Цели работы

- Изучить основные понятия машинного обучения (признаки, предположение IID, функции потерь, метрики, параметры и гиперпараметры) и их применение в задаче распознавания изображений.
- Научиться использовать простые классификаторы (kNN и наивный Байес) на примере датасета MNIST, реализуя их в Jupyter/Colab.
- Оценить на практике влияние нормализации признаков и выбора метрики расстояния на качество классификации.
- Исследовать, как нарушение предположения о независимости и одинаковом распределении данных влияет на работу алгоритмов.

Задание

В этом домашнем задании необходимо последовательно реализовать и проанализировать два простых метода классификации на примере задачи распознавания рукописных цифр (датасет MNIST). Работу следует выполнить в Jupyter Notebook или Google Colab. После каждого этапа реализованных экспериментов необходимо оформлять выводы и комментарии в Markdown-отчёте внутри тетрадки.

1. **Загрузка и исследование данных.** Загрузите датасет MNIST (например, через `sklearn.datasets` или другую библиотеку). Посмотрите на примеры изображений и статистику данных (размер выборок, число классов, размеры изображений). Определите, что вы используете в качестве признаков объекта.
    - Какие значения принимают отдельные признаки (например, интенсивности пикселей)? Нуждается ли набор признаков в нормализации?
    - Объясните, что в контексте задачи компьютерного зрения представляют собой признаки и зачем их возможно нормализовать.
2. **Постановка задачи.** Определите целевую переменную (что мы предсказываем) и выберите метрики качества классификации (например, точность, F1-меру и т.п.). Укажите и опишите функцию потерь, которой руководствуются выбранные метрики.

- Почему выбраны именно эти метрики для задачи MNIST?
- В чем разница между функцией потерь и метрикой качества в задаче классификации?

3. **Классификатор kNN.** Реализуйте или используйте готовую реализацию kNN-классификатора. Разбейте данные на обучающую и тестовую выборки.

   - Проведите классификацию без нормализации признаков и с нормализацией (например, стандартизация или масштабирование пикселей к интервалу $[0, 1]$). Сравните результаты. Как нормализация влияет на качество работы kNN?
   - Попробуйте несколько значений параметра k и разные метрики расстояния (например, евклидово и манхэттенское). Как это сказывается на точности классификации?
   - Какие параметры модели (например, веса признаков) и какие гиперпараметры (например, k, выбор метрики) вы использовали и настраивали в kNN? Приведите примеры.

4. **Наивный байесовский классификатор.** Реализуйте наивный байесовский классификатор (например, GaussianNB или BernoulliNB из `sklearn`). Опишите предположение о независимости признаков, которое делает этот метод.

   - Нужно ли нормализовать данные перед применением наивного байеса? Как масштабирование признаков повлияет на этот алгоритм?
   - Оцените качество классификации на тестовой выборке. Как результаты наивного байеса соотносятся с результатами kNN?

5. **Проверка предположения IID.** Организуйте эксперимент, в котором нарушается предположение о том, что обучающая и тестовая выборки имеют одинаковое распределение. Например, разделите данные так, что в обучении будут присутствовать далеко не все классы или искусственно измените распределение признаков.

   - Как изменяется качество (точность) работы kNN и наивного байеса при нарушении предположения о одинаковом распределении данных?

- Какой из методов (kNN или наивный байес) оказался более устойчивым к таким изменениям и почему?

6. **Итоги и выводы.** На основе проведенных экспериментов сформулируйте выводы о том, как нормализация признаков, выбор метрики и структура данных (IID vs не-IID) влияют на работу методов kNN и наивного байеса. Сравните эти два классификатора по качеству на задаче распознавания цифр.

Вопросы для отчета (объяснительная часть)

- Что называется признаком (feature) в задаче компьютерного зрения? Почему нормализация признаков важна для kNN?
- Что понимается под предположением *IID* и как оно используется при оценке алгоритмов машинного обучения? Как можно специально нарушить это предположение в экспериментах?
- Что такое функция потерь? Приведите пример функции потерь для задачи классификации. Как она связана с метрикой качества?
- Что такое параметры модели и гиперпараметры? Приведите примеры параметров и гиперпараметров для kNN и для наивного байеса.
- Чем метрика качества (например, точность) отличается от функции потерь, и почему мы смотрим на метрику при оценке модели?