

Машинное обучение: Классификатор Кто?

Предыстория

Иван обучил шесть различных моделей машинного обучения на наборе данных для классификации изображений: Дерево решений (Decision Tree), Метод k-ближайших соседей (K-Nearest Neighbors), Логистическая регрессия (Logistic Regression), Многослойный перцептрон (Multi-Layer Perceptron), Метод опорных векторов (Support Vector Machine) и Метод случайного леса (Random Forest).

Он оставил свой ноутбук разблокированным во время обеда. Элис, решив пошалить, перемешала названия файлов с предсказаниями и немного подкрутила гиперпараметры обучения.

— Элис! Теперь я не могу понять, какие предсказания от какой модели! — возмутился Иван.

— Ты же эксперт в ML, — ухмыльнулась Элис. — Разве ты не должен уметь определять это по их поведению?

Иван понял, что она права. Это интересная детективная задача — и идеальное задание для олимпиады.

Задача

Вам дан набор изображений `x_train.csv` вместе с их метками `y_train.csv`, на которых обучались 6 разных моделей.

Модели сделали свои предсказания на тестовом наборе `x_test.csv` (и `y_test.csv`), и эти предсказания сохранены в файлах `preds_0.csv`, `preds_1.csv`, ..., `preds_5.csv`. Используемые модели:

- Дерево решений (Decision Tree, DT)
- Метод k-ближайших соседей (K-Nearest Neighbors, KNN)
- Логистическая регрессия (Logistic Regression, LR)
- Многослойный перцептрон (Multi-Layer Perceptron, MLP)
- Метод опорных векторов (Support Vector Machine, SVM)
- Метод случайного леса (Random Forest, RF)

Подсказки

- Используйте `sklearn` для обучения моделей и получения предсказаний.
- Проверьте, какие модели дают наиболее схожие прогнозы с заданными файлами прогнозов.
- Вы можете использовать PCA из `sklearn` для уменьшения размерности данных для Логистической регрессии (LR) и Многослойного перцептрона (MLP).
- Если вы не знакомы с некоторыми моделями или с библиотекой `sklearn` в целом, мы рекомендуем начать с нашего основного руководства по введению в ML. Оно будет полезно для всех задач по машинному обучению.

Оценивание

Оценивание производится на основе точности определения моделей (например, 2 правильно найденные модели = 2/6).