Homework 4

(Advanced) Data Mining: Algorithms and Applications-Winter 2024

Due on Mar 30, 11.59PM

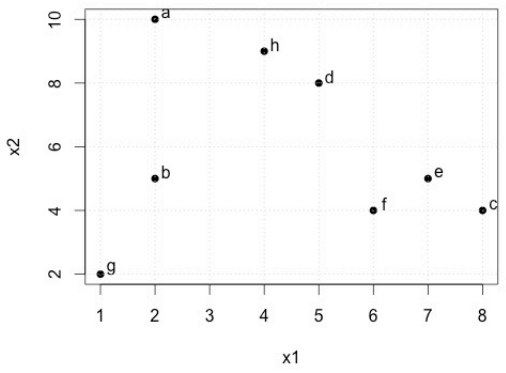
Important

- Please type your answers for the calculations.
- Only submissions through Canvas will be accepted.

department	age	salary	status	count
sales	31_35	46K_50K	senior	30
sales	26_30	26K_30K	junior	40
sales	31_35	31K_35K	junior	40
systems	21_25	46K_50K	junior	20
systems	31_35	66K_70K	senior	5
systems	26_30	46K_50K	junior	3
systems	41_45	66K_70K	senior	3
marketing	36_40	46K_50K	senior	10
marketing	31_35	41K_45K	junior	4
secretary	46_50	36K_40K	senior	4
secretary	26_30	26K_30K	junior	6

1. Given a data tuple having the values "systems", "26_30", and "46K_50K" for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status according to the data above? Notice that *Count* column is **NOT** an attribute. It just tells how many times a row occurs in our database and *status* is our target variable. (No need to write a program) (25pts)
2. By using Python Notebooks on Canvas, split your diabetes data into two parts for training and testing purposes. Namely, reserve last 10 rows of the diabetes_train.csv for the test set. Then fit a SVM classifier on the bigger portion of this data and test it on these 10 rows you had reserved. Please feel free to modify existing codes. Notice that you're not going to read diabetes_test.csv anymore since you're going to split the bigger data. Please submit your Python code and your prediction results. (25pts)
3. Please use the data shown for questions below. (25pts)

	x1	x2
a	2	10
b	2	5
c	8	4
d	5	8
e	7	5
f	6	4
g	1	2
h	4	9



(a) If *h* and *c* are selected as the initial centers for your *k*-means clustering, assign memberships for other points, and compute the means (centroids) of your initial clusters. You can use Manhattan distance.

(b) Based on the centroids you found above reassign the memberships by using Manhattan distance.

4. Given the distance matrix below answer the following questions. Notice that this is a distance matrix, meaning the distance between any pair of points can be found by checking the corresponding cell them. (25pts)

	a	b	c	d	e	f	g
b	5						
c	8	6					
d	4	4	5				
e	7	5	1	4			
f	7	4	2	4	1		
g	8	3	7	7	7	5	
h	2	4	6	1	5	5	8

(a) Perform hierarchical clustering using *single link* measure for the above and draw the final dendrogram.

(b) Determine whether a point is *core* based on $\epsilon = 6$ and $minPts = 2$. (Recall that a point p is a core point if at least $minPts$ points are within distance ϵ of it (including p).)