Credit Card

Name: Srinivas Kulkarni

Language Used: Python (Jupyter Notebook, Numpy, Scikit-Learn, Pandas)

* Given problem had ‘csv’ dataset file with 690 samples, 15 features and label column (which gave information whether credit card is approved or not)
* Some features were continuous and others were categorical. And in some feature column there were missing data values.
* As we cannot work with missing values, median of each column of continuous features were calculated, and missing values were replaced by calculated medians.
* Each continuous feature had different range and standard deviation, hence it was standardized by subtracting by mean and dividing by std\_dev.
* For categorical feature, missing values (‘?’) were replaced by unique values each column might take, considering the probability of occurrence of those unique value.
* Later categorical features were converted to binary type. So that it can be used for scikit-learn module.
* After data pre-processing, dataset was split into two. One for training and other for testing.
* SVM (support vector machine) classifier was used to classify the data.

1. As dataset is small, it’s hard to avoid overfitting, and noise due to outliers.
2. With small data set, we have to work using simple model. As more complex model leads to more weights and biases (which is hard to learn, considering small number of samples).
3. Using regularization in model helps to fight against overfitting, for which SVM does a good job.
4. SVM has less number of hyper parameter, C (regularization term) and kernel (choice of kernel). C sets the penalty for error i.e. to avoid misclassification. Kernel used was radial basis function.
5. Once we fit the training data, predict output for test data using trained model and calculate the accuracy.

* Accuracy obtained was: 0.907514450867.