового объекта как наиболее часто встречающийся среди его ближайших соседей.

## **Псевдокод метода**

|  |
| --- |
| 1.Для каждого объекта в тестовой выборке:   * Вычислить расстояние до каждого объекта в обучающей выборке. * Отсортировать объекты обучающей выборки по возрастанию расстояния. * Выбрать k ближайших соседей. * Определить класс как наиболее часто встречающийся среди этих k соседей.   2. Вернуть предсказанные классы для всех объектов в тестовой выборке. |

## **Результаты выполнения**

Для того, чтобы оценить качество модели, построенной с помощью метода k-ближайших соседей, необходимо построить матрицы точности, где ячейки с одинаковым индексом по столбцам и строчкам будут показывать нам верно предсказанные классы, а все остальные будут указывать нам на промахи нашей модели. Вот что мы получили для нашего варианта:

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Изображение выглядит как снимок экрана, текст, Прямоугольник, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

## **Примеры использования кода**

Метод k-ближайших соседей может быть полезен в следующих ситуациях:

* Медицина: Классификация заболеваний на основе симптомов пациента, где каждый новый пациент сравнивается с историческими данными больных с похожими симптомами.
* Рекомендательные системы: Определение предпочтений пользователя на основе предпочтений его ближайших "соседей" (пользователей с похожими интересами).
* Маркетинг: Сегментация клиентов на основе их покупательского поведения для предсказания будущих покупок.

Этот метод был выбран для задачи классификации из-за его простоты в реализации и эффективности при небольших наборах данных.

**Лабораторная работа 3. Деревья решений**

## **Введение**

Рассмотрим теперь третий метод, который также нужен для регрессии и классификации данных в сфере МО. В данном методе мы будем подходить к разделению данных на категории с помощью дерева решений.

## **Описание метода**

Метод деревьев решений — это алгоритм машинного обучения для задач классификации и регрессии. Он строит модель в виде дерева, где каждая внутренняя вершина соответствует признаку данных, каждая ветвь — это условие на значение признака, а каждый лист — это итоговый класс или предсказание. Основная идея заключается в разделении данных на подмножества с минимальной примесью, например, с помощью критерия Джини или энтропии.

## **Псевдокод метода**

|  |
| --- |
| 1. Если все объекты принадлежат одному классу, вернуть этот класс.  2. Если достигнута максимальная глубина дерева, вернуть наиболее часто встречающийся класс.  3. Иначе для каждого признака:  a) Разбить данные по уникальным значениям признака.  b) Вычислить критерий примеси (Джини, энтропия).  c) Выбрать признак с наибольшим приростом информации.  4. Разделить данные на подмножества по выбранному признаку.  5. Рекурсивно строить поддеревья для каждого подмножества.  6. Вернуть дерево. |

## **Результаты выполнения**

Для выявления показателей качества, построенной модели нарисуем кривые PR и ROC:

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Автоматически созданное описание

## **Примеры использования кода**

Метод деревьев решений может быть полезен в задачах, где важно объяснить процесс принятия решения. Например:

* В медицине для диагностики заболеваний на основе симптомов.
* В маркетинге для сегментации клиентов по признакам покупательского поведения.
* В финансовой сфере для оценки кредитного риска.

Метод деревьев решений выбран за его простоту и интерпретируемость, что делает его полезным для задач, где важно понимать, почему было принято определенное решение.

**Лабораторная работа 4. Логистическая регрессия**

## **Введение**

В этой лабораторной работе нам нужно обучить модель, способной предсказывать вероятности принадлежности объекта к одному из двух классов. То есть строим модель для бинарной классификации.

## **Описание метода**

Логистическая регрессия — это метод классификации, который используется для прогнозирования вероятности принадлежности объекта к одному из двух классов. В основе метода лежит использование логистической функции (сигмоиды), которая преобразует линейную комбинацию входных признаков в значения вероятности. Метод подходит для решения задач бинарной классификации, когда целевая переменная может принимать только два значения, например, "да" или "нет", "1" или "0".

Принцип работы метода:

1. На основе входных данных формируется линейная комбинация признаков.
2. Полученное значение пропускается через логистическую функцию, которая ограничивает результат значениями от 0 до 1.
3. На основе порога (например, 0.5) принимается решение о том, к какому классу принадлежит объект.

## **Псевдокод метода**

|  |
| --- |
| 1. Инициализация весов модели случайными значениями  2. Для каждой итерации:  a. Рассчитать линейную комбинацию признаков: z = w0 + w1 \* x1 + ... + wn \* xn  b. Применить сигмоидную функцию для вычисления вероятности: p = 1 / (1 + exp(-z))  c. Вычислить ошибку модели: E = - (y \* log(p) + (1 - y) \* log(1 - p))  d. Обновить веса модели на основе градиентного спуска: w = w - α \* ∇E  e. Проверить критерий остановки (например, если изменение в весах незначительно)  3. Возвращать веса, соответствующие минимальной ошибке |

## **Результаты выполнения**

Изображение выглядит как текст, снимок экрана, меню, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, дизайн

Автоматически созданное описание

## **Примеры использования кода**

Метод логистической регрессии широко применяется в следующих ситуациях:

1. **Медицина**: для предсказания вероятности наличия заболевания на основе симптомов и анализов пациента.
2. **Маркетинг**: для прогнозирования вероятности того, что клиент совершит покупку, исходя из данных о его предыдущих действиях.
3. **Кредитный скоринг**: для оценки вероятности того, что клиент не вернет кредит на основе его финансовой истории и характеристик.

Метод логистической регрессии был выбран в этих случаях из-за его интерпретируемости и простоты. Модель позволяет легко понять, как каждый признак влияет на вероятность того или иного исхода, что важно для принятия решений в различных областях, таких как медицина и финансы.

**Сравнение методов**

Методы машинного обучения, такие как линейная регрессия, метод k-ближайших соседей (k-NN), деревья решений и логистическая регрессия, имеют свои уникальные особенности, преимущества и ограничения.

1. Линейная регрессия:
   * Преимущества: Простота в реализации и интерпретации. Хорошо работает на линейно зависимых данных. Быстрая вычислительная скорость.
   * Ограничения: Предполагает линейную зависимость между входными и выходными переменными. Чувствительна к выбросам.
2. Метод k-ближайших соседей (k-NN):
   * Преимущества: Простота и эффективность. Не требует предварительной тренировки модели. Хорошо справляется с многомерными данными.
   * Ограничения: Высокая вычислительная сложность при предсказании (особенно на больших выборках). Чувствителен к выбору параметра k и к масштабу признаков.
3. Деревья решений:
   * Преимущества: Простота интерпретации и визуализации. Могут обрабатывать как числовые, так и категориальные данные. Не требуют масштабирования признаков.
   * Ограничения: Склонны к переобучению, особенно при глубоком дереве. Чувствительны к небольшим изменениям в данных.
4. Логистическая регрессия:
   * Преимущества: Простота и эффективность. Хорошо работает для бинарной классификации. Интерпретируемость коэффициентов модели.
   * Ограничения: Предполагает линейную зависимость между признаками и логарифмом шансов. Меньшая гибкость по сравнению с более сложными методами.

Примеры лучшего использования каждого метода

* Линейная регрессия: Эффективна для задач, где необходимо прогнозировать непрерывные значения, такие как цена недвижимости или температура, при наличии линейных зависимостей.
* Метод k-ближайших соседей (k-NN): Наиболее эффективен в задачах, где нет явных границ между классами, например, в распознавании образов или в рекомендационных системах.
* Деревья решений: Идеальны для задач с разнородными данными, таких как диагностика заболеваний, где данные могут содержать как категориальные, так и числовые признаки. Подходят для обработки больших объемов данных.
* Логистическая регрессия: Наилучший выбор для бинарной классификации, например, в задачах кредитного скоринга или медицинской диагностики, когда необходимо оценить вероятность принадлежности к классу.

**Заключение**

В заключение выбор метода машинного обучения зависит от конкретной задачи, доступных данных и требований к интерпретации модели. Линейная регрессия и логистическая регрессия предпочтительны для простых задач, где важно понимание взаимосвязей между переменными. В то время как k-NN и деревья решений предлагают большую гибкость и могут справляться с более сложными структурами данных. Каждый из методов имеет свои сильные и слабые стороны, что делает их более или менее подходящими в зависимости от контекста применения.