CS5512 Machine Learning

August-December 2021

Programming Assignment for ML Theory

Follow the instructions carefully before attempting:

- You must submit your code in a single python .ipynb notebook with naming format as follows: Firstname_Lastname_programming_assignment_2.ipynb
- For each question, create a separate text block containing the question followed by a code block containing the solution.
- Follow each and every instruction given in each question carefully.
- Your code must be properly commented explaining each step clearly.
- If any of the above instructions are not followed, penalty will be there for the same
- Your code and answers will be checked for plagiarism and if found plagiarised, zero marks will be provided for this assignment. So make sure you actually code and solve the questions rather than noting down the answers.

Task-1:[3 marks]

Use make_blobs() to generate 100 datapoints with 5 centers and implement K means algorithm from scratch(do not use Scikit learn). Also show the optimal number of clusters using ELBOW method.

Task-2:[4 marks]

Load the 'dataset_facebook.csv' file using the link https://archive.ics.uci.edu/ml/datasets/Facebook+metrics, go through the abstract(including Data Set Information, Attribute Information).

Perform necessary pre-processing steps to clean this dataset and subsequently implement dimensionality reduction(with, n_components = 2) paradigm using KernelPCA :

1. a ‘polynomial’ kernel of degree = 3
2. a ‘gaussian’ kernel of gamma value as 0.05

Report the results.

Task-3:[3 marks]

Apply DBSCAN on the dataset given in data.csv to group the data into clusters and also predict the outliers(noise points). Plot in 2D the clusters obtained along with the noise points for a clear visualization.

Add a column “Outliers” in data.csv which will contain the value -1 for noise points and value 0 otherwise and submit this .csv file also.