CS909/CS429 Data Mining 2023 Assignment 1: Classification

Heads Up, Future Innovators! 🌟

Guess what? This year's assignment might seem like a déjà vu from the past, but hold your horses! It's packed with fresh, unique data that's itching for your original solutions. That's right, no rehashed codes from yesteryears here.

Think of it as your launchpad into the exciting world of machine learning. Who knows? The skills you hone here might be your secret sauce in cracking complex challenges tomorrow - maybe you'll be the one to outsmart cancer using data analysis!

So, gear up and dive in with all your creativity. This isn't just an assignment; it's your stepping stone to making a real-world impact. Let's make learning not just insightful but downright fun. Ready to create some magic? 🎩✨

🎯 **Objective:**

Your mission, should you choose to accept it, is to craft a classic machine learning solution for an object recognition task. Picture this: each object is a 28x28 pixel image. You'll get these images as 'flattened' 784-dimensional vectors, each tagged with a label (+1 or -1).

Data Sources:

Training Data (Xtrain): Rows of images for you to train your model.

Training Labels (Ytrain): The label of each image.

Test Data (Xtest): More rows of images for you to test your model's savvy.

Grab this data from this URL: https://github.com/foxtrotmike/CS909/tree/master/2024/A1

Use Xtrain and Ytrain for training, validating, and selecting your model. You can load the data with np.loadtxt.

📝 **Submission Guide:**

Whip up a SINGLE Python Notebook containing all your code and answers.

Make sure it includes:

1. All prediction metrics, presented neatly.
2. The output of every cell executed, so markers can verify your work.
3. A summary table of performance metrics to spotlight the star model.
4. Stick to these libraries: sklearn, numpy, pandas, scipy. If you explore beyond these, include installation code (!pip install xxx).
5. Submit your solution as a single Ipython Notebook through Tabula, complete with comments explaining your code.
6. Also, turn in a prediction file for the test data, formatted as prescribed.

🚫 Important:

No recycling old solutions, please! This year's dataset is a whole new game compared to previous years, demanding fresh answers.

Question No. 1: (Exploring data) [10% Marks]

Start by loading the training and test data. Once you have it ready, let's explore with these questions:

i. **Dataset Overview**
   a. How many examples are in the training set? And in the test set?
   b. Within the training data, count the positive and negative examples. What's the distribution like? Does this distribution signify any potential issues in terms of design of the machine learning solution and its evaluation?

ii. **Visual Data Exploration**
   a. Pick 10 random objects from each class in the training data and display them using plt.matshow. Reshape the flattened 28x28 arrays for this. What patterns or characteristics do you notice?
   b. Do the same for 10 random objects from the test set. Are there any peculiarities in the data that might challenge your classifier's ability to generalise?

iii. **Choosing the Right Metric**
   Which performance metric would be best for this task (accuracy, AUC-ROC, AUC-PR, F1, Matthews correlation coefficient, mean squared error etc.)? Share your reasoning for this choice.

iv. **Benchmarking a Random Classifier**
   Imagine a classifier that randomly guesses labels. What accuracy would you expect it to achieve on both the training and test datasets? Show this through a calculation, statistical proof, or a coding experiment.

v. **Understanding AUC Metrics for Random Classifier**
   What would be the AUC-ROC and AUC-PR for a random classifier in this context? Again, support your answer with a mathematical or statistical argument, or a practical demonstration.

Question No. 2: (Nearest Neighbour Classifier) [10% 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-nearest neighbour (kNN) classifier and answer the following questions:

i. Start with a k = 5 nearest neighbour classifier. **Define and calculate** the accuracy, balanced accuracy, AUC-ROC, AUC-PR, F1 and Matthews Correlation Coefficient 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? What part of the ROC curve is more important for this problem?

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 or training speed? Why do you think this is the case?

Question No. 3: [20% Marks] Cross-validation of SVM and RFs

Use 5-fold stratified cross-validation over training data to choose an optimal classifier between: SVMs (linear, polynomial kernels and Radial Basis Function Kernels) and Random Forest Classifiers. Be sure to tune the hyperparameters of each classifier type (C and kernel type and kernel hyper-parameters for SVMs, the number of trees, depth of trees etc. for the Random Forests 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 in 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 classifiers 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 colour. 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 colour (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. You may want to select different principal components for the classification (not necessarily the first few). What is the (optimal) cross-validation performance of a Kernelized SVM classification with PCA? Remember to perform hyperparameter optimization!

Question No. 5 Optimal Pipeline [20% 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 Another classification problem [20% 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.

i. What does the value of this AUC-ROC tell you about any differences between training and test sets? Show code for this analysis and clearly explain your conclusions with supporting evidence.
ii. How can you use this experiment to identify and eliminate any systematic differences between training and test sets?