Machine Learning (PG)

Monsoon 2020

TOTAL MARKS: 100 ASSIGNMENT 1 DUE DATE: 27 SEPT, 2020

Instructions:

(1) The assignment is to be attempted individually
(2) You can use only Python as the programming language.
(3) You are free to use math libraries like Numpy, Pandas; and use Matplotlib, Seaborn library for plotting.
(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 whereever 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 '**A1_rollno_name.zip**.' 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.

---

(1) In this question, you will explore the data visualization, a very essential task for the machine learning.
Datasets: 'dataset_1', 'dataset_2' (attached with the assignment)
You can use the pyplot or other related libraries for this question. You have to include the legends in the plots to denote the class labels and other relevant information.
(a) Visualize the 10 samples of each class from 'dataset_1' in the form of images. What are your observations? **5 marks**
(b) Usually to explore the data complexity, we visualize the scatter plot of the data. In the scatter plot, the samples of all the classes are visualized simultaneously, which provides information about the class separation. Visualize the 'dataset_2' in the form of scatter plot. What is your inference about this dataset? **10 marks**
(c) We can visualize only the 2D or 3D data using the scatter plots. For features dimensions higher that three, we may use T-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the number of features. Use the t-SNE to reduce the 'dataset_1' to 2 dimensions and visualize the scatter plot. What is your inference regarding the class separation ? **10 marks**
(d) Now, reduce the original 'dataset_1' to the three dimensions and visualize the scatter plot again. Is there any important distinction in inference as compared to (c) above. **10 marks**
You can use the t-sne from the sklearn library for part (c) and (d).

(2) For this question, you can use the decision tree classifier from sklearn.
Dataset: 'dataset_2' (attached with the assignment).
Use first 70% of the samples for training and remaining 30% for testing. Implement the function to split the dataset and calculate accuracy. You cannot use any inbuilt version. Though, you can use Numpy, Pandas, Random etc.
(a) Take depth as a hyperparameter, and perform a grid search for finding its optimal value. Plot a curve between depth and testing accuracy. Comment on the effect

of depth on the performance of the classifier. You have to perform grid search for at least 15 values of the depth. Is your best performance consist ant with the visualization in (1.b). You have to implement grid search and cannot use any inbuilt implementation for the algorithm. **15 marks**

(b) For part (a), prepare a table representing train accuracy and validation accuracy for each value of the depth. Comment on overfittimg, and underfitting for each entry in the table. **10 marks**

(c) Replicate part (b) with sklearn 'accuracy function'. Is there any deviation between the results from your implementation and inbuilt function? **5 marks**

(3) For this question, you can use the decision tree classifier from sklearn.
Dataset: PM2.5 Data UCI Archive
Target Variable: Month
You will have to handle null values in the data.
(Remove "No" column as that is index. This information is shared as this might be the first ML model for many people. From next assignment, data analysis and feature selection will be a part of the exercise.)
Split the data into training and testing set (80:20 ratio) using the function you created in previous question. Use the same training set for training the following models. You can not use sklearn for splitting the dataset.

(a) Train a decision tree using both gini index and entropy. Don't change any of other default values of the classifier. In the following models, use the criteria which gives better accuracy on test set. **5 marks**

(b) Train decision trees with different maximum depths [2, 4, 8, 10, 15, 30]. Find the best value of depth by using testing and training accuracy. Plot the curve between training and testing accuracy and depth to support your analysis. **10 marks**

(c) Ensembling is a method to combine multiple not-so-good models to get a better performing model (more in upcoming lectures). Create 100 **different** decision stumps (max depth 3). For each stump, train it on randomly selected 50% of the training data, i.e., select data for each stump separately. Now, predict the test samples' labels by taking take majority vote of the output of the stumps. How is the performance effected as compared to part (a) and (b)? **10 marks**

(d) Now, try to tune the decision stumps by changing the max-depth [4, 8, 10, 15, 20, best achieved from (b)] and number of trees. Analyze the effect on the training and testing accuracy. Use majority vote for final prediction on the test data. **10 marks**

Compare the results of the classification models created above on the test set. Rank the models and analyze if there is a statistically significant difference.
Add all the analysis to the report.