CS909/CS429 2021 Assignment 1: Classification

(by Fayyaz Minhas)

Due: **12 noon 16th Februrary 2022 (UK time).**

Grades: 25% of final grade

In this assignment, the objective is to develop object classification solutions 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/2022/A1

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

Ytrain: Training labels

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

You will use Xtrain and Ytrain for training, validation and model selection.

Submission: 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 your output. **Also submit a consolidated table of your performance metrics within the notebook to indicate which model performs the best (MANDATORY).**

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**. Please do not use the solutions to previous years' assignments. Even though the questions may appear the same, the expected answers are different as the dataset is significantly different.

Question No. 1: (Showing data) [5 Marks]

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

i. How many training and test examples are there? You can use np.loadtxt for this purpose. Show at least 10 randomly selected objects of each class using plt.matshow.
ii. How many positive and negative examples are there in the training dataset?
iii. Which performance metric (accuracy, AUC-ROC and AUC-PR) should be used? Give your reasoning.
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 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 why this would be the case.

Question No. 2: (Nearest Neighbor Classifier) [5 Marks]

Perform 5-fold stratified cross-validation (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html) over the training dataset using the k = 1 nearest neighbour classifier and answer the following questions:

i. What is the prediction accuracy, AUC-ROC and AUC-PR for each fold using this classifier? Show code to demonstrate the results.
ii. What is the mean and standard deviation of each performance metric (accuracy, AUC-ROC and AUC-PR)across all the folds for this classifier? Show code to demonstrate the results.
iii. What is the impact of various forms of pre-processing (https://scikit-learn.org/stable/modules/preprocessing.html) on the cross-validation performance? Show code to demonstrate the results.
iv. Use 5-fold cross-validation over training data to calculate the optimal value of k for the k-Nearest neighbour classifier. What is the optimal value of k and what are the cross-validation accuracy, AUC-ROC and AUC-PR? Show code to demonstrate the results.

Question No. 3: [5 Marks] CV

Use 5-fold stratified cross-validation over training data to choose an optimal classifier between: k-nearest neighbour, Perceptron, Naïve Bayes Classifier, Logistic regression, Linear SVM and Kernelized SVM. Be sure to tune the hyperparameters of each classifier type (k for k-nearest neighbour, C and kernel type and parameters for SVM and so on). Report the cross validation results (mean and standard deviation of 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. Show code to demonstrate the results. Also show the comparison of these classifiers using a single table.

Question No. 4 [5 Marks] PCA

i. Reduce the number of dimensions of the data using PCA to 2 and plot a scatter plot of the training data. What are your observations about the data based on data?
ii. Plot the scree graph of PCA and find the number of dimensions that explain 95% variance in the training set.
iii. 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 [5 Marks]

Develop an optimal pipeline for classification based on your analysis (Q1-Q4). You are free to use any tools at your disposal. However, no external data sources may be used. Describe your pipeline and report your results 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!). Your prediction file should be named by your student ID, e.g., u100011.csv.