Assignment 3: Support Vector Machine classifier

1 Instructions

- Use python programming language for your implementation.
- Use appropriate approach if you find some attribute is missing in your data.
- Report must contain step-wise description of your implementation and analysis of results. Since data analysis is a crucial task for any machine learning algorithm, report should demonstrate detailed analysis of results and conclusion. It should also clearly mention the steps to run your code.
- Learn the projection matrix for any dimension reduction technique using the train split only. Once the projection matrix has been trained using the train split, use that matrix to reduce the dimension of validation and test splits.
- You can use any python library function to complete the assignment.

2 Dataset:

Download Census-Income (KDD) Data Set:
https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)

3 Problem statement: Support Vector Machine classifier

In this assignment, you will learn to use several dimensionality reduction techniques to reduce the feature dimension of a data. Then you will train a SVM classifier on the reduced dimension feature space.

1. Read the dataset and randomly split it into train, validation and test part. The ratio of the train, validation and test splits should be $70 : 10 : 20$ respectively. **5 marks**

2. Reduce the feature dimension of the above data into a two dimensional feature space using Principle Component Analysis (PCA). Plot the reduced dimensional data of the train split in a 2d plane. In the plot, all data points of a single class should have same color and data points from different classes should have different colors. **15 marks**

3. Train an SVM classifier (sklearn.svm.SVC) on the reduced dimensional data generated from the step 2. Try different kernel type by varying the appropriate hyperparameters of the classifier and compute the classification accuracy on the validation split. Show the validation accuracy for each combination in a tabular form. Choose the kernel for which the validation accuracy is highest and compute the test accuracy using that kernel. Print the test accuracy. **20 marks**

4. Reduce the feature dimension of the above data into a one dimensional feature space using Linear Discriminant Analysis (LDA). Plot the reduced dimensional data of the train split. In the plot, all data points of a single class should have same color and data points from different classes should have different colors. **15 marks**

5. Repeat Step 3 on the data obtained from Step 4. **20 marks**

6. Is there any significant difference between the final test accuracy obtained from Step 3 and Step 5. If so, justify the results with proper reason. **5 marks**

7. Prepare a report clearly describing the process followed and showing the results of the above steps. **20 marks**