Домашнее задание по курсу “Машинное обучение”

Задачи

Часть 1. Линейная регрессия

Датасет: https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset

Для приведенного выше датасета, построить модель линейной регрессии для оценивания стоимости аренды недвижимости. Для решения задачи предлагается использовать следующие подходы:

1. Получить решение задачи в замкнутом виде с помощью псевдообратной матрицы. Рассмотреть случаи без регуляризации и с L_2 регуляризацией.
2. Получить решение задачи методом градиентного спуска. Получить формулу для градиента, используя матрицу признаков. Сравнить полученные решение с п.1.

При решении задачи необходимо ответить на следующие вопросы:

1. Какие признаки оказывают наибольший вклад в точность определения стоимости аренды? Предложить способы отбора наиболее важных признаков
2. Какая модель имеет наименьшее значение функции потерь на тестовой выборке? Помогает ли регуляризация избежать эффекта переобучения в данном примере?

Часть 2. Задача классификации

Датасет: https://www.kaggle.com/competitions/titanic

Задачи:

А. Решение задачи классификации

1. Построить модель линейной классификации. При решении задачи методом градиентного спуска, необходимо вычислить градиент с помощью матрицы признаков.
2. Построить модель на основе случайного леса классифицирующих деревьев. Определить параметры классификатора (количество деревьев, максимальная глубина дерева), при которых точность классификации максимальна.

Б. Анализ ROC/PR-кривых

1. Рассмотреть простейшую задачу двухклассовой классификации, в которой элементы выборки имеют нормальное распределение с дисперсией 1 и средними значениями $+\mu$ для экземпляров одного класса и $-\mu$ для экземпляров

другого класса. Экземпляры каждого класса появляются в обучающей выборке с вероятностью ½. Для данной задачи необходимо определить оптимальный классификатор, а также построить ROC-кривую для различных значений μ.

2. Для двух моделей, полученных в п.А., построить ROC и PR (precision-recall) кривые. Получить значения площадей под этими кривыми. Какая метрика (ROC-AUC или PR-AUC) является более предпочтительной в данном примере и почему?

Часть 3. Задача кластеризации

Датасет: https://www.kaggle.com/datasets/arjunbhasin2013/ccdata
Для представленного датасета необходимо решить задачу кластеризации методом k-средних. При решении задачи необходимо ответить на следующие вопросы:

1. Как выбор начальных центров кластеров влияет на результат кластеризации
2. Какую метрику выбрать для оценки качества кластеризации?

Часть 4. Обучение нейронных сетей

Датасет: https://www.kaggle.com/competitions/digit-recognizer/data

Для представленного датасета: необходимо:

1. Решить задачу многоклассовой классификации с помощью многослойной нейронной сети, состоящей из полносвязных слоев.
2. Вычислить градиент целевой функции по обучаемым параметрам нейронной сети, сравнить полученные значения со значениями, вычисленными с помощью библиотечных функций
3. Определить параметры нейронной сети, при которых точность классификации максимальна

Требования к решению

Решение должно состоять из двух частей:

- исходного кода (рекомендуемый язык - Python)
- краткого отчета, содержащего вывод основных результатов

Критерии оценки

Каждая из четырех задач оценивается в 2 балла