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 the number of rows and columns, check sanity, and examine the features (e.g. with histograms or plots).

2. [10 pts] Check the class balance, choose an evaluation metric, and justify the choice.

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

4. [10 pts] Split the dataset 50-50 for training and testing. Then, without any tree pruning or regularization, run classifiers of the following types:
   - `SVC`
   - `DecisionTreeClassifier`
   - `MLPClassifier`
   - `RandomForest`

   Report each one's classification performance.

5. [10 pts] Now use tree pruning and/or regularization to run classifiers of the following types:
   - `SVC`
   - `DecisionTreeClassifier`
   - `MLPClassifier`

   (Hint: you might use `GridSearchCV` to optimize the regularization parameters, or simply run a few pilot tests). Report each one's classification performance. Make sure to use the same subsets as above to train and test.

6. [30 pts] Script a PyTorch neural network with a hidden layer. (You could also experiment with 2 hidden layers, with sizes between 20 and 40). Report its classification performance, using the same 50-50 subsets. (Expect a similar performance to the neural network in Q5.)

7. [10 pts] Add dropout to the PyTorch neural network and repeat the previous step. Note that a robust model, even with a performance comparable to Q5. or Q6.'s neural networks, is always preferred. Why?

8. [10 pts] Train a Random Forest classifier with 10-fold CV; revisit the two PyTorch neural network from Q6. and Q7. And train them with 10-fold CV as well. Comment on the results.