Machine Learning: Classifier Who

Problem Story

Ivan trained seven different ML models on an image classification dataset: Decision Tree, K-Nearest Neighbors, Logistic Regression, Gaussian Process Classifier, Multi-Layer Perceptron, Support Vector Machine, and Random Forest.

He left his laptop unlocked during lunch. Alice, feeling mischievous, scrambled the prediction file names and tweaked his training hyperparameters.

"Alice! Now I can't tell which predictions came from which model!" Ivan protested.

"You're the ML expert," Alice grinned. "Shouldn't you be able to figure it out by their behavior?"

Ivan realized she had a point. This was actually an interesting forensics problem — and a perfect olympiad task.

Problem

You are given a dataset of images `X_train.csv` along with their labels `y_train.csv` that the 6 models were trained on.
The models made their predictions on the test set `X_test.csv` (and `y_test.csv`), and the predictions are stored in files `preds_0.csv`, `preds_1.csv`, ..., `preds_6.csv`. The models used are:

- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Logistic Regression (LR)
- Multi-Layer Perceptron (MLP)
- Support Vector Machine (SVM)
- Random Forest (RF)

Hints

- Use sklearn to train the models and make predictions.
- Check which models makes most similar predictions to given prediction files.
- Use PCA from sklearn to reduce dimensionality for LR and MLP.
- If you are not familiar with some models or with sklearn in general, we recommend starting with our main guide on getting started in ML. You will find it useful for all ML tasks.

Grading

Scored with accuracy in identifying the models (e.g. 2 model found correctly = 2/6).