Machine Learning (PG)

Monsoon 2020

TOTAL MARKS: 160 | ASSIGNMENT 3 | DUE DATE: 10 NOV, 2020

Instructions:

(1) The assignment is to be attempted in groups.
(2) You can use only Python as the programming language.
(3) You are free to use math libraries like Numpy, Pandas, SciPy etc.; any library is allowed for visualizations; and utility libraries like os, pickle etc. are fine.
(4) Usage instructions regarding the other libraries is provided in the questions. **Do not use any ML module that is not allowed.**
(5) Create a '.pdf' report that contains your approach, pre-processing, assumptions etc. Add all the analysis related to the question in the written format in the report, **anything not in the report will not be marked**. Use plots wherever required.
(6) Implement code that is modular in nature. Only python (*.py) files should be submitted.
(7) Submit code, readme and analysis files in ZIP format with naming convention '**A3_groupno.zip**' (one submission per group). This nomenclature has to be followed strictly.
(8) You should be able to replicate your results during the demo, failing which will fetch zero marks.
(9) There will be no deadline extension under any circumstances. According to course policies, no late submissions will be considered. So, start early.

Save your best models using 'joblib' or 'pickle'. During demo, you must be able to load your saved models and replicate the reported results.

Question 1: (60 Points)

Use the CIFAR-10 dataset for all the experiments. Choose hyper-parameters in a *systematical manner*.

(1) (a) Perform PCA using sklearn on the dataset such that 90% of the total variance is retained - feature descriptor(1) **5 Points**
(b) Combine HOG and color histogram (must be implemented from scratch) on a whole ie., (hog + color_hist) - feature descriptor(2) **15 Points**

Now perform the following for both these feature descriptors.

(2) Visualize the 2D t-SNE plot. State your observations. **10 Points**
(3) Use GridSearchCV (cv=5) to find the best parameters (C, *kernel*, γ in case of gaussian kernel) of SVM using the train set. Report the accuracies (train, test) and the run-times on the best parameters obtained. State your observations (if any) on the obtained best parameters. **15 Points**
(4) Develop a new training set by extracting the support vectors from the SVM fitted in (3). Now fit another SVM with the new training set and report the accuracies(train, test). Compare the accuracies from (3) and (4). State your observations. **15 Points**

Question 2: (50 Points)

Dataset: 'dataset_a' attached with the assignment.
Divide the dataset into 80-20 ratio. The 20% is the test set. You can only use the *fit()* from the sklearn for the SVM. Other tasks are to implemented from the scratch using the attributes of the model created. You can write a class that can utilize the fit() of the sklearn, and other methods of the class (such as predict()) can use the resultant attributes to achieve the other tasks.

(1) Visualize the (whole dataset) and state your observations **5 Points**

(2) Use a linear SVM to classify this dataset. Perform the grid search for parameter C to obtain its optimal value. Report the accuracy on the test set. For three different values of C, visualize the decision boundary, margins and the support vectors. There will be three different plots, each for a different value of the C. Specifically, highlight the support vectors in the plots. State your observations from this analysis. **20 Points**

(3) Now use a SVM with RBF kernel to classify this dataset. Perform the grid search for parameters C and γ to obtain their optimal values. Report the accuracy on the test set. For three different values of C and γ, visualize the decision boundary, margins and the support vectors. There will be six different plots, each for a different combination of the C and γ. Specifically, highlight the support vectors in the plots. State your observations from this analysis. **20 Points**

(4) For the optimal values of the parameters obtained in (2) and (3), use the sklearn to make the predictions on the test set. Is there any deviation in the performance? **5 Points**

Question 3: (50 Points)

Dataset: 'dataset_b' attached with the assignment. Use five fold cross validation.
You can only use the *fit()* from the sklearn for the SVM. Other tasks are to implemented from the scratch using the attributes of the model created. You can write a class that can utilize the fit() of the sklearn, and other methods of the class (such as predict()) can use the resultant attributes to achieve the other tasks.

(1) Visualize the (whole dataset) and state your observations **5 Points**

(2) Use a SVM with RBF kernel to classify this dataset through *one vs rest* approach. Perform the grid search for parameters C and γ to obtain their optimal values. Report the accuracy over the five folds along with the mean accuarcy. Also, report the mean class accuracy. **20 Points**

(3) Use a SVM with RBF kernel to classify this dataset through *one vs one* approach. Perform the grid search for parameters C and γ to obtain their optimal values. Report the accuracy over the five folds along with the mean accuarcy. Also, report the mean class accuracy. **20 Points**

(4) For the optimal values of the parameters obtained in (2) and (3), use the sklearn to make the predictions on the test set. Is there any deviation in the performance? **5 Points**

The OVO and OVR also have to implemented from the scratch.