Data mining exam questions

Points -> Q11: 20, Q10: 15, Q9: 15, Q8: 20, Q7: 20, Q6: 20, Q5: 20, Q4: 25, Q3: 10, Q2:20 , Q1: 15

Total points: 200

Duration: 3 hours

Date: 23 - 06 - 2023

→ look at the memory of your calculator

Q1 (full page): Do the first iteration for logistic regression. X1 = -1, Y = 1, w0 = 1, w1 = -1, learning parameter = 1, regularization = 0.1

- P(yi = 1 | w, xi) = 0.5
- w0(t+1) =
- w1(t+1) = -1 + 1*(-.1 * -1 + (-1*(1 - 0.5))) = -1.4

Q2 (full page): Gradient boosting question. Given: 3 trees + 4 examples with 4 X parameters (X1 = numeric, X2 = T or F, X3 = numeric, X4 = T or F) and a Y value (numeric). Predict the y value for the next iteration for X1 and X2 with learning parameter of 1. Do the same for learning parameter of 0.5 but for X3.

Tree 1 Tree 2 Tree 3

X1 X3 X4

| |
|<10 |>= 10
X? X?
| | | |
| | | |

	x1	x2	x3	x4	Y
1					17

2					26
3					20
4					

Answer X1: 16 -2 +1 = 15

Answer X2:

Answer X3: 10*.5 + 8*.5 + 2 * 0.5 = 10 (??)

Q3: Recommender system?

Q4:

Q5:

Q6:

Q7:

Q8:

Q9: Logistic regression. What is the shape of L1 and L2 on the plot with #non-zero coefficients on the Y axis and size of lambda 10^-6 to 10^6 on the X axis.

Q10 (full page): Given is a one dimensional 10 datapoints dataset. What is the maximal accuracy one can obtain when you train a logistic classifier on the data?

- Maximum accuracy:
- Please explain by drawing the decision boundary, or with maximum 25 words excluding formulas.

Q11 (full page): Question about Toivonen's algorithm applied to a dataset → Given are some itemsets and should give the itemsets on the negative border

---

Questions about (please put them in the correct order above if you still remember where which one belonged to):

One the same page:

[one question] Association rules

- What is the confidence of {} -> _
- What is the interest of H -> B
- What is the lift of {} → _ (⅔)/(¾) = 0.8889?

- What could be items in the conditional __ of AH if the support threshold is 2? [Frequent pattern growth]
  - Answer

    First count support: E: 8, F: 7, B:6 , A:5, C:3, D: 2, G:1, H:2, I:2, J:1, L:1

    Then Prune: E:8, F: 7, B:6, A:5, C:3, D:3, H:2, I:2

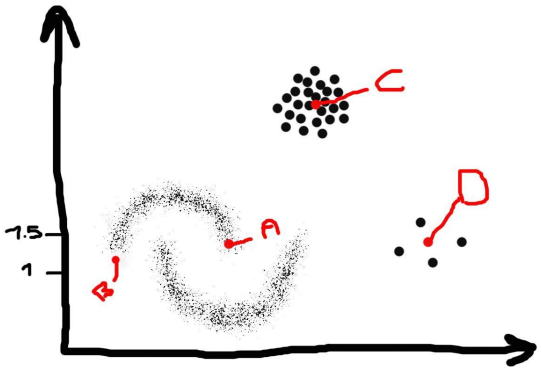
    Make tree

    E, F, B all go before A and H so them? Can be more?

One the same page :

[one question] Clustering

- DBSCAN. MinPts = 5, e = 0.75:



  - what does ‘C’ represent
  - what does ‘D’ represent

- Does ‘B’ belong to the same cluster as ‘A’? Write down your answer in 1-2 sentences taking into account the criteria of DBSCAN

[another question] Time series

- You want to classify a time series data to do DTW with 1 KNN which has a certain shape and *amplitude* what preprocessing do you need to do for this task? (the question is asked in a way so it is not clear if it should be only one or multiple preprocessing steps)

On the same page:

[one question] Sequence

- Prefix with <e> projection and threshold of support 3.

[another question] Clustering

- Clustering with kmeans++: given are 5 clusters and also 5dots (A to E) selected in these clusters. If we run kmeans++ with k=2 starting at A initially, which point will it most likely select next out of the other four. List them from most to less likely. Assume Manhattan distance.