УНИВЕРСИТЕТ ИТМО

Факультет программной инженерии и компьютерной техники

Направление подготовки 09.03.04 Программная инженерия

Дисциплина «Системы искусственного интеллекта»

**Лабораторная работа 6**

Студент

*Селянта Олег Дмитриевич*

*P3314*

Преподаватель

*Королёва Юлия Александровна*

Санкт-Петербург, 2024 г.

**Введение**

В данной лабораторной работе была реализована модель логистической регрессии для предсказания выживаемости пассажиров "Титаника" на основе предоставленных данных. Целью работы было изучение основных этапов построения модели машинного обучения: от предобработки данных до оценки качества модели с использованием различных метрик.

**1. Загрузка и предобработка данных**

* **Загрузка данных:** Данные были загружены из файлов train.csv и test.csv, предоставленных для соревнования "Titanic: Machine Learning from Disaster" на платформе Kaggle.
* **Выбор признаков:** Для обучения модели были выбраны следующие признаки:
  + **Pclass:** Класс каюты пассажира (1, 2 или 3).
  + **Sex:** Пол пассажира (мужской/женский).
  + **Age:** Возраст пассажира.
* **Обработка пропущенных значений:** Пропущенные значения в столбце Age были заменены средним значением возраста всех пассажиров.
* **Преобразование категориальных признаков:** Категориальный признак Sex был преобразован в числовой формат:
  + male -> 0
  + female -> 1
* **Разделение данных:** Данные из train.csv были разделены на матрицу признаков (X\_train) и вектор целевой переменной (y\_train). Данные из test.csv были преобразованы в матрицу признаков (X\_test).

**2. Визуализация данных**

Для каждого из выбранных признаков были построены гистограммы с отображением основных статистических характеристик:

* **Среднее значение:** Показывает среднее значение признака.
* **Медиана:** Показывает значение, которое делит набор данных пополам.
* **Квантили (25% и 75%):** Показывают значения, которые ограничивают 25% и 75% набора данных, соответственно.
* **Стандартное отклонение:** Показывает разброс значений признака относительно среднего значения.

Визуализация данных позволила получить представление о распределении значений признаков и выявить потенциальные выбросы.

A graph with a bar and text

Description automatically generated with medium confidence

A graph with a bar and a number of numbers

Description automatically generated with medium confidence

A graph with a number of different colored squares

Description automatically generated with medium confidence

A graph of a number of people

Description automatically generated

**3. Реализация логистической регрессии**

Для построения модели логистической регрессии были реализованы следующие функции:

* **sigmoid(z):** Функция активации sigmoid, преобразующая линейную комбинацию признаков и весов в вероятность (значение от 0 до 1).
* **log\_loss(y\_true, y\_pred):** Функция потерь log loss (логарифмическая функция потерь), используемая для оценки качества модели во время обучения.
* **logistic\_regression(X, y, learning\_rate, num\_iterations):** Функция, реализующая алгоритм обучения логистической регрессии с помощью градиентного спуска. Функция принимает на вход матрицу признаков (X), вектор целевой переменной (y), коэффициент обучения (learning\_rate) и количество итераций обучения (num\_iterations). В результате работы функция возвращает вектор весов (w) и смещение (b), определяющие обученную модель.
* **predict(X, w, b):** Функция, выполняющая предсказание меток класса (0 или 1) для заданного набора данных с использованием обученной модели.

**4. Обучение и оценка модели**

Модель логистической регрессии была обучена на обучающем наборе данных с использованием различных комбинаций гиперпараметров:

* **learning\_rate:** 0.1, 0.01, 0.001, 0.0001
* **num\_iterations:** 100, 250, 500, 891

Для каждой комбинации гиперпараметров были рассчитаны следующие метрики качества модели на тестовом наборе данных:

* **Accuracy:** (Доля правильных ответов)
* **Precision:** (Точность)
* **Recall:** (Полнота)
* **F1-score:** (Среднее гармоническое Precision и Recall)

**5. Результаты**

Анализ результатов показал, что наилучшее качество модели было достигнуто при следующих гиперпараметрах:

* **learning\_rate = 0.1**
* **num\_iterations = 250**

При этих значениях модель продемонстрировала следующие метрики:

* **Accuracy: 0.8301**
* **Precision: 0.7120**
* **Recall: 0.8947**
* **F1-score: 0.7930**

**Заключение**

В ходе лабораторной работы была построена модель логистической регрессии для предсказания выживаемости пассажиров "Титаника".

**Основные выводы:**

* Предсказание выживаемости - комплексная задача, на которую влияют многие факторы.
* Выбор и предобработка данных играют ключевую роль в построении эффективной модели.
* Подбор оптимальных гиперпараметров позволяет значительно улучшить качество модели.

Дальнейшие исследования могут быть направлены на:

* Использование дополнительных признаков (например, информация о семье, номере каюты)
* Применение более сложных моделей машинного обучения (например, случайный лес, градиентный бустинг)
* Глубокий анализ важности признаков для интерпретации результатов модели.