Machine Learning (PG)

Monsoon 2020

Total Marks: 80 — Assignment 4 — Due Date: 18 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, sklearn, 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 '**A4_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.

Question 1: KMeans Algorithm

Use the IRIS dataset for this question.

(1) Load the dataset and perform splitting into training and validation sets with 70:30 ratio. **5 Points**
(2) Implement the Kmeans algorithm using sklearn. You need to find the optimal number of clusters using the **elbow method**. Plot the error vs number of clusters graph while using the elbow method. Report the optimal number of cluster found. **25 Points**
(3) Use Scatter plot to visualize the dataset to depict the clusters formed(optimal). **10 Points**
(4) Report the training and the validation accuracy. Comment on the accuracy obtained for both the sets. **10 Points**

Question 2: Naive Bayes

For this question, use the yelp sentiment ananlysis dataset available here.

(1) Load the dataset. Split the dataset using sklearn's stratify split into 70:30 ratio.
(2) Preprocess the dataset by -
(a) Removing punctuation signs
(b) Lowercasing all words
(c) Removing stopwords (use nltk library)
5 Points
(3) Create a vocabulary of unique words from the training set. Use this vocabulary to design word count feature matrices where the (d,w) entry corresponds to the number of occurrences of word w in document d. The feature matrices should be separate for the train and validation sets. **10 Points**
(4) Implement the multinomial Naive Bayes Algorithm using the sklearn library. Apply *add-1 smoothing.* **10 Points**

(5) Report the training and validation accuracy. Give some examples of the misclassified samples and comment as to why they may have been misclassified. **5 Points**