Assignment 12

Applied Machine Learning

Credit card fraud costs about 1% of their revenue to the banks, an amount which customers (us) eventually pay. Let's find those anomalies which might reveal a fraud. Download the popular credit card dataset from Kaggle.

1. [10 pts] Explore the dataset, list number of rows and columns, check sanity, examine features (e.g. histograms/plots).

2. [10 pts] Check the class balance and pick an evaluation metric.

3. [10 pts] Check if you need normalization or standardization, and justify. Complete pre-processing.

4. [20 pts] Split the dataset 50-50 for training and testing. Then run `SVC`, `DecisionTreeClassifier`, `MLPClassifier`, `RandomForest` without any tree pruning or regularization. Report the classification performance. Then run `SVC`, `DecisionTreeClassifier`, `MLPClassifier` with tree pruning and regularization (Hint: might use `GridSearchCV` to optimize the regularization parameters; or simply run a few pilot tests). Report the classification performance.

5. [30 pts] Script a PyTorch neural network with a hidden layer (might experiment with 2 hidden layers, size might be 20 to 40). Report the classification performance on the previous 50-50 dataset. Expect a similar performance to the neural network in Q4.

6. [10 pts] Add dropout to the PyTorch neural network and repeat the previous step. Note that a robust model with a comparable performance to Q4. or Q5. is always preferred. Why?

7. [10 pts] Evaluate the 10-fold cross validation of Random Forest and two PyTorch neural network from Q5. and Q6. Comment about results.