UBS Data Science Task

Ismet Ferdi SEN

Data Pre-processing

- Check whether there is missing value?
  - If yes eliminate them whether deleting if the number not to much
  - or impute them if keeping them more logical.
- Check whether the classes are evenly sampled?
  - If not try to duplicate the less one to reach a more balanced sampling.
- Split Data into Training and Validation parts, use **stratify** parameter to preserve proportion of classes balanced in the train and validation set.
- Perform Feature Normalization (Data standardization) using **only TRAINING DATA** *for both Training set and Validation Set*, to sure our model **generalize well** on new, unseen data.

Classifier preparation

- Define a function to train many classifier at once and to print their performance.

- Search on scikit-learn for different classifier.
- Chosen ones:
  - Linear SVM, SVM Classifier, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes

Results for Task-1 (with 130 features)

Classifier	Training-Set Accuracy	Validation-Set Accuracy
Linear SVM	84.64%	85.80%
SVM Classifier	**95.85%**	92.55%
Decision Tree	92.77%	91.30%
Random Forest	85.49%	81.95%
Neural Net	94.17%	**92.75%**
AdaBoost	92.10%	92.00%
Naive Bayes	82.93%	83.45%

Challenges

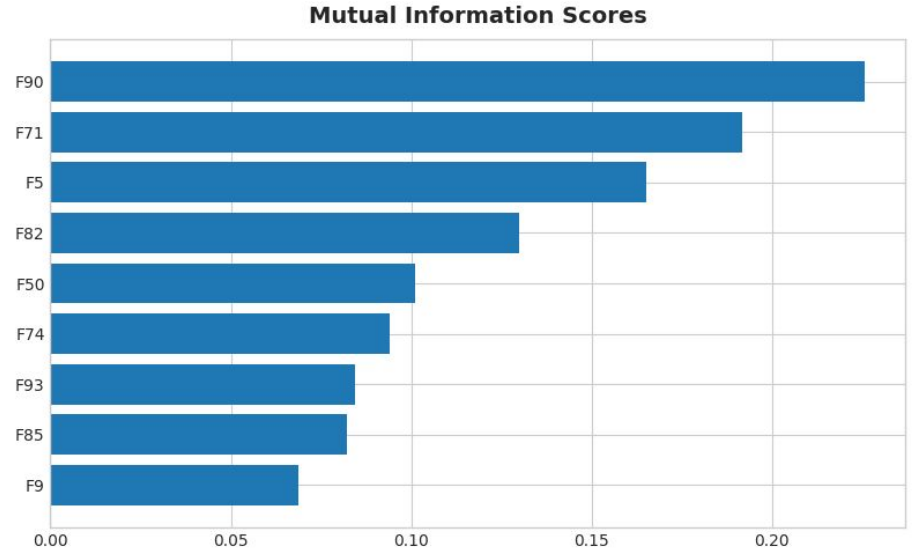
- Neural Network tend to be overfitted with default parameters.
  - Activated the ***early stopping*** parameter for Neural Network Classifier to prematurely stop the training at an optimal epoch. It solved the overfitting-problem.

Neural Network Classifiers	**Training-Set Accuracy**	**Validation-Set Accuracy**
Without early stopping	98.66%	91.55%
With early stopping	94.17%	92.75%

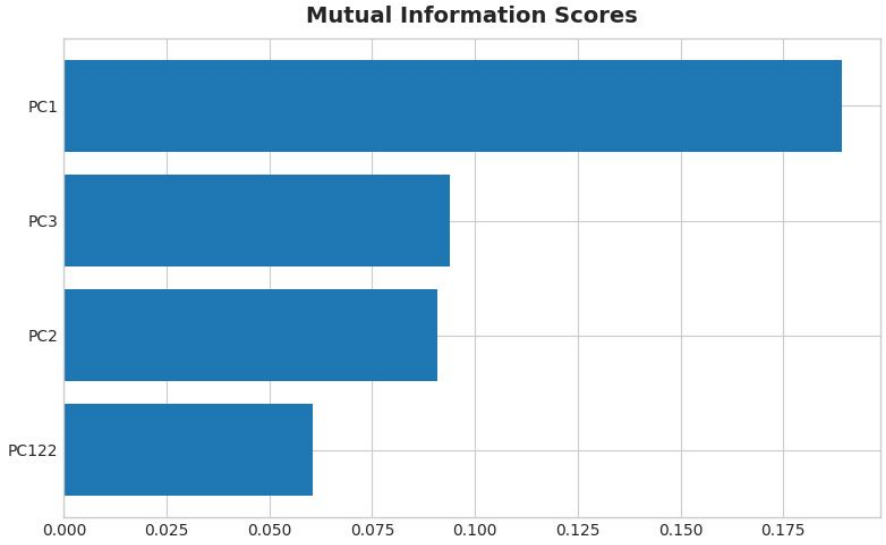
Feature Selection

- Measure associations between a feature and the target using Mutual Information. **Mutual Information (MI)** detect ***any*** kind of relationship, while correlation only detects ***linear*** relationships.
- **Selected 9 features** with the highest MI-scores had show better performance.
- In order to speed up the model, I reduced the number of hidden layers of the neural network as the number of feature sets decreased.
- Other than that, I created new features with **PCA** from the strongest features and checked their success. I got reasonable results with **only 4 principal component.**

Feature-Mutual Information and PCA-Mutual Information



Feature Set



Principal Components

Results for Task-2

Classifier	Validation-Set Accuracy Without feature selection (130 features)	Validation-Set Accuracy Feature selection with Mutual Information (10 features)	Validation-Set Accuracy Feature selection with PCA (4 PCs)
Linear SVM	85.80%	86.40%	83.10%
SVM Classifier	92.55%	**94.80%**	**92.60%**
Decision Tree	91.30%	92.05%	88.65%
Random Forest	81.95%	91.35%	90.10%
Neural Net	**92.75%**	92.45%	91.65%
AdaBoost	92.00%	92.60%	88.75%
Naive Bayes	83.45%	83.55%	88.10%

Thank you for giving me chance to show my skills.

Ismet Ferdi SEN