Lab Course Machine Learning

Exercise Sheet 5

Prof. Dr. Dr. Lars Schmidt-Thieme
Kiran Madhusudhanan

HiWi: Basharat Mubashir Ahmed
Sophia Damilola Lawal
Submission deadline : December 9, 2022

General Instructions

1. Perform a data analysis, deal with missing values and any outliers.
2. **Use Scikit-Learn for this exercise sheet**.
3. Data should be normalized.
4. Train to Test split should be 80-20
5. Convert any non-numeric values to numeric values. For example you can replace a country name with an integer value or more appropriately use one-hot encoding.

1 High-Dimensional Data (5 points)

In the previous exercise sheets, you have implemented various algorithms from scratch; this exercise focuses on introducing you to a high-level Machine Learning library (sklearn).

1. Load the **20news-group-vectorized** dataset from sklearn. This dataset consists of 130107 predictors with 20 classes. Perform the following experiments:
   a) Multiclass Classification
   b) Use Logistic Regression Algorithm and perform
      i. one vs. rest
      ii. one vs. one
   c) Use Linear Discriminant Analysis Algorithm and perform
      i. one vs. rest
      ii. one vs. one
2. Report an accuracy of $\geq 55\%$ on the test set for each configuration.

2 Variable Selection via Forward and Backward Search (5 points)

Load the dataset **regression.npy**, the dataset consists of over 100 predictors. We generated the regression dataset such that only a few predictors are relevant. Perform the following experiments using the least angle regression algorithm

1. Forward Search
2. Backward Search

Print out the indices of the selected features, compare the outputs of the two methods. Are the indices the same?

3 Regularization (5 points)

Variable selection via forward and backward search drops some predictors; in some cases, we don't want to remove these predictors. Rather we want their coefficients to be small as possible. We are going to test the effect of the regularization term alpha. Try the following alpha values: [10, 1, 0.1, 0.0001, 0.00001], use **regression.npy** dataset

1. GridSearch
   a) Ridge regression
   b) Lasso
   c) Elastic-Net
2. RandomSearch
   a) Ridge regression
   b) Lasso
   c) Elastic-Net
3. Briefly discuss the effect of high and low values of alpha. Then, return the best three models with their respective alpha values using the appropriate metric
4. Compare the best three models to the models from question 2. What do you observe?
5. Get the indices of the top k coefficient of these three models and compare them with the features selected via the forward and back search method. Feel free to try different values of k

4 Custom Linear regression using Inheritance (5 points)

Sklearn is a very powerful high-level Machine learning API. It has a clean implementation of almost all the popular machine-learning algorithms. This task will introduce you to how to write a custom estimator in sklearn.

1. Create a python class called MyLinearRegression
2. Ensure your class inherit from sklearn BaseEstimator and RegressorMixin
3. Implement fit(X,Y) method, and returns self
4. Implement predict(X) method
5. Use check_estimator() method to know if your estimator(MyLinearRegression) is valid
6. Fit the dataset below using your custom estimator. Remember 80:20 split

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	6.0	4.83	3.7	3.15	2.41	1.83	1.49	1.21	0.96	0.73	0.64

Note: Your custom estimator should implement the simple-linear regression from Exercise sheet 1