CS 5525 Assignment 5

Due 3rd Dec 2020

Problem 1. K-means clustering **[20 points]**

Consider the following set of one-dimensional data points: {0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9}.

1) Suppose we apply k-means clustering to obtain three clusters, A, B, and C. If the initial centroids are located at {0, 0.25, 0.6}, respectively, show the cluster assignments and locations of the centroids after the first three iterations.
2) Compute the SSE of the k-means solution (after 3 iterations).
3) Apply bisecting k-means (with k=3) on the data.
   First, apply k-means on the data with k=2 using initial centroids located at {0.1, 0.9}.

	Cluster assignment of data points							Centroid	
Iter	0.10	0.20	0.40	0.50	0.60	0.80	0.90	A	B
0	-	-	-	-	-	-	-	0.10	0.90
1									
2									

   Next, compute the SSE for each cluster (make sure you indicate the SSE values in your answer). Choose the cluster with larger SSE value and split it further into 2 sub-clusters. You can choose the two data points with the smallest and largest values as your initial centroids. For example, if the cluster to be split contains data points (0.20, 0.40, 0.60, and 0.80), then the centroids should be initialized to 0.20 and 0.80. Show the clustering solution produced obtained applying bisecting k-means. (Note: In case of a tie, assign to the larger centroid.)
4) Compare the results of k-means clustering against bisecting k-means. Which clustering method is more effective for the given data set?

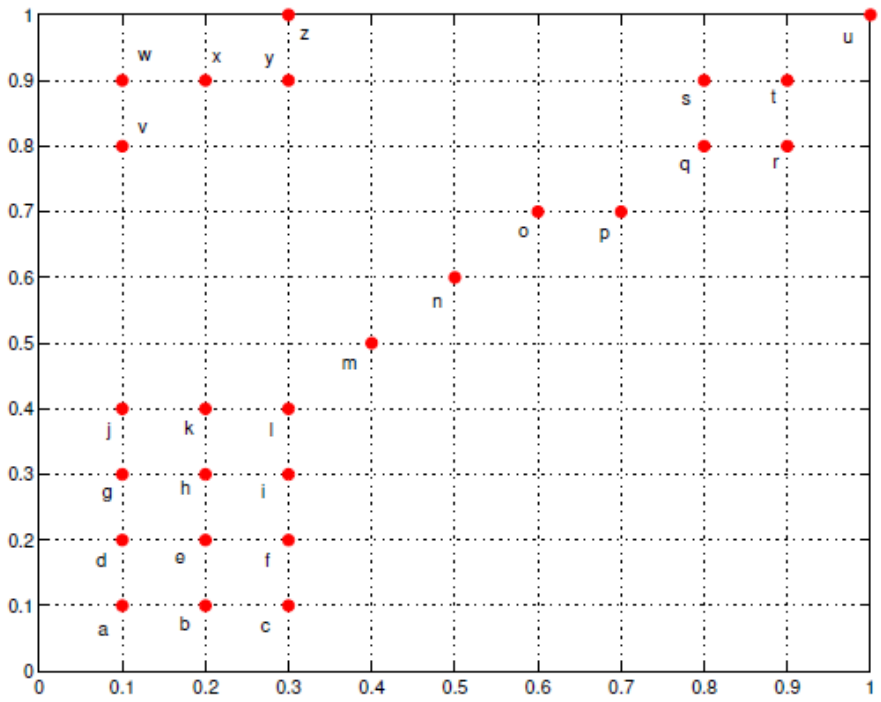
Problem 2. Dendograms **[15 points]**

Use the distance matrix shown in the table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged and the y-axis show the distance between pairs of clusters being merged at each iteration.

	p1	p2	p3	p4	p5
p1	0	0.5840	0.1955	0.3815	0.1127
p2	0.5840	0	0.6132	0.4956	0.5733
p3	0.1955	0.6132	0	0.2390	0.3067
p4	0.3815	0.4956	0.2390	0	0.4694
p5	0.1127	0.5733	0.3067	0.4694	0

Problem 3. DBSCAN Clustering **[15 points]**

Consider the data set shown in Figure 1. Suppose we apply DBSCAN algorithm with Eps=0.15 (in Euclidean distance) and MinPts=3.



1) List all the core points in the diagram (you can use the labels of the data points in the diagram). Note: a point is considered a core point if there are more than MinPts number of points (including the point itself) within a neighborhood of radius Eps.
2) List all the border points in the diagram.
3) List all the noise points in the diagram.
4) Using the DBScan algorithm, how many clusters will be obtained from the data set?

Problem 4. **[20 points]**

Consider the confusion matrices for two clustering solutions as shown below, where the rows correspond to the clusters and the columns correspond to the ground truth classes. Note that solution 2 simply partitions the first cluster of solution 1 into two smaller sub-clusters.

	Solution 1	
	Ground truth class	
	Class 1	Class 2
Cluster 1	40	20
Cluster 2	10	30

	Solution 2	
	Ground truth class	
	Class 1	Class 2
Cluster 1	35	15
Cluster 2	5	5
Cluster 3	10	30

Each entry n_{ij} in the matrix corresponds to the number of data points assigned to cluster i that belong to class j. Furthermore, let $n_{i+} = \sum_j n_{ij}$ (i.e., the sum of all entries in row i) be the number of points in cluster i, $n_{+j} = \sum_i n_{ij}$ (i.e., the sum of all entries in column j) be the number of data points that belong to class j, and $N = \sum_{ij} n_{ij}$ (i.e., the sum of all entries in the table) be the total number of data points. In this exercise, you will compare the performance of the two clustering solutions using the following measures:

- Entropy, $e = \sum_i \frac{n_{i+}}{N} e_i$, where $e_i = -\sum_j \frac{n_{ij}}{n_{i+}} \log \frac{n_{ij}}{n_{i+}}$ is the entropy of cluster i
- Purity, $p = \sum_i \frac{n_{i+}}{N} p_i$, where $p_i = \max_j \frac{n_{ij}}{n_{i+}}$ is the purity of cluster i
- Normalized mutual information

$$NMI = \frac{2 \sum_{i,j} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{n_{i+} n_{+j}}}{H_1 + H_2}$$

where $H_1 = -\sum_i \frac{n_{i+}}{N} \log \frac{n_{i+}}{N}$, and $H_2 = -\sum_j \frac{n_{+j}}{N} \log \frac{n_{+j}}{N}$.

Answer the following questions:

1) Compute the values of entropy, purity, and NMI when the clusters are pure (i.e., contains only data points from one class). Assume number of clusters is the same as number of classes (k = 2).
2) Compute the entropy for both solutions. Which solution is better?
3) Compute the purity for both solutions. Which solution is better?
4) Compute the NMI for both solutions. Which solution is better?
5) Based on your answers above, state which supervised measure do you think is better and why?

Problem 5. K-means Clustering **[15 points]**

In this problem, you will use Python and Scikit-Learn package to write a simple program for text clustering by following these steps:

1) Pre-process and load text data to your program.
   Please sample 300 pieces of news from 20newsgroups dataset.
   If you are using scikit-learn package, you do not have to load the data manually.
2) Extract features from a text corpus.
   This step will convert text to numerical vectors. You can use TfidfVectorizer.
3) Use scikit-learn to build a k-means clustering model and train it on your data sample.
   Set the number of clusters to 5.
4) Print top representative terms per cluster.
5) Find one piece of news from any website and predict which cluster it belongs to.

Additional notes:

Jupyter notebook is recommended.
Include your codes and intermediate output of each step (Less than 3 pages) in your report.
If you use the public/open source codes, please provide the links to the sources.

Sources:

https://pythonprogramminglanguage.com/kmeans-text-clustering/
http://jonathansoma.com/lede/algorithms-2017/classes/clustering/k-means-clustering-with-scikit-learn/
https://scikit-learn.org/stable/modules/feature_extraction.html
https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

Problem 6. Regression Analysis **[15 points]**

The Energy Efficiency Dataset (ENB2012_data.xlsx) in this problem is a public dataset from UCI (https://archive.ics.uci.edu/ml/datasets/Energy+efficiency). In this data, the energy analysis is performed using 12 different building shapes simulated in Ecotect. They are different in several aspects, such as including the glazing area and the orientation. In the dataset, there are 768 samples and 8 features. In this problem, we will use 7 features, and the goal is to predict the cooling load (Y), which is converted to a binary variable in this problem.

1) Fit a multivariate linear regression model using 5-fold cross validation to predict the cooling load Y.
2) Fit a logistic regression model using 5-fold cross validation to predict the cooling load Y.
3) Provide the confusion matrix along with the mean squared error (MSE) or Accuracy to evaluate the performance of each model.

Additional notes:
Jupyter notebook is recommended.
Include your codes and intermediate output of each step (Less than 3 pages) in your report.
If you use the public/open source codes, please provide the links to the sources.
Sources:
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
https://realpython.com/logistic-regression-python/