Individual Exercise 2 - Linear Machine Learning Classifiers

Import the credit data file `/lecture2/CreditData.data` from the `ECON5130` repository `https://github.com/agarwalankush/ECON5130/tree/main/lecture3`. The credit dataset has three class labels : `0,1,2`. Class label `0` signifies the customer is not credit worthy, Class label `1` signifies the customer is credit worthy and Class label `2` signifies the customer is risky.

To evaluate the performance of a trained model on unseen data, split the dataset into separate training and test datasets. You will notice that the data is imbalanced. Use the `train_test_split` function from Lecture 2 to create the training and test dataset.

Once the training set is created, create a balanced training set with equal number of examples from Class label `0`, class label `1` and class label `2` using resampling, that is, randomly select examples from an under-represented class to create a subset with the same number of samples as the most represented class. Justify the need to create a balanced training set.

Once the balanced training dataset is created, perform standardization as the feature values are on different scales. Next, fit the perceptron learning rule, logistic regression model, support vector machine (SVM) classifier with `linear` and `rbf` kernels. Only use Scikit-learn implementations of the three models. Experiment with different model parameter values to obtain the best fit. - For the perceptron learning rule, experiment with different values of `eta0`. - For the logistic regression, experiment with different values of `C`. - For the SVM, experiment with different values of `gamma` and `C`.

Generate test accuracy scores using the three fitted models with best combination of model paramters and the plot respective decision boundaries. Repeat the procedure for different combinations of feature pairs.