CS909/CS429 Data Mining 2023 Assignment 1: Classification

Version: 1.1

Due: **12 noon 15th February 2023 (UK Time).**

Grade: 25% of final grade

In this assignment, the objective is to develop a classification solution using classical machine learning methods. Specifically, we shall be solving an object recognition task. Each object is represented by a 28x28 dimensional image in a single 'flattened' 784 dimensional vector with an associated label (+1 or -1). The training data (with labels) and test data (without labels) are available to you at the URL https://github.com/foxtrotmike/CS909/tree/master/2023/A1

Xtrain: Training Data (each row is a single image)

Ytrain: Training labels

Xtest: Test Data (each row is a single image)

You will use Xtrain and Ytrain for training, validation and model selection. The data can be loaded with np.loadtxt.

Submission Requirements: You are expected to submit a **single Python Notebook** containing all answers and code. Include all prediction metrics in a presentable form within your notebook and include the output of the execution of all cells in the notebook as well so that the markers can verify the output. **Also submit a consolidated table of your performance metrics within the notebook to indicate which model performs the best (MANDATORY).** If you use any libraries other than sklearn, numpy, pandas and scipy, please include the code for its installation e.g., (!pip install xxx).

Your solution will be a single Ipython Notebook (with comments on your code) submitted through Tabula **together with a single prediction file for the test data in the format prescribed below**.

Do not use the solutions to previous years' assignments. Even though the questions in this assignment may appear to be similar to the ones from previous years' assignments, the expected answers are different as the dataset is significantly different.

Question No. 1: (Exploring data) [20% Marks]

Load the training and test data files and answer the following questions:

i. How many training and test examples are there? How many positive and negative examples are there in the training dataset?
ii. Show at least 10 randomly selected objects of each class using plt.matshow by reshaping the flattened array to 28x28. What are your observations about the nature of the data? Also show 10 randomly selected objects from the test set. Do you see any issues in the data that may limit the generalization performance of your classifier?
iii. Which performance metric (e.g., accuracy, AUC-ROC and AUC-PR) should be used for this problem? Give the reasoning behind your choice(s).
iv. What is the expected accuracy of a random classifier (one that generates random labels for a given example) for this problem over the training and test datasets? Demonstrate (either by a mathematical or statistical proof or a programming experiment) why this would be the case.

v. What is the AUC-ROC and AUC-PR of a random classifier for this problem over the training and test datasets? Demonstrate (either by a mathematical or statistical proof or a programming experiment) why this would be the case.

Question No. 2: (Nearest Neighbor Classifier) [15% Marks]

Perform 5-fold stratified cross-validation (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html) over the training dataset using a k=3 nearest neighbour (kNN) classifier and answer the following questions:

i. What is the accuracy, balanced accuracy, AUC-ROC and AUC-PR for each fold using this classifier? Show code to demonstrate the results. Calculate the average and standard deviation for each metric across all folds and show these in a single table. As the KNN classifier in sklearn does not support decision_function, be sure to understand and use the predict_proba function for AUC-ROC and AUC-PR calculations or plotting.

ii. Plot the ROC and PR curves for one fold. What are your observations about the ROC and PR curves?

iii. What is the impact of various forms of pre-processing (https://scikit-learn.org/stable/modules/preprocessing.html) (e.g., mean-standard deviation or standard scaling or min-max scaling) on the cross-validation performance? Show code to demonstrate the results and write a summary of your findings. Do any pre-processing techniques improve predictive performance? Why do you think this is the case?

Question No. 3: [20% Marks] CV

Use 5-fold stratified cross-validation over training data to choose an optimal classifier between: Perceptron, Naïve Bayes Classifier, Linear SVM and Kernelized SVM. Be sure to tune the hyperparameters of each classifier type (C and kernel type and kernel hyper-parameters for SVM etc). Report the cross validation results (mean and standard deviation of accuracy, balanced accuracy, AUC-ROC and AUC-PR across fold) of your best model. You may look into grid search as well as ways of pre-processing data.

i. Write your strategy for selecting the optimal classifier. Show code to demonstrate the results for each classifier.

ii. Show the comparison of these classifiers using a single consolidated table.

iii. Plot the ROC curves of all classifiers on the same axes for easy comparison.

iv. Plot the PR curves of all classifier on the same axes for comparison.

v. Write your observations about the ROC and PR curves.

Question No. 4 [20% Marks] PCA

i. Reduce the number of dimensions of the training data using PCA to 2 and plot a scatter plot of the training data showing examples of each class in a different color. What are your observations about the data based on this plot?

ii. Reduce the number of dimensions of the training and test data together using PCA to 2 and plot a scatter plot of the training and test data showing examples of each set in a different color (or marker style). What are your observations about the data based on this plot?

iii. Plot the scree graph of PCA and find the number of dimensions that explain 95% variance in the training set.

iv. Reduce the number of dimensions of the data using PCA and perform classification. What is the (optimal) cross-validation performance of a Kernelized SVM classification with PCA? Remember to perform hyperparameter optimization!

Question No. 5 [15% Marks]

Develop an optimal pipeline for classification based on your analysis (Q1-Q4). You are free to use any tools or approaches at your disposal. However, no external data sources may be used. Describe your pipeline and report your outputs over the test data set. (You are required to submit your prediction file together with the assignment in a zip folder). Your prediction file should be a single column file containing the prediction score of the corresponding example in Xtest (be sure to have the same order as the order of the test examples in Xtest!). Your prediction file should be named by your student ID, e.g., u100011.csv.

Question No. 6 [10% Marks]

Using the data given to you, consider an alternate classification problem in which the label of an example is based on whether it is a part of the training set (label = -1) or the test set (label = +1). Calculate the average and standard deviation of AUC-ROC using 5-fold stratified cross-validation for a classifier that is trained to solve this prediction task. What is the implication of this AUC-ROC value? Show code for this analysis and clearly explain your conclusions with supporting evidence.