Lab - 3
CSL2010: Introduction To Machine Learning
AY 2022-23
Bayes Classification, KNN

General Instructions

1. You need to upload a zip ***<Lab3_Your_Roll_No>.zip***, which contains one file for the task in ***<Lab3_Your_Roll_No>.py*** format and the report for the entire assignment in ***<Lab3_Your_Roll_No>.pdf*** format.
2. Provide your colab file link in the report. **Make sure that your file is accessible**.
3. Submit a single report mentioning your observations for all the tasks.
4. Report/Cite any resources you have used while attempting the assignment.
5. Attempt either Q2 or Q3 during the lab.

Q1) Data pre-processing [10 marks]

a) Download the dataset from the link below. https://drive.google.com/file/d/1s2IhEwbbSAGEtVuPpLwLq_wQ8P3Svf0l/view?usp=sharing
(The 'Outcome' column is the target). Use this dataset for both Q2 and Q3.

b) Perform data preprocessing steps according to the previous lab (remove Null/NaN values, reduce noise, normalize/scale the features etc.). [5 marks]

c) Split the dataset into training-validation-test splits in the ratio of 70:20:10. [5 marks]

Q2) kNN [30 marks]

a) Using sklearn library, perform KNN classification (take n_neighbours=4). You may use any distance/similarity function of your choice. [5 marks]

b) Vary the value of n_neighbours (k) from 4 to 10, and calculate the classification accuracy score for each value of k. Discuss in your report the significance of k and how the value of k affects the classification accuracy. [10 marks]

c) Implement KNN from scratch (without using sklearn library, although the use of NumPy and pandas is allowed) using the same distance/similarity function as before. For different values of 'k', compare its results with those obtained in part (e). [15 marks]

Q3) Bayes Classification [30 marks]

a) Identify the two real-valued variables/features which have the largest standard deviation among all the features. Create a smaller dataset using these two features and the target column “outcome” for use in the next parts [5 Marks].

b) For each of the features obtained in the previous part, you are required to predict the target variable using only one variable at a time and using the Bayes Classification method. For a given feature, you need to choose an appropriate bin-size to discretize the continuous variables into bins (of histogram) such that the classification accuracy is maximum. [15 Marks].

c) Compare the results obtained in part (b) with those obtained in Q2. [10 Marks]

Resources:

- https://scikit-learn.org/stable/modules/naive_bayes.html
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://towardsdatascience.com/knn-using-scikit-learn-c6bed765be75