CS 4980/6980: Introduction to Data Science © Spring 2018

Lecture 10: Review

Instructor: Daniel L. Pimentel-Alarcón Scribed by: Samin Khan and Jason Abrahams

This is preliminary work and has not been reviewed by instructor. If you have comments about typos, errors, notation inconsistencies, etc., please email the scribers.

10.1 Introduction

This lecture was used for Exam review. This was the problem given during class.

10.2 K-Mean Clustering Problem

Consider the following data matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & -7 & 5 & 8 \\ 2 & 4 & 9 & 0 \\ 3 & -2 & -6 & -1 \end{pmatrix}$$

and the following intial centers: $\mu_1 = \begin{pmatrix} -4 \\ 3 \\ -1 \end{pmatrix}$ $\mu_2 = \begin{pmatrix} 9 \\ -4 \\ 6 \end{pmatrix}$

After 2 iterations of K-means clustering, how would the colums in $\mathbf{X}$ be clustered?

10.3 Eucliden Distance Formula

d(p,q)= $\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$

This will be used to find the distance between the object and center of each cluster.

10.4 Iterations

Using the Eucliden Distance method, you complete the first iteration by plugging in each X value with centriod values. You then assign each new object to group based on minimum distance. Ex) Out of pair (6.5, 10.4) for column X1, $6.5 < 10.4$. Therefore, there will be a 1 put in for 6.5 and 0 put in for 10.4 in the Object Clustering table.

Distance Matrix	X1	X2	X3	X4	
	6.5	3.3	11.9	12.4	C_1
D_0=	10.4	19.6	18.1	8.1	C_2

Object Clustering	X1	X2	X3	X4	
	1	1	1	0	C_1
G_0=	0	0	0	1	C_2

10.5 Clusters

New Cluster 1 = X1, X2, X3 = $\frac{(1-7+5)}{3}$, $\frac{(2+4+9)}{3}$, $\frac{(3-2-6)}{3}$ = -0.33, 5, -1.67 = $\mu_1 = \begin{pmatrix} -0.33 \\ 5 \\ -1.67 \end{pmatrix}$

New Cluster 2 = X4 = 8, 0, -1 = $\mu_2 = \begin{pmatrix} 8 \\ 0 \\ -1 \end{pmatrix}$

10.6 Second Iteration

10.7 ANSWER

Since $G_1 = G_0$, no more iterations need to be made.

Final Cluster 1 = X1, X2, X3 = $\frac{(1-7+5)}{3}$, $\frac{(2+4+9)}{3}$, $\frac{(3-2-6)}{3}$ = -0.33, 5, -1.67 = $\mu_1 = \begin{pmatrix} -0.33 \\ 5 \\ -1.67 \end{pmatrix}$

Final Cluster 2 = X4 = 8, 0, -1 = $\mu_2 = \begin{pmatrix} 8 \\ 0 \\ -1 \end{pmatrix}$

X1, X2, X3 are clustered together and X4 is clustered by itself

Distance Matrix	X1	X2	X3	X4	
	5.7	6.8	7.9	9.7	C_1
D_1=	8.3	15.6	10.7	0	C_2

Object Clustering	X1	X2	X3	X4	
	1	1	1	0	C_1
G_1=	0	0	0	1	C_2