CSE / ECE 343/543:
ML ASSIGNMENT 3

Due Date: 23/11/2021. 11.59 PM

Max Marks: 50

Instructions:

• You are allowed to discuss but the final answer should be your own. Keep collaborations at high level discussions. Copying/Plagiarism will be dealt with strictly.

• Late submission penalty: As per course policy.

• Your submission should be a single zip file Admission No_HW3.zip. Including only the relevant files arranged with proper names. A single .pdf report explaining your codes with details of EDA and pre-processing, relevant graphs, visualization and solution to theory questions. Anything not in the report will not be considered for evaluation. The structure of submission should follow:
Admission No HW3

|-Q1.py |- | Report.pdf | Weights (folder) |- plots (folder)

• Restrict to using only Python for coding assignments.

• You are free to use math libraries: Numpy, Pandas; and use Matplotlib library for plotting.

• Use of inbuilt function for any evaluation metric is not allowed. Each of the metric needs to be implemented from scratch.

• Remember to turn in after uploading on Google Classroom. No excuses or issues 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.

• Document your code. Lack of comments and documentation would result in loss of 20% of the obtained score.

• The assessment will be done on basis of the following components: - Working codes

- Analysis and clarity of results (drawing comparisons across different parts) clarity of the report
- Understanding the theoretical concepts/viva

Q1

Feature selection, Feature extraction, and Clustering (40 marks)
In this question, you will apply unsupervised learning techniques to identify the segment of the population that has greater than 50k earnings per year. You are given two sets of data, Dataset 1. Population.csv contains information about the general population, while Dataset 2. more_than_50k.csv contains the same information about the people who are making more than 50k per year. You are required to do feature selection, feature extraction, and clustering. The goal of the problem is to cluster data in the general population and more_than_50k population and analyze which clusters are over-represented in the general population vs the more_than_50k population and vice versa.
1. For this part, you are allowed to use any library at your convenience.
2. Submit a well-organized IPython/Jupyter Notebook as per sections given in the steps you are required to do. After each section report observation and analysis for the given section. The submitted notebook will act as the report for this question.
3. Do submit the IPython/Jupyter Notebook exported in .py format too.

Dataset Folder Contains:
1. population.csv : General Population Data
2. more_than_50k.csv : Dataset for Population having more than 50k Annual Income
3. Data Description.csv : Contains description for the features in the dataset.

In this question, you will apply unsupervised learning techniques to identify the segment of the population that has greater than 50k earnings per year. You are given two sets of data, Dataset 1. Population.csv contains information about the general population, while Dataset 2. more_than_50k.csv contains the same information about the people who are making more than 50k per year. You are required to do feature selection, feature extraction, and clustering. The goal of the problem is to cluster data in the general population and more_than_50k population and analyze which clusters are over-represented in the general population vs the more_than_50k population and vice versa.

Steps you are required to do:
On Dataset Population.csv

1. Preprocessing **[2+2 marks]**
    1.1. Replace missing data with NaN where the missing data is marked by '?'
    1.2. Perform an assessment of how much missing data there is in each column of the dataset, based on that remove columns with more than 40% data missing.

2. Feature Analysis **[2+2 marks]**

2.1. Plot histogram of values for each feature (both categorical as well as numerical features)

2.2. Drop features in which most of the data is in one column and there is almost no data in the remaining columns, for example, feature 'GRINST'. Make sure you also convert numerical data to categorical data using bins for better analysis in later parts.

3. Imputation, Bucketization, One-Hot Encoding **[2+2+2 marks]**

3.1. Replace missing values in each column with mode for the column, make sure you store mode for each feature as you will need to replace missing features in the more_than_50k dataset with the same values.

3.2. Bucketize Numerical features

3.3. One hot encode features

4. Clustering **[5+3+3 marks]**

4.1. Apply K-median clustering with varying values of k in the range [10,24] and draw avg within-cluster distance vs a number of clusters graph.

4.2. Based on the elbow in the graph, choose the best value for k

4.3. Apply K-median clustering with the best value chosen above.

5. Handling more_than_50k data **[5 marks]**

5.1. Apply all the steps you did on general population data to more than 50k population data. While doing this make sure you don't perform operations on this data which do not align with operations done with population data.

6. Compare more_than_50k data with Population Data **[3+3+3+1 marks]**

6.1. Compare the proportion of data in each cluster for the more_than_50k data to the proportion of data in each cluster for the general population.

6.2. Find out which clusters are over-represented in the general population vs more_than_50k population and vice versa

6.3. What kinds of people are part of a cluster that is overrepresented in the more_than_50k data compared to the general population? For this, you may need to inverse transform PCA to map to original features and then analyze the value of centroid of the clusters. You may use features that have the highest magnitude for the first principal component to analyze the values for the centroid.

6.4. Similarly analyze a cluster that is overrepresented in the more_than_50k data compared to the general population?

Note: It is advised to make extensive use of Pandas and Numpy

Q2

SVM (Theory): **[1+1+3 marks]**

Q3. If we have non-linearly separable data, we need to modify the SVM algorithm by introducing an error margin that must be minimized. You are required to use an l2 norm soft margin SVM. The optimization problem for the same is:

$$min_{w,b,\epsilon} \; 1/2||w||^2 + C/2 \sum_{i=1}^{m} \epsilon^2$$

$$s.t.\; y^{(i)}(w^T x^{(i)} + b) \geq 1 - \epsilon_i, \quad i = 1,..., m$$

a. Should there be a non-negativity constraint on ϵ, i.e. $\epsilon \geq 0$ for the l2 norm soft margin SVM optimization problem? Why or why not?
b. Find the Lagrangian of the above optimization problem.
c. Find the dual of the above optimization problem.

Q3

SVM (Theory) **[2+3 marks]**

You are given the task of training an SVM machine using the Gaussian Kernel, K(x, z) = exp(−‖x −z‖2/τ 2). We wish to show that as long as there are no two identical points in the training set, we can always find a value for the bandwidth parameter τ such that the SVM achieves zero training error.

a) The decision function learned by the SVM can be written as,

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x) + b$$

Assume that there is a minimum separation of distance ϙ between the points of training data {(x(1), y(1)), . . . , (x(m), y(m))}. That is, |x(j) −x(i)|| ≥ ϙ for any i ≠ j. You need to find the values for the set of parameters {α1, . . . , αm, b}and Gaussian kernel width τ such that x(i) is correctly classified for all i = 1, . . . , m. [Hint: Let αi = 1 for all i and b = 0. Now notice that for y ∈ {−1, +1} the prediction on x(i) will be correct if |f (x(i)) −y(i)|< 1, so find a value of τ that satisfies this inequality for all i.]

(b) Suppose we run an SVM with slack variables using the parameter τ you found in part (a). Will we obtain zero training error on the resulting classifier? Why or why not? A short explanation (without proof) will suffice.