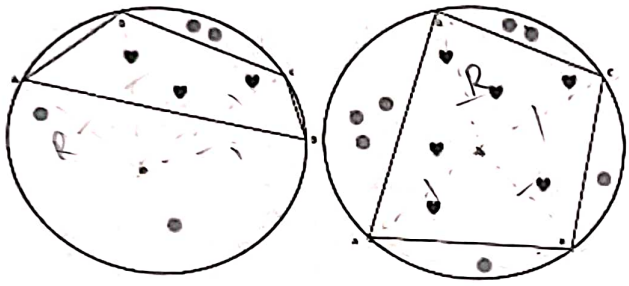
CS21B059

CS303L/CS519L Machine Learning End Semester Examination

Date: 25/Nov/2023, Duration: 2 hours, Marks: 70, Instructor: Dr. Y. Kalidas

Q1) ***(10 marks)*** Consider the scenario where a cyclic quadrilateral (any shape, example figures below) is encapsulating *hearts* (`label = -1`) and outside of the quadrilateral but inside the circle, the points are *circles* (`label = +1`). Each corner is represented by a 2D coordinate. The quadrilateral is represented by 4 coordinates of its points (A, B, C, D) *in any* order.

Given a point (x, y) how will you determine if the point is lying inside or outside the quadrilateral? Please provide a very clear answer with properly numbered equations. Also you need to address all scenarios why your algorithm should work.



Q2) ***(5 marks)*** In continuation of (Q1) above, given are the centre of the circumcircle as (c_x, c_y) and radius as R.
How do you generate some random 100 points some of which lie inside the quadrilateral and some lie outside the quadrilateral but inside the circle? Please provide very clear answers with neatly numbered equations and steps.

Q3) ***(5 marks)*** In continuation of (Q1) and (Q2).
What is the loss function for the data set $D = \{ (x_1, y_1), \dots, (x_N, y_N) \}$?

Q4) ***(5 marks)*** There are 5 sets of points that are well separated but have thin lines of noisy patterns connecting them. Which of the clustering algorithms DBSCAN, K Means and Agglomerative is prone to error, which are not and why?

Q5) ***(5 marks)*** Consider the loss function, $l(w) = (\max(w \times x, 0) - y)^2$. Given are the initial values of $x, y, w \in R$ as 1, 2, 2.1 respectively. Consider gradient descent with learning rate 1. What are the updated values of w after two iterations?

Q6) ***(10 marks)*** Compare Bagging, Boosting and Random Forest algorithms in terms of data rows, columns, underfitting and overfitting behaviour?

Q7) Consider the following table T for studying 2-variate decision tree regression.

Assume availability of the following 4 functions:

- `var(T, col_idx)` computes variance of a specified column `col_idx` on table T.
- `mean(T, col_idx)` computes the mean value of a specified column `col_idx` on table T.
- `subset(T, col_idx, comparator_str, v)` selects a subset of rows from T that satisfy a condition. `comparator_str` provides standard comparisons such as "<", "=", "<=", ">", ">=".
  For example, `T1 = subset(T, 4, "<", 12)` or `T2 = subset(T, 2, "<=", 1.2)` or `T3 = subset(T, 1, "=", a)` etc.
  Output is T type
- `size(T)` gives number of rows in T

Sample data table T:

Col Idx ->	1	2	3	4	5
Row Idx	X1	X2	X3	Y1	Y2
1	a	1	1	1	-6
2	a	-1	1	-1	-5
3	b	4	1	2	-4
4	c	3	1	2	4
5	c	7	1	1	5
6	c	1	1	-3	6

Q7A) ***(5 marks)*** **What is the weighted variance for X1 in terms of the functions defined?**

Q7B) ***(5 marks)*** **What is the weighted variance of X2 in terms of the functions defined?**

Q8) *(10 marks)* What are the two important loss functions for a **K class** classification using decision trees? Provide a clearly understandable standard formula for the same.

Q9) *(5 marks)* What are any five non trivial and important types of operations that can be used in a transformation pipeline?

Q10) *(5 marks)* Which of the following 5 options are most appropriately True?

A) PCA selects a subset of columns
B) PCA performs dot product against eigenvectors
C) Convolution requires positional relationship between data points
D) Convolution requires positional relationship between columns of a data point
E) The output of a convolution operation is called a feature map
F) The output of a convolution operation is called dimension
G) Deconvolution is used for reducing dimensions
H) Deconvolution increases dimensions
I) PCA can be used in a data transformation pipeline

All The Best