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

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:

```
R^2 для первой модели: 0.47842815120849735
R^2 для второй модели: 0.5184622405521854
R^2 для третьей модели: 0.5522503959624603
Сравнение моделей:
Модель 1 R^2: 0.47842815120849735
Модель 2 R^2: 0.5184622405521854
Модель 3 R^2: 0.5522503959624603 (с синтетическим признаком)
R^2 для тестовых данных первой модели: 0.47971026264117556
R^2 для тестовых данных второй модели: 0.517897215988669
R^2 для тестовых данных третьей модели: 0.5533400004490683
Сравнение моделей на тестовых данных:
Модель 1 R^2: 0.47971026264117556
Модель 2 R^2: 0.517897215988669
Модель 3 R^2: 0.5533400004490683 (с синтетическим признаком)
```

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.

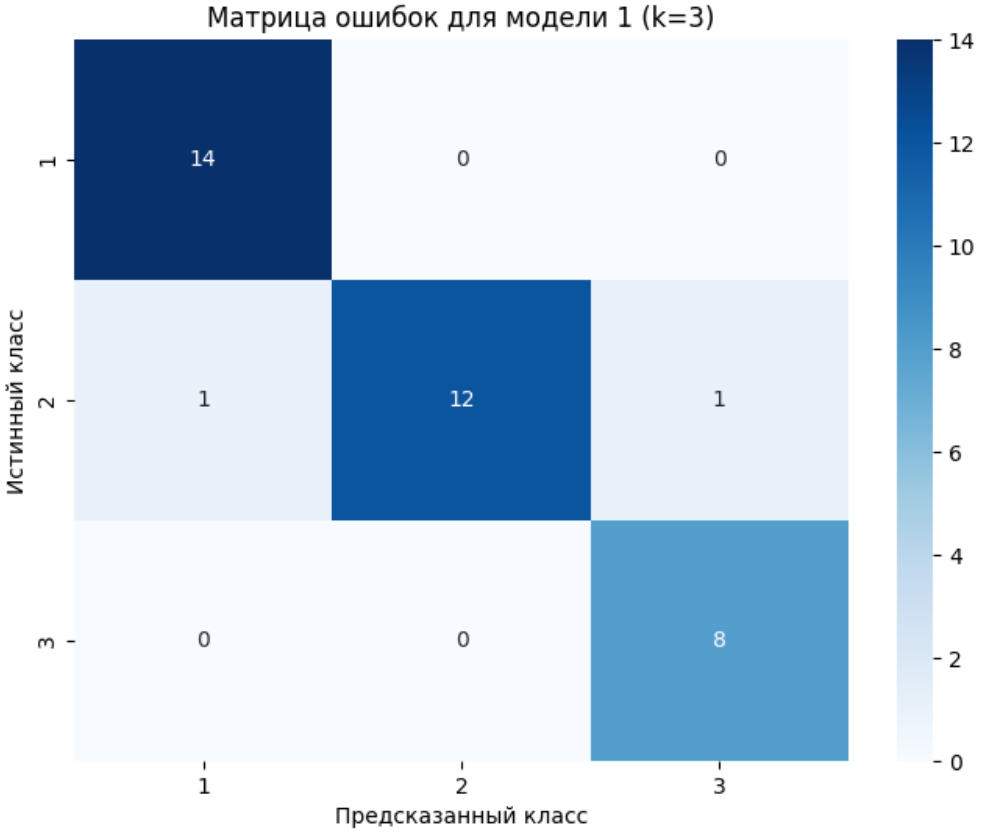
Псевдокод метода

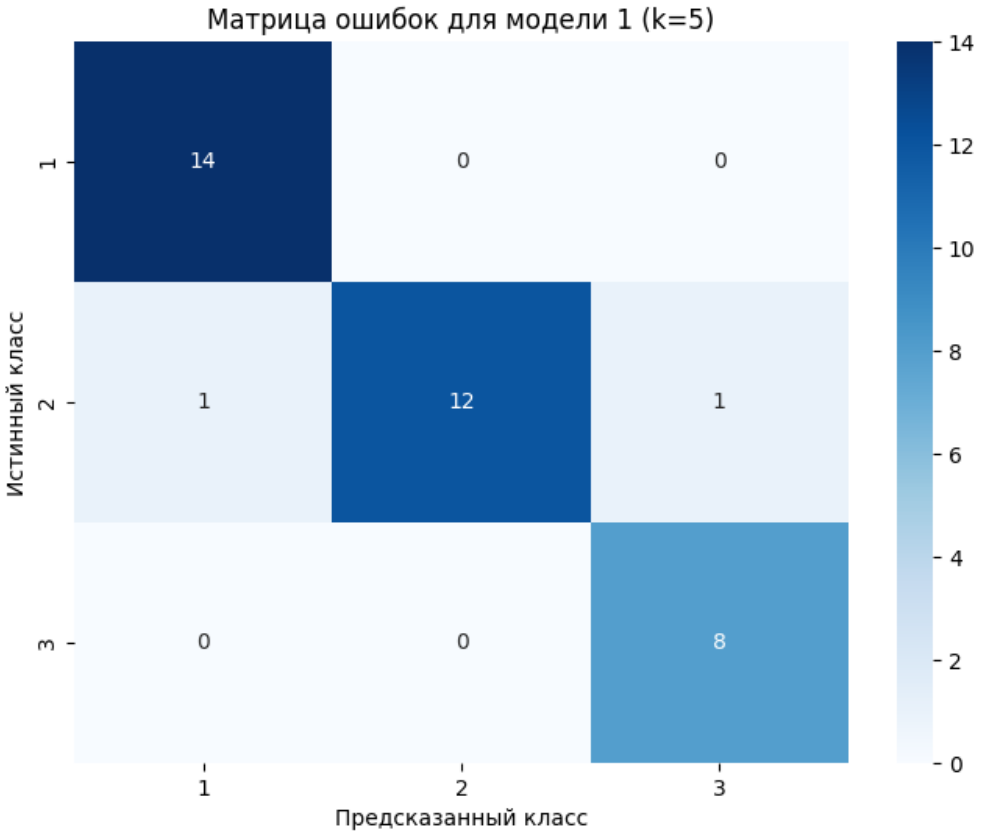
1.Для каждого объекта в тестовой выборке:
- Вычислить расстояние до каждого объекта в обучающей выборке.
- Отсортировать объекты обучающей выборки по возрастанию расстояния.
- Выбрать k ближайших соседей.
- Определить класс как наиболее часто встречающийся среди этих k соседей.

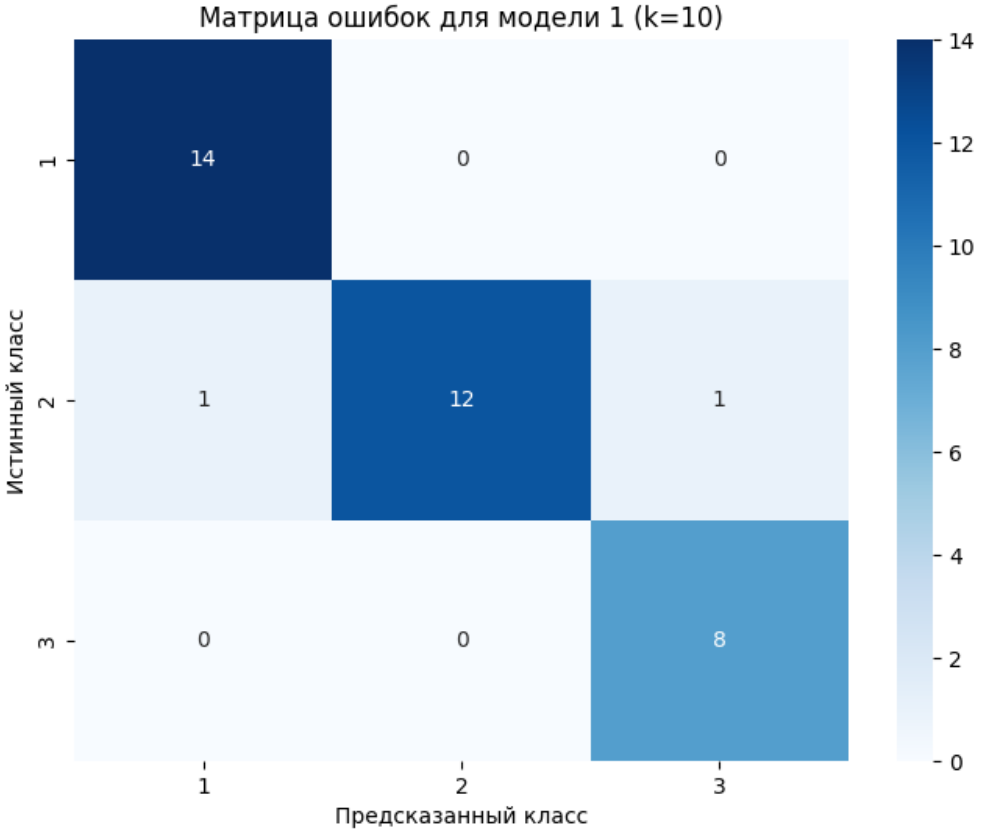
2. Вернуть предсказанные классы для всех объектов в тестовой выборке.

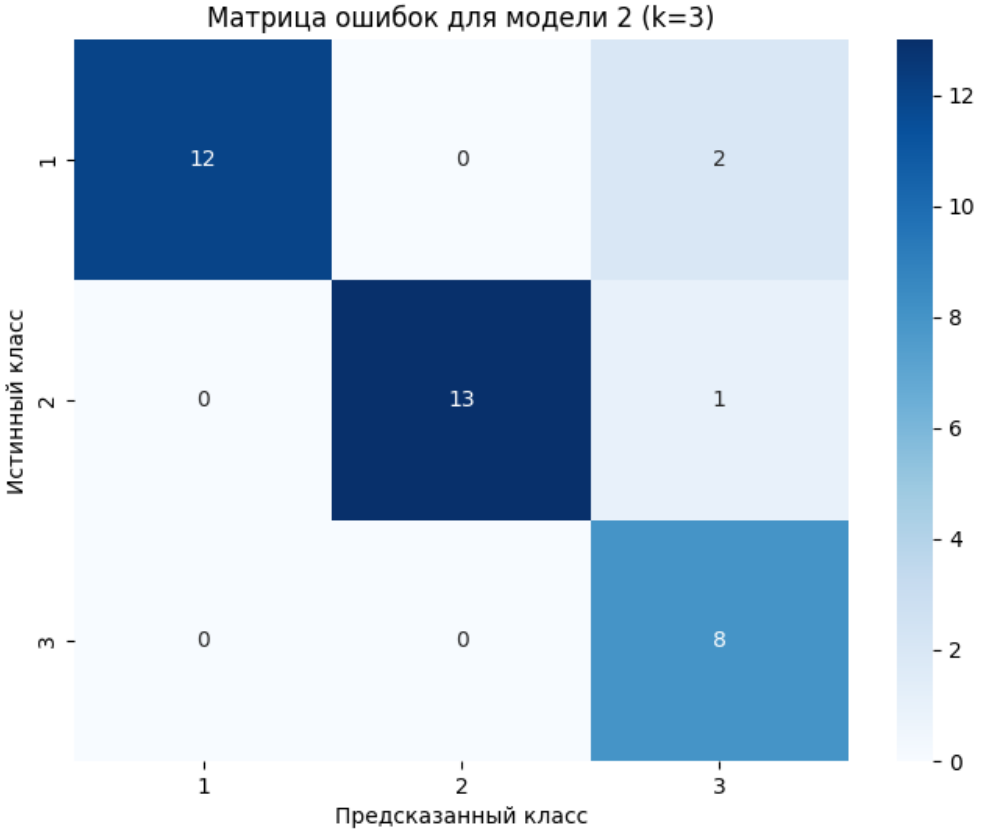
Результаты выполнения

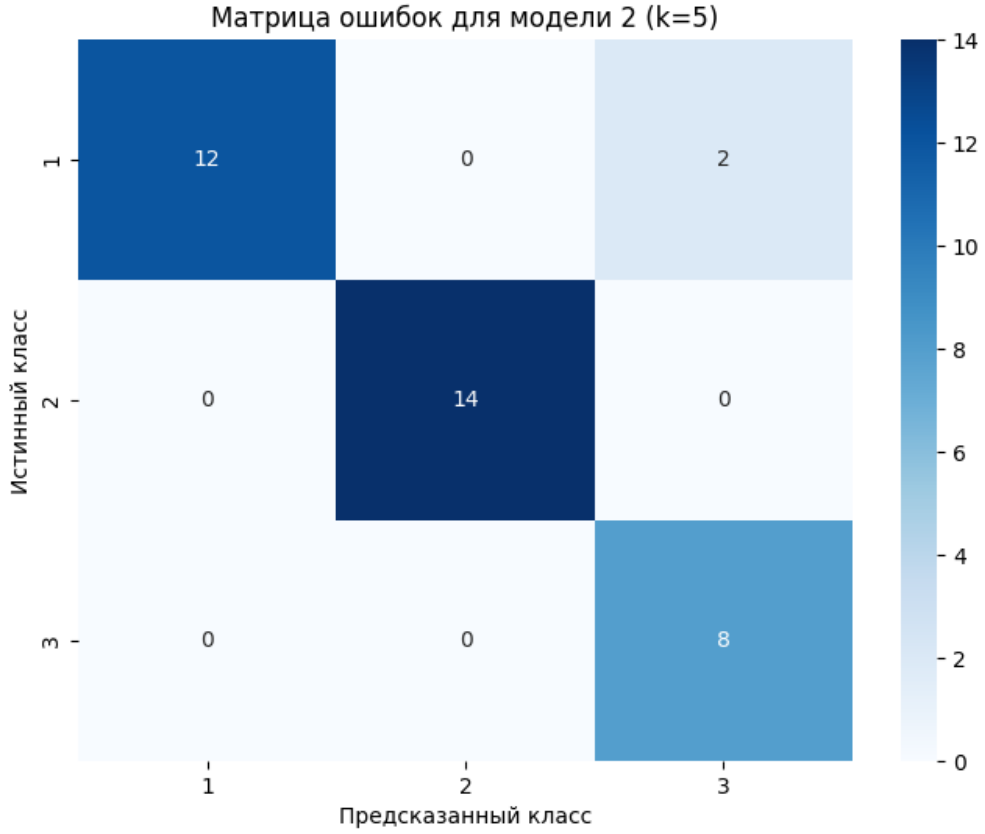
Для того, чтобы оценить качество модели, построенной с помощью метода k-ближайших соседей, необходимо построить матрицы точности, где ячейки с одинаковым индексом по столбцам и строчкам будут показывать нам верно предсказанные классы, а все остальные будут указывать нам на промахи нашей модели. Вот что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

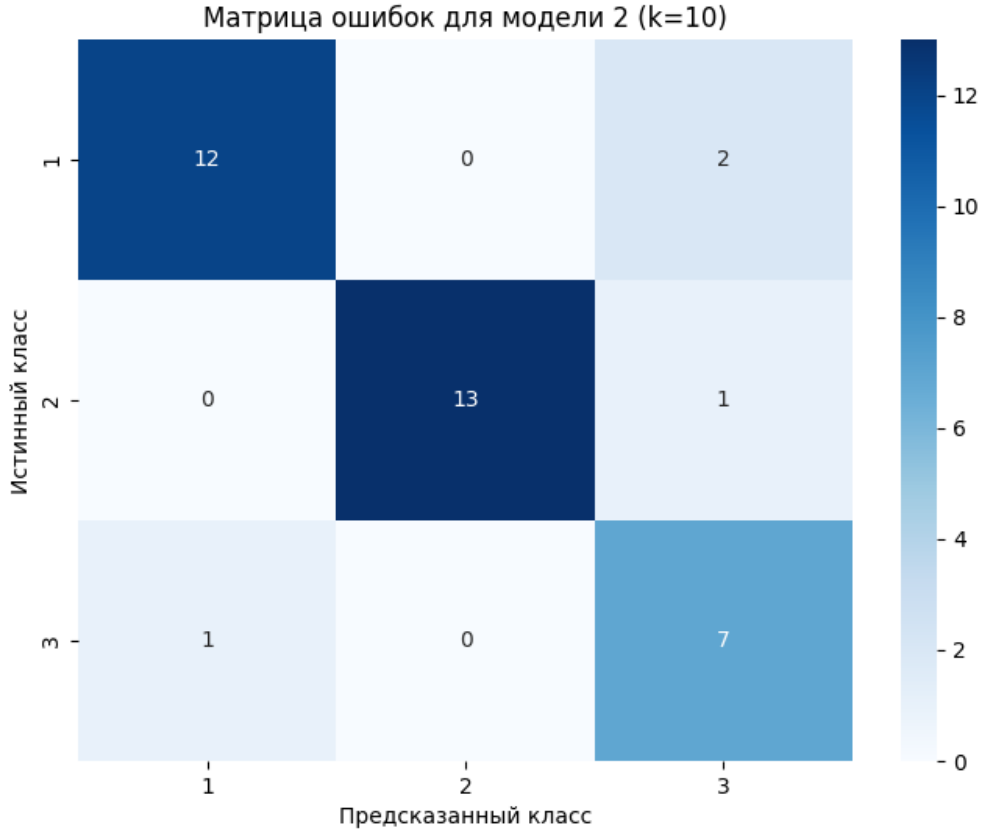

Матрица ошибок для модели 1 (k=3)
Истинный класс
Предсказанный класс
1
2
3
14
0
0
1
12
1
0
0
8
0
2
4
6
8
10
12
14


Матрица ошибок для модели 1 (k=5)
Истинный класс
Предсказанный класс
1
2
3
14
0
0
1
12
1
0
0
8
0
2
4
6
8
10
12
14


Матрица ошибок для модели 1 (k=10)
Истинный класс
Предсказанный класс
14
0
0
1
12
1
0
0
8
1
2
3
14
12
10
8
6
4
2
0


Матрица ошибок для модели 2 (k=3)
Истинный класс
Предсказанный класс
12
0
2
0
13
1
0
0
8
1
2
3
12
10
8
6
4
2
0


Матрица ошибок для модели 2 (k=5)
Истинный класс
Предсказанный класс
1
2
3
12
0
2
0
14
0
0
0
8
0
2
4
6
8
10
12
14


Матрица ошибок для модели 2 (k=10)
Истинный класс
Предсказанный класс
1
2
3
12
0
2
0
13
1
1
0
7
0
2
4
6
8
10
12

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

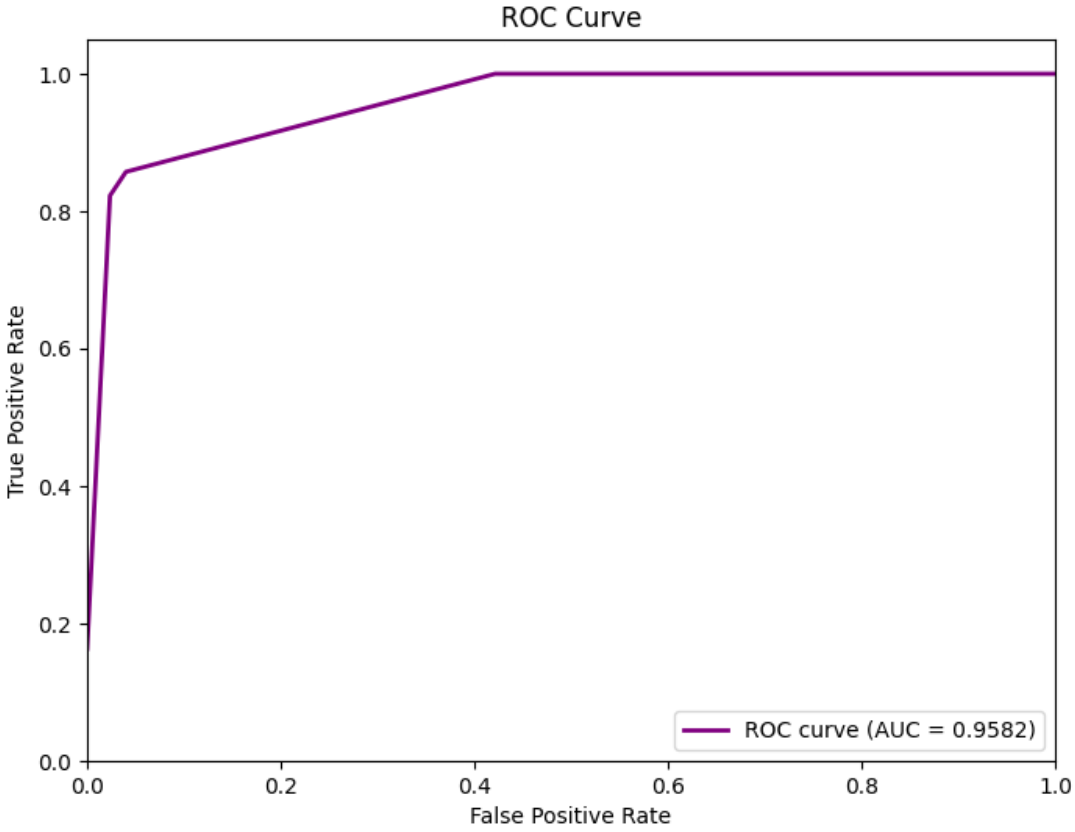
Метод деревьев решений — это алгоритм машинного обучения для задач классификации и регрессии. Он строит модель в виде дерева, где каждая внутренняя вершина соответствует признаку данных, каждая ветвь — это условие на значение признака, а каждый лист — это итоговый класс или предсказание. Основная идея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

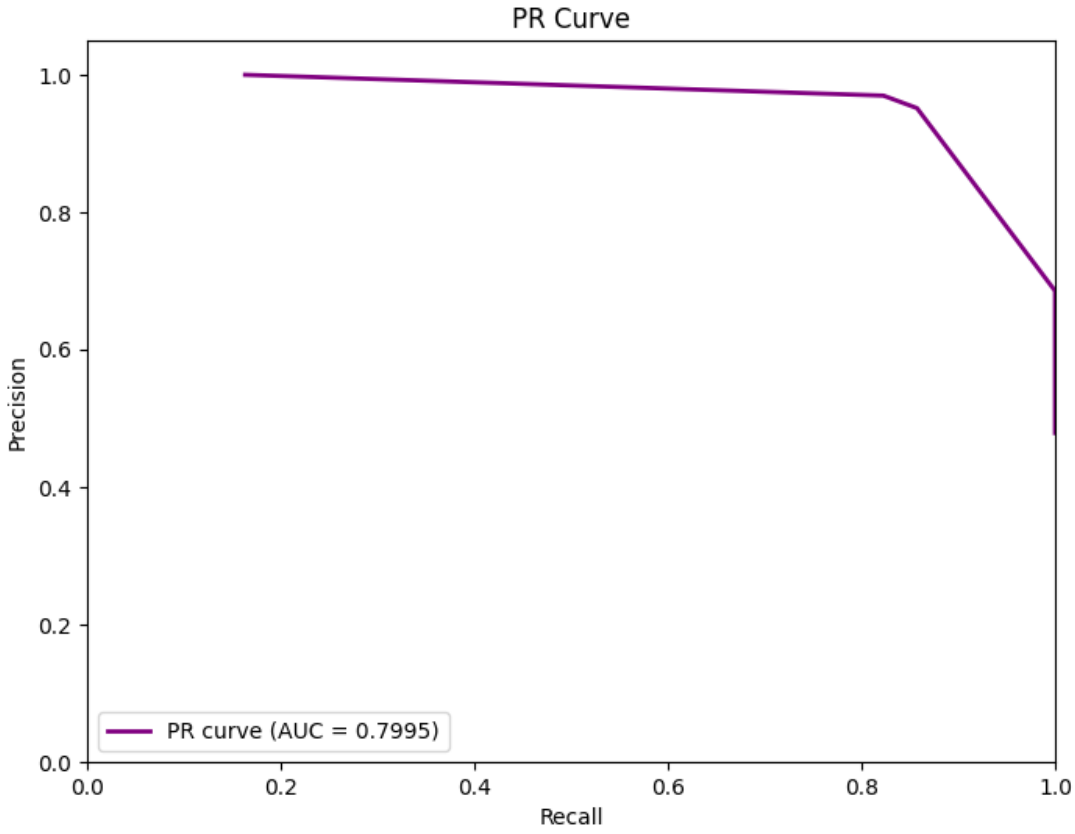
Псевдокод метода

```
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:


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.9582)



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

```
   Survived  Pclass  Sex   Age
0         0       3    0  22.0
1         1       1    1  38.0
2         1       3    1  26.0
3         1       1    1  35.0
4         0       3    0  35.0
Размер обучающего набора: (891, 3)
Размер тестового набора: (418, 3)
Gradien descent, 1000 iterations, step = 0.
Iteration 100, Loss: 0.7385061715236964
Iteration 200, Loss: 0.7196917099118204
Iteration 300, Loss: 0.7024729990011856
Iteration 400, Loss: 0.686157908299604
Iteration 500, Loss: 0.6704773851000836
Iteration 600, Loss: 0.6553252997317115
Iteration 700, Loss: 0.6406529441114973
Iteration 800, Loss: 0.6264292497106833
Iteration 900, Loss: 0.61262758930655
Iteration 1000, Loss: 0.5992227952735406
Accuracy: 0.80
Precision: 0.66
Recall: 0.97
F1-Score: 0.78
```

```
Newton method, 1000 iterations
Iteration 100, Loss: 0.45190225248406307
Iteration 200, Loss: 0.45190225248406307
Iteration 300, Loss: 0.4519022524840631
Iteration 400, Loss: 0.45190225248406307
Iteration 500, Loss: 0.45190225248406307
Iteration 600, Loss: 0.4519022524840631
Iteration 700, Loss: 0.45190225248406307
Iteration 800, Loss: 0.45190225248406307
Iteration 900, Loss: 0.4519022524840631
Iteration 1000, Loss: 0.45190225248406307
Accuracy: 0.96
Precision: 0.94
Recall: 0.96
F1-Score: 0.95
```

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.