Machine Learning (PG) M2021
Assignment-4

Max Marks: 130 **Due Date**: 21/11/2021, 11:59 PM

Instructions

- The assignment is to be attempted in groups of atmost 2 students.
- You can only use python as the programming language.
- 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.
- Usage instructions regarding the other libraries is provided in the questions. **Do not use any ML module that is not allowed.**
- 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 where ever required. **Submit code, readme and report files in ZIP format with naming convention** `A4_rollno1_name1_rollno2_name2.zip`. This nomenclature has to be followed strictly. (If only one member: `A4_rollno_name.zip`)
- Remember to **turn in** after uploading on Google Classroom. No justifications would be taken regarding this after the deadline.
- Start the assignment early. Resolve all your doubts from TAs in their office hours **two days before the deadline.**
- You should be able to replicate your results during the demo, failing which will fetch zero marks.

---

1. (30 points) **Naive Bayes**
   Implement Naive Bayes (Use sklearn) using this dataset. Use 80:20 train-test ratio.
   - Load and preprocess the dataset. Mention all the preprocessing steps in the report. You can use nltk library here, if required. **(5 Points)**
   - 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)**
   - Implement the multinomial Naive Bayes Algorithm using the sklearn library. Apply add-1 smoothing. **(10 points)**
   - Report Accuracy, F1 score for training and test set. **(5 points)**

2. (40 points) **KNN**
   Use the this dataset for this question.

   - Load the dataset and perform splitting into training and test sets with 75:25 ratio. Use tsne plot to visualise the dataset. **(5 points)**
   - Implement the kNN algorithm from **scratch** . You need to find the optimal number of k using the grid search. You may use sklearn for grid search. Plot the error vs number of neighbours graph (k). Report the optimal number of neighbours. **(30 points)**
   - Report the training and the validation accuracy only with optimal value of k using sklearn kNN function. Comment on the accuracy obtained for optimal value of k for both the methods i.e, your implementation and the inbuilt sklearn function. **(5 points)**

3. (60 points) **KMeans**

   - **Load the dataset and visualize through tsne. (5 points)**
   - Implement the Kmeans algorithm from **scratch**. 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. **(35 points)**
   - Use Scatter plot to visualize the dataset to depict the clusters formed(optimal). **(10 points)**
   - Report the training and the test set accuracy. Comment on the accuracy obtained for both the sets. Compare with sklearn. **(10 points)**