

Homework 2

DSA1101

Introduction to Data Science

September 7, 2018

Name:

Matriculation card number:

Problem 1 (15 points). *The k -means clustering algorithm*

Suppose we have data for six objects on two features:

object	x_1	x_2
A	1	1
B	1.5	2
C	3	4
D	3.5	5
E	4.5	5

We set $k = 2$ to cluster the six data points into two clusters, \mathcal{P} and \mathcal{Q} , and initialize the algorithm with the centroids $(x_{1,\mathcal{P}}, x_{2,\mathcal{P}}) = (2, 2)$ and $(x_{1,\mathcal{Q}}, x_{2,\mathcal{Q}}) = (4, 4)$.

- (a) Fill up the following table to identify the objects in each cluster during the first iteration of the k -means algorithm:

cluster	object(s)
\mathcal{P}	
\mathcal{Q}	

- (b) Compute the new centroids for the two clusters based on cluster assignment in (a). Please show your working.

- (c) Based on the centroids computed in (b), fill up the following table to identify the objects in each cluster during the second iteration of the k -means algorithm:

cluster	object(s)
\mathcal{P}	
\mathcal{Q}	

- (d) Calculate the *Within Sum of Squares* (WSS) for the clustering assignment in (c). Please show your working.

Problem 2 (10 points). *The k -nearest neighbor classifier*

Suppose we have labelled training data for three objects with two features:

object	x_1	x_2	y
A	4	1	1
B	4.5	4	0
C	2.5	2	0

Here y is a categorical outcome with only two levels, $y = 1$ or $y = 0$.

- (a) Predict the value of the outcome y for the following objects, using the k -nearest neighbor classifier with $k = 1$, based on the training data set.

object	x_1	x_2	Predicted y
D	2	2	
E	3	2.5	
F	4	1.5	

(b) The actual value of outcome y for the objects in (a) are

object	x_1	x_2	Actual y
D	2	2	0
E	3	2.5	1
F	4	1.5	0

Compute the *accuracy*, *true positive rate*, *false positive rate* and *false negative rate* of the classifier based on the actual and predicted values of y computed in (a). The definitions are as follows:

Actual y		Predicted y	
		1	0
	1	True Positives (TP)	False Negatives (FN)
	0	False Positives (FP)	True Negatives (TN)