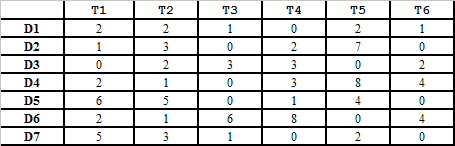
# Homework 6

Your Name: Srujan Shekar Shetty

Student ID: A20529733

**1. (40 points) Clustering by the K-Means approach**

Consider the following document-term matrix, where each entry represents the raw frequency of a term Ti in document Dj. We would like to apply clustering to automatically group these documents into 3 classes (clusters). Note: you can use a spreadsheet program such as Microsoft Excel to facilitate computation in intermediate steps. But you need to show the results in each step, until the clustering is converged



Suppose we initially assign D2 to Cluster 1, D4 and D6 to Cluster 2, and D5 and D7 to Cluster 3. Using the K-means clustering method discussed in class, compute the final contents of the 3 Clusters. Use the Cosine similarity of two vectors (NOT only the dot product) as your similarity measure. Show the details of your computation, including intermediate steps in each iteration of the algorithm.

Note: Recall that the Cosine similarity of two vectors is their dot product divided by the product of their norms. For example, Consider the two vectors X and Y:

X = <3, 0, 1, 2, 0, 3>

Y = <2, 0, 0, 3, 8, 4>

The dot product is given by sum of the coordinate-wise multiples:

dot-product(X, Y) = 3\*2 + 0\*0 + 1\*0 + 2\*3 + 0\*8 + 3\*4

= 6 + 0 + 0 + 6 + 0 + 12

= 24.

The norm of each vector is the square-root of the sum of the squares of its dimension values. So, the norms of X and Y are:

http://facweb.cs.depaul.edu/mobasher/classes/ect584/HW/norm-d1.gif     http://facweb.cs.depaul.edu/mobasher/classes/ect584/HW/norm-d4.gif

and the Cosine similarity of X and Y is given by:

http://facweb.cs.depaul.edu/mobasher/classes/ect584/HW/sim.gif

Your answer:

In the approach for K means clustering, we shall follow the following steps

Before we start, let us initialize the K clusters.

Here the clusters are given initially.

C1= D2

C2= D4, D6

C3= D5, D7

**Step 1: Calculate the centroids of K Clusters**

To calculate we shall use Excel.

As instructed Cosine similarity is applied by using the below formula

=SUMPRODUCT(C1:C3, D1:D7) / (SQRT(SUMSQ(C1:C3)) \* SQRT(SUMSQ(D1:D7)))

**Step 2: Assign the data points using distance between centroids and data as reference**

**Step 3: Compare the new K clusters with the previous ones**

**Step 4: Repeat the learning iterations till there is no movement in cluster points(convergence)**

The following shows the screenshot of First Iteration:

A screenshot of a spreadsheet

Description automatically generated

From the above screenshot, we can infer that the new centroids for iteration 2 are.

C1=D2, D4

C2= D3, D6

C3= D1, D5, D7

After comparing the new K cluster values with the previous one we proceed to the next iteration.

The below screenshot shows the Iteration 2 results.

Comparing the New K cluster values to the previous ones, we can observe the convergence of data points indicating to stop the iterations.

A table with numbers and symbols

Description automatically generated

From the above screenshot, we can observe the following after the final iteration.

C1=D2, D4

C2= D3, D6

C3= D1, D5, D7

**2. (30 points) Practice with hierarchical clustering**

Perform a hierarchical clustering of the following data points:

D1: <1, -2>

D2: <-1, 9>

D3: <-6, 4>

D4: <0, -5>

D5: <1, -9>

clusters are represented by their centroid (use means), and at each step the clusters with the closest centroid are merged. And you should use the “Bottom-to-Top” approach to construct your hierarchical tree. Use the Euclidean distance to measure the distance between instances or clusters.

You must draw a hierarchical tree structure and give the sequence (such as, ①②③…) to each merging operation. For example, number ① means this merge is the 1st merge operation. In addition, you must use the centroid value or vector to represent the new value or vector for each group of points. Write down the centroid value or vector to the side of the ① notations.

Your answer:

A piece of paper with writing on it

Description automatically generated

A paper with writing on it

Description automatically generated

**A drawing on a piece of paper

Description automatically generated**

**3. (30 points) Run K-Means on Malware\_MultiClass.csv data**

**Notes**

* **When you run K-Means clustering, you should exclude the label in the classification task**
* **Set maximal learning iteration = 500**
* **Use elbow method to find the optimal K**
* **Evaluate the K-Means results by using at least two methods**