Отчет по модулю №2 по предмету

“Системы искусственного интеллекта”

**Выполнил**: Кочнев Роман

**Группа**: P33081

**Преподаватель:** Авдюшина Анна

Евгеньевна

Санкт-Петербург

2023

Оглавление

[Лабораторная работа 1. Метод линейной регрессии 3](#_Toc152542132)

[Введение 3](#_Toc152542133)

[Описание метода 3](#_Toc152542134)

[Псевдокод метода 3](#_Toc152542135)

[Результаты выполнения 4](#_Toc152542136)

[Первая модель: 4](#_Toc152542137)

[Вторая модель: 4](#_Toc152542138)

[Третья модель: 4](#_Toc152542139)

[Четвертая модель: 4](#_Toc152542140)

[Примеры использования метода 4](#_Toc152542141)

[Примеры использования метода 4](#_Toc152542142)

[Лабораторная работа 2. Метод k-ближайших соседей (k-NN) 5](#_Toc152542143)

[Введение 5](#_Toc152542144)

[Описание метода 5](#_Toc152542145)

[Псевдокод метода 5](#_Toc152542146)

[Результаты выполнения 6](#_Toc152542147)

[Примеры использования метода 6](#_Toc152542148)

[Лабораторная работа 3. Деревья решений 6](#_Toc152542149)

[Введение 6](#_Toc152542150)

[Описание метода 7](#_Toc152542151)

[Псевдокод метода 7](#_Toc152542152)

[Результаты 7](#_Toc152542153)

[Примеры использования метода 8](#_Toc152542154)

[Лабораторная работа 4. Логистическая регрессия 8](#_Toc152542155)

[Введение 8](#_Toc152542156)

[Описание метода 8](#_Toc152542157)

[Псевдокод метода 8](#_Toc152542158)

[Результаты выполнения 9](#_Toc152542159)

[Примеры использования метода 10](#_Toc152542160)

[Сравнение методов 10](#_Toc152542161)

[Сравнительный анализ методов 10](#_Toc152542162)

[Примеры лучшего использования каждого метода 10](#_Toc152542163)

[Заключение 11](#_Toc152542164)

[Приложения 11](#_Toc152542165)

[LAB4 11](#_Toc152542166)

[LAB5 11](#_Toc152542167)

[LAB6 11](#_Toc152542168)

[LAB7 11](#_Toc152542169)

# Лабораторная работа 1. Метод линейной регрессии

## Введение

Цель данного исследования заключается в разработке и оценке трех моделей линейной регрессии для прогнозирования результатов обучения студентов. Мы стремимся выявить наилучшие признаки, которые способствуют эффективной модели прогнозирования. Для достижения этой цели мы выполним анализ данных, предварительную обработку, построение моделей и оценку их производительности.

## Описание метода

Линейная регрессия - это метод, который используется для моделирования отношений между зависимой переменной (в данном случае, результаты обучения студентов) и набором независимых переменных (признаки). Основной принцип линейной регрессии заключается в поиске оптимальных коэффициентов, минимизирующих сумму квадратов разностей между фактическими и предсказанными значениями.

## Псевдокод метода

1. Загрузка данных:

- Загрузить данные, содержащие зависимую переменную (y) и независимые переменные (x).

2. Предобработка данных:

- Обработать отсутствующие значения, если они есть.

- Нормализовать признаки, если необходимо.

3. Инициализация параметров:

- Задать начальные значения параметров модели: веса (weights) и смещение (bias).

4. Определение функции гипотезы:

- Определить функцию гипотезы h(x), которая представляет собой линейное уравнение: h(x) = w1\*x1 + w2\*x2 + ... + wn\*xn + b, где w - веса, x - независимые переменные, и b - смещение.

5. Определение функции стоимости:

- Определить функцию стоимости J(w, b), например, среднеквадратичную ошибку: J(w, b) = (1/2m) \* ∑(h(xi) - yi)^2, где m - количество обучающих примеров, xi - i-й обучающий пример, и yi - соответствующая ему метка.

6. Обучение модели (градиентный спуск):

- Пока не достигнут критерий останова:

- Рассчитать градиенты частных производных функции стоимости по отношению к весам и смещению.

- Обновить параметры модели с использованием градиентного спуска: w := w - α \* ∂J/∂w, b := b - α \* ∂J/∂b, где α - скорость обучения.

7. Возвращение обученной модели:

- Вернуть обученные параметры w и b.

8. Прогноз:

- Для новых данных x, используя обученные параметры, рассчитать прогнозную переменную y\_pred: y\_pred = w1\*x1 + w2\*x2 + ... + wn\*xn + b.

9. Оценка модели (опционально):

- Если есть правильные метки для тестового набора, оценить качество модели, например, среднеквадратичную ошибку или коэффициент детерминации.

10. Итерация:

- Повторить шаги 4-9, если необходимо настроить параметры или улучшить модель.

## Результаты выполнения

### Первая модель:

Коэффициенты: Hours Studied (22.823), Previous Scores (60.005), Sleep Hours (2.367), Sample Question Papers Practiced (1.726).

MSE: 4.182.

R^2: 0.989.

### Вторая модель:

Коэффициенты: Hours Studied (22.847), Previous Scores (60.025), Sleep Hours (2.367).

MSE: 4.545.

R^2: 0.988.

### Третья модель:

Коэффициенты: Hours Studied (2.524), Previous Scores (59.752).

MSE: 58.537.

R^2: 0.842.

### Четвертая модель:

Коэффициенты: Hours Studied (22.820), Previous Scores (60.002), Sleep Hours (2.385), Sample Question Papers Practiced (1.726), Extracurricular Activities\_Yes (0.609).

## Примеры использования метода

На основе предоставленных данных и результатов моделей, признаки можно расположить по убыванию их эффективности для прогнозирования производительности студентов:

Previous Scores (Предыдущие баллы):

Присутствует во всех четырех моделях.

Считается одним из наиболее важных признаков.

Hours Studied (Количество часов учебы):

Присутствует во всех четырех моделях.

Является ключевым признаком для прогнозирования производительности.

Sleep Hours (Количество часов сна):

Присутствует в двух моделях (третьей и четвертой).

Указывает на важность правильного режима сна для успеха студентов.

Sample Question Papers Practiced (Практика на образцах вопросов):

Присутствует только в четвертой модели.

Практика на образцах вопросов положительно влияет на производительность.

Extracurricular Activities\_Yes (Участие во внеклассных мероприятиях):

Присутствует только в четвертой модели.

Участие во внеклассных мероприятиях также положительно связано с успехом

## Примеры использования метода

Компания хочет предсказать объем продаж своего продукта на основе различных факторов, таких как бюджет на рекламу, время года, экономический рост и другие.

Применение линейной регрессии: Метод линейной регрессии может быть использован для построения модели, предсказывающей объем продаж на основе линейной комбинации различных факторов. Например, модель может выглядеть следующим образом: Продажи = a \* Бюджет + b \* Время года + c \* Экономический рост + ....

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

## Введение

Цель данной работы заключается в проведении анализа и построении моделей на основе метода k-ближайших соседей (k-NN) для двух различных датасетов: о вине и о диабете. Мы будем выполнять предварительную обработку данных, визуализацию статистики, а также реализацию и оценку моделей k-NN с разными наборами признаков.

## Описание метода

Метод k-ближайших соседей (k-NN) является алгоритмом классификации, основанным на простой идее: объект присваивается тому классу, который является наиболее распространенным среди k его ближайших соседей. Расстояние между объектами может измеряться различными метриками, такими как евклидово расстояние.

## Псевдокод метода

1. Загрузка данных:

- Загрузить обучающий набор данных.

- Загрузить тестовый набор данных (если требуется).

2. Предобработка данных:

- Нормализовать признаки, если необходимо.

- Обработать отсутствующие значения, если они есть.

3. Выбор значения k:

- Выбрать количество соседей (k) для алгоритма.

- Определить метрику расстояния (например, евклидово расстояние).

4. Расчет расстояний:

- Для каждого объекта в тестовом наборе данных:

- Рассчитать расстояние до каждого объекта в обучающем наборе.

5. Сортировка и выбор соседей:

- Отсортировать объекты обучающего набора по расстоянию к тестовому объекту.

- Выбрать k ближайших соседей.

6. Принятие решения:

- В случае классификации: определить класс большинства среди k соседей.

- В случае регрессии: вычислить среднее или медиану целевых значений k соседей.

7. Возвращение результата:

- Вернуть предсказание для каждого тестового объекта.

8. Оценка модели (опционально):

- Если есть правильные метки для тестового набора, оценить точность или другие метрики.

9. Итерация:

- Повторить шаги 3-8, если необходимо настроить параметры или улучшить модель.

## Результаты выполнения

Матрица ошибок для модели 1 с k=3:

[[64 34]

[37 18]]

Матрица ошибок для модели 1 с k=5:

[[78 20]

[38 17]]

Матрица ошибок для модели 1 с k=10:

[[85 13]

[36 19]]

Матрица ошибок для модели 2 с k=3:

[[83 15]

[27 28]]

Матрица ошибок для модели 2 с k=5:

[[84 14]

[27 28]]

Матрица ошибок для модели 2 с k=10:

[[90 8]

[33 22]]

## Примеры использования метода

**Классификация объектов в медицинских диагнозах:**

Представим, что у нас есть множество медицинских данных с информацией о пациентах, включая результаты различных тестов.

Применение k-NN: Метод k-NN может быть использован для классификации нового пациента в группу с определенным заболеванием на основе его ближайших соседей. Например, если у большинства соседей есть диагноз диабета, то новый пациент, чьи параметры близки к этим соседям, также, вероятно, имеет диабет.

# Лабораторная работа 3. Деревья решений

## Введение

Лабораторная работа направлена на разработку и оценку алгоритма построения дерева решений для задачи классификации. Для этого используются два различных датасета: один связан с классификацией грибов, а другой с данными об оценках студентов на инженерном и педагогическом факультетах. Цель работы - реализовать дерево решений без использования сторонних библиотек, отобрать случайные признаки и провести оценку модели с использованием метрик Accuracy, Precision, Recall, а также построить кривые AUC-ROC и AUC-PR.

## Описание метода

Дерево решений - это алгоритм машинного обучения, который используется для принятия решений на основе серии вопросов и условий. Он строит дерево, в каждом узле которого находится вопрос, разделяющий данные на подгруппы. Цель - разбить данные таким образом, чтобы в каждой подгруппе объекты имели схожие значения целевой переменной.

## Псевдокод метода

1. Определение структуры узла дерева:

- Каждый узел содержит вопрос (или условие) и два потомка: один для случаев, когда условие истинно, и другой - когда условие ложно.

- Листовые узлы содержат прогноз или классификацию.

2. Определение критерия разделения:

- Выбор критерия, например, индекса Джини или энтропии, для определения того, как разделить данные на каждом узле.

3. Рекурсивное построение дерева:

- На каждом узле выбрать признак и пороговое значение для разделения данных.

- Создать два поддерева для случаев, когда условие верно и ложно.

- Рекурсивно повторять этот процесс для каждого поддерева до достижения критерия останова (например, максимальной глубины дерева).

4. Прогноз/классификация новых данных:

- Пройти по дереву, от корня к листьям, используя значения признаков нового объекта, чтобы принять решение.

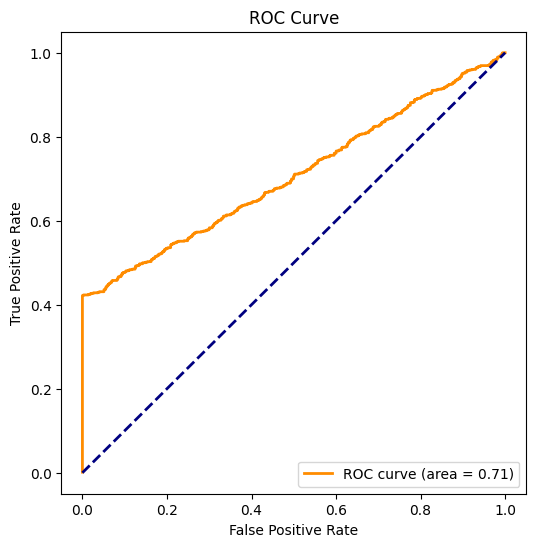
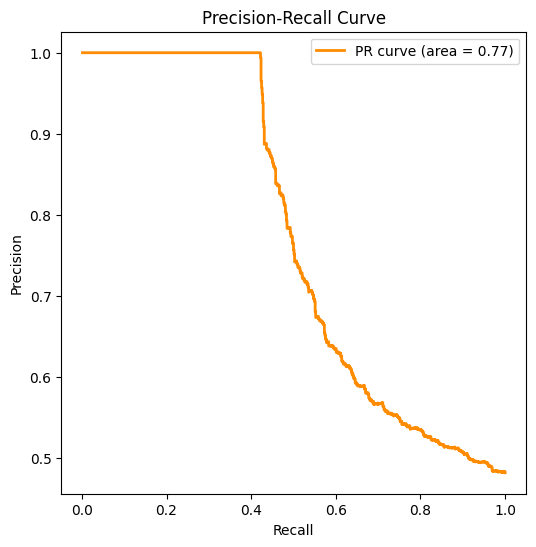
5. Оценка модели (опционально):

- Если есть правильные метки для тестового набора, оценить точность или другие метрики.

6. Итерация:

- Повторить шаги 2-5, если необходимо настроить параметры или улучшить модель.

Результаты выполнения

AUC-ROC: 0.7068

AUC-PR: 0.7708

## Примеры использования метода

**Классификация грибов:**

Имеется датасет с описанием грибов и их классификацией на ядовитые и съедобные.

Применение дерева решений: Дерево решений может быть использовано для построения модели, способной классифицировать грибы на ядовитые и съедобные на основе их характеристик.

# Лабораторная работа 4. Логистическая регрессия

## Введение

Цель данной работы состоит в построении и оценке модели логистической регрессии для задачи классификации на выбранном датасете. В качестве данных используются информация о пассажирах Титаника. Основные шаги включают в себя предварительную обработку данных, визуализацию статистики, разделение на обучающий и тестовый наборы, а также реализацию логистической регрессии "с нуля" с возможностью варьирования гиперпараметров.

## Описание метода

Логистическая регрессия - это метод машинного обучения, применяемый для решения задач бинарной классификации. Он использует логистическую функцию для оценки вероятности принадлежности объекта к одному из двух классов. Параметры модели настраиваются с использованием метода максимального правдоподобия.

## Псевдокод метода

1. Загрузка данных:

- Загрузить данные, содержащие метки классов (0 или 1) и независимые переменные (признаки).

2. Предобработка данных:

- Обработать отсутствующие значения, если они есть.

- Нормализовать признаки, если необходимо.

3. Инициализация параметров:

- Задать начальные значения весов (weights) для каждого признака и смещения (bias).

4. Определение функции гипотезы (sigmoid):

- Определить функцию гипотезы h(x) = 1 / (1 + exp(-(w1\*x1 + w2\*x2 + ... + wn\*xn + b))), где exp - функция экспоненты.

5. Определение функции стоимости:

- Определить функцию стоимости J(w, b) = (-1/m) \* ∑[y\*log(h(x)) + (1-y)\*log(1-h(x))], где m - количество обучающих примеров, y - метка класса, h(x) - предсказанная вероятность.

6. Обучение модели (градиентный спуск):

- Пока не достигнут критерий останова:

- Рассчитать градиенты частных производных функции стоимости по отношению к весам и смещению.

- Обновить параметры модели с использованием градиентного спуска: w := w - α \* ∂J/∂w, b := b - α \* ∂J/∂b, где α - скорость обучения.

7. Возвращение обученной модели:

- Вернуть обученные параметры w и b.

8. Прогноз:

- Для новых данных x, используя обученные параметры, рассчитать предсказанную вероятность h(x).

- Если h(x) > 0.5, классифицировать как 1; иначе, классифицировать как 0.

9. Оценка модели (опционально):

- Если есть правильные метки для тестового набора, оценить точность или другие метрики.

10. Итерация:

- Повторить шаги 4-9, если необходимо настроить параметры или улучшить модель.

## Результаты выполнения

Learning Rate:

Learning Rate 0.1 приводит к быстрому расхождению функции потерь (loss) и, вероятно, переполнению (overflow) в вычислении сигмоиды. Learning Rate 0.01 также приводит к расхождению и высокой потере точности. Learning Rate 0.001 показывает лучшие результаты, и при этом достигается приемлемая точность.

Iterations:

При Learning Rate 0.1 и 500 итерациях обучения происходит быстрое расхождение, что может быть связано с слишком большими обновлениями весов. Увеличение количества итераций (даже до 1500) при Learning Rate 0.1 не приводит к улучшению производительности, скорее всего из-за больших шагов градиентного спуска.

Метод оптимизации:

Градиентный спуск и стохастический градиентный спуск показывают схожие результаты, что может быть обусловлено небольшим размером данных.

## Примеры использования метода

**Прогноз выживания на Титанике:**

Имеется информация о пассажирах Титаника, включая их возраст, класс, пол и другие параметры.

Применение логистической регрессии: Логистическая регрессия может быть использована для предсказания вероятности выживания пассажира на основе предоставленных данных.

# Сравнение методов

## Сравнительный анализ методов

1. Линейная регрессия:

**Преимущества**:

* Простота и понятность.
* Эффективность в задачах, где зависимость между признаками и целевой переменной линейна.

**Ограничения**:

* Неспособность моделировать сложные нелинейные зависимости.
* Чувствительность к выбросам.

2. Метод k ближайших соседей:

**Преимущества**:

* Хорошая адаптация к сложным нелинейным зависимостям.
* Не требует предварительного обучения.

**Ограничения**:

* Зависимость от выбора числа соседей и метрики расстояния.
* Высокая вычислительная сложность при большом объеме данных.

3. Деревья решений:

**Преимущества**:

* Способность моделировать сложные зависимости и взаимодействия признаков.
* Интерпретируемость результатов.

**Ограничения**:

* Тенденция к переобучению, особенно на больших данных.
* Нестабильность при небольших изменениях данных.

4. Логистическая регрессия:

**Преимущества**:

* Эффективность в задачах классификации.
* Возможность интерпретации коэффициентов.

**Ограничения**:

* Предположение о линейной зависимости между признаками и логарифмом шансов.
* Неспособность моделировать сложные нелинейные отношения.

## Примеры лучшего использования каждого метода

**1. Линейная регрессия:**

Прогнозирование цены на недвижимость на основе факторов, таких как количество спален, площадь и расстояние до центра города.

**2. Метод k ближайших соседей:**

Классификация текстовых документов по темам на основе их векторного представления.

**3. Деревья решений:**

Прогнозирование, будет ли клиент банка погашать кредит, основываясь на различных финансовых и личных характеристиках.

**4. Логистическая регрессия:**

Определение вероятности заболевания пациента на основе медицинских параметров для задачи бинарной классификации (болен/не болен).

## Заключение

В общем, выбор метода зависит от характера данных и поставленной задачи. Линейная регрессия эффективна при линейных зависимостях, k ближайших соседей хорош для нелинейных данных, деревья решений подходят для моделирования сложных взаимодействий, а логистическая регрессия применима в задачах бинарной классификации. Экспериментирование с различными методами и их гиперпараметрами может привести к оптимальным результатам в конкретной задаче машинного обучения.

## Приложения

### [LAB4](https://colab.research.google.com/drive/1DL7q4ASwMVtYmO0dgdyooNjTeM_1rboO)

### [LAB5](https://colab.research.google.com/drive/14XRnXsNJvNtCZwdcDmmbRNNlXDK1VZTf)

### [LAB6](https://colab.research.google.com/drive/10VJyUREq0ppyCY068wzr7lpg6e6h1UVq#scrollTo=5D7icx_GTAjI)

### [LAB7](https://colab.research.google.com/drive/1-1FnzpVN2uETJ-GW3X88VtwO0WhEe7Mf#scrollTo=poloti9kBcBB)