CS 6350 – Big Data Analytics and Management

Spring 2018

Assignment# 4

Due: April 23 by 11.59 p.m.

Part I: Clustering

Question A (K-means algorithm)

Consider the following eight points in 2-dimensional space: (2,10); (2,5); (8,4); (5,9); (7,5); (6,4); (1,2); (4,9); (10,10). Suppose we plan to use the Euclidean distance metric and that we are interested in clustering these points into 3 clusters.

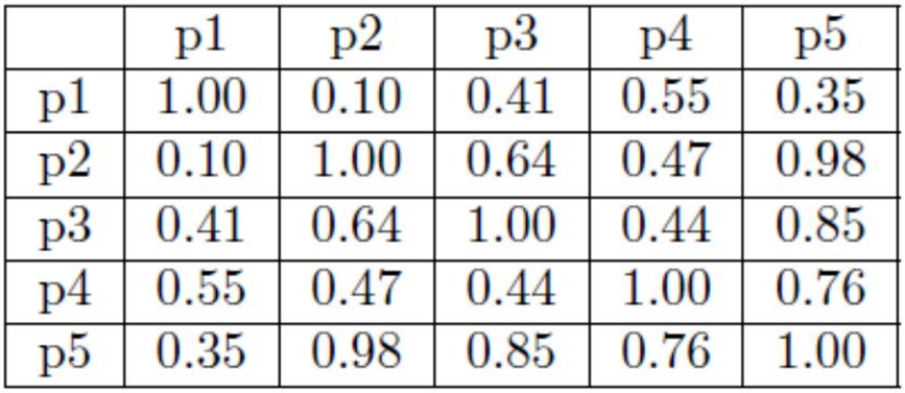
(i).Plot the data to see what might be appropriate clusters.

(ii) Beginning with the points (2,5), (5,8) and (4,9) as initial cluster centers, form the three initial clusters.

(iii) Use the k-means clustering algorithm to get the final three clusters. What are the resulting centers and resulting clusters? (Here K = 3)

Question B

(i). Use the similarity matrix in Table 1 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.



	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Question C.

(ii) Hierarchical clustering is sometimes used to generate K clusters, K > 1 by taking the clusters at the Kth level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: {6, 12, 18, 24, 25, 28, 30, 42, 48}.

(a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

1) { 5, 7.5}

2) {15, 25}

b) Do both sets of centroids represent stable solutions, i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

c) What are the two clusters produced by MIN? (MIN is single-link clustering, also called minimum method)

d) Which technique, K-means or MIN, seems to produce the "most natural" clustering in this situation?

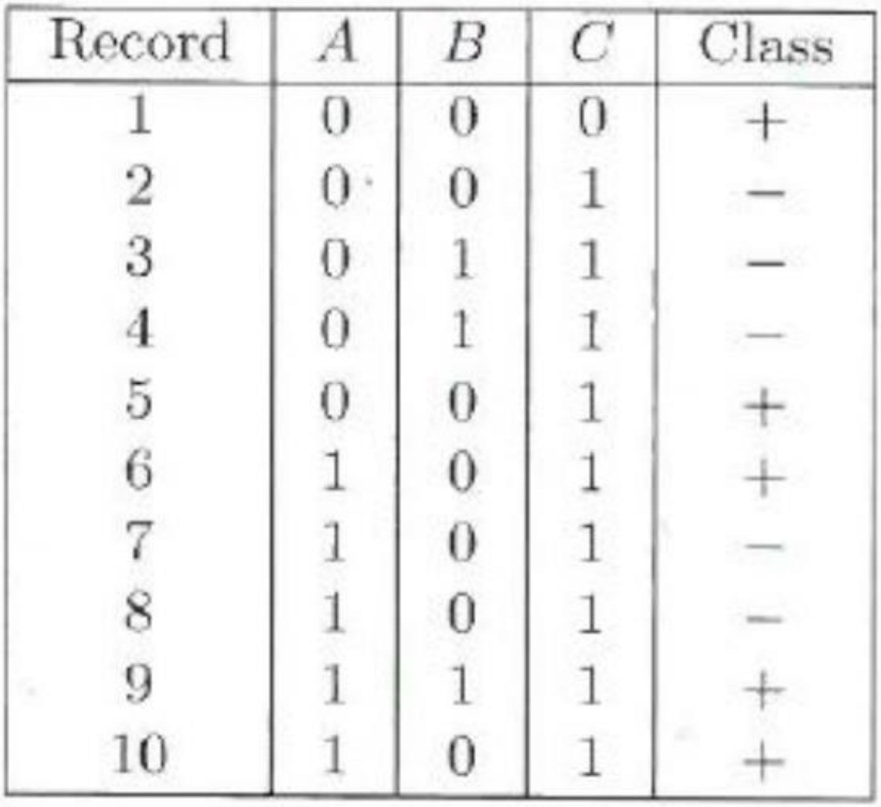
e) What well-known characteristic of the K-means algorithm explains the previous behavior?

Part II: Classification

Question D

i) Naive Bayes.

Consider the data set shown in Table 2:



Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

(1) Estimate the conditional probabilities for P (A|+), P (B|+), P (C|+), P (A|−), P (B|−), and P (C|−).

(2) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A = 1, B = 1, C = 0) using the naive Bayes approach.

(3) Estimate the conditional probabilities using the m-estimate approach, with p = 1/2 and m = 4.

(4) Repeat part (2) using the conditional probabilities given in part (3).

(5) Compare the two methods for estimating probabilities. Which method is better and why?

(ii) Adaboosting

Consider the following dataset, plotted in Figure 1:

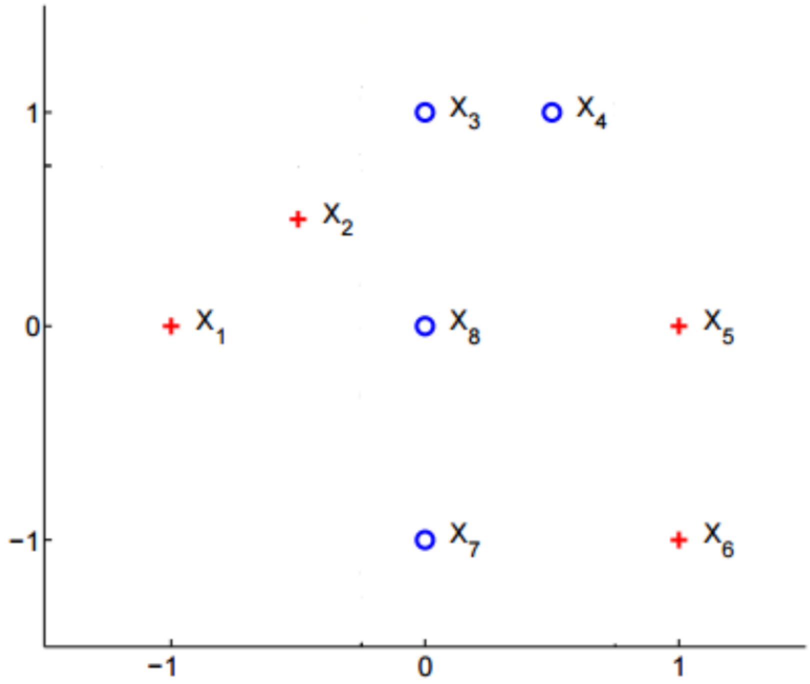


Figure 1: A small dataset, for binary classification with AdaBoost.

X1 = (−1, 0, +), X2 = (−0.5, 0.5, +), X3 = (0, 1, −), X4 = (0.5, 1, −), X5 = (1, 0, +), X6 = (1, −1, +), X7 = (0, −1, −), X8 = (0, 0, −).

In this problem, you'll run through T = 3 iterations of AdaBoost with decision stumps (axis-aligned half planes) as weak learners. (Attention: Z_t 's calculate function is Z = 2 √ $\varepsilon_t\,(1-\varepsilon_t)$)

1): For each iteration t = 1, 2, 3, compute ε_t, $\alpha_{t,}$ Z_t, $W_t(i)$ (i = 1,2...8), and draw the decision stump (on Figure 1). Recall that Z_t is the normalization factor to ensure that the weights W_t sum to 1.

t	εt	αt	Zt	Wt(1)	Wt(2)	Wt(3)	Wt(4)	Wt(5)	Wt(6)	Wt(7)	Wt(8)
1											
2											
3											

2). What is the training error of AdaBoost? Give a one-sentence reason for why AdaBoost outperforms a single decision stump.

Q5: Decision Tree

a. Given the following training data and attributes with their respective possible values: X1 ∈{a,b}, X2 ∈{c,g,u,w}, X3 ∈{k,s,v}. The class is Y ∈{+1,-1}.

x1	x2	x3	y
a	c	k	-1
a	w	k	-1
b	w	v	+1
a	c	v	+1
b	w	k	-1
a	c	s	+1
b	w	s	+1
a	u	v	-1
b	c	v	-1
b	c	s	+1
b	g	v	-1

Learn a decision tree using the ID3 algorithm (information gain heuristic with entropy) and draw the resulting decision tree (with all parts labeled accordingly). Please show all calculations justifying your answer.

b. Given the following training data and attributes with their respective possible values: X_1 ∈{a,b}, X_2 ∈N, X_3 ∈{e,f}, X_4 ∈{c,d}. The class is Y ∈{+1, -1}.

X_1	X_2	X_3	X_4	Y
b	185	f	d	+1
b	180	f	c	+1
b	170	f	c	-1
b	140	e	d	-1
a	176	f	d	+1
b	179	f	d	-1

i. Draw a decision tree with only 2 attribute node 2 attribute nodes and 3 leaf class nodes that will get 100% accuracy on the training data. No need to show calculations, but please make sure all parts of the tree are labels accordingly.

ii. What is the accuracy of the following test data using your learned decision tree?

X_1	X_2	X_3	X_4	Y
b	170	f	d	-1
a	150	f	d	+1
b	60	f	d	+1