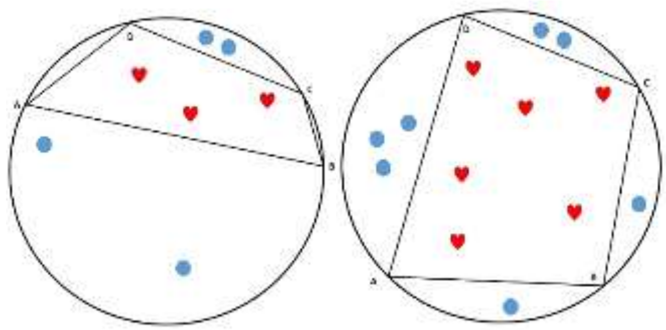
CS303L/CS519L Machine Learning End Semester Examination

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.



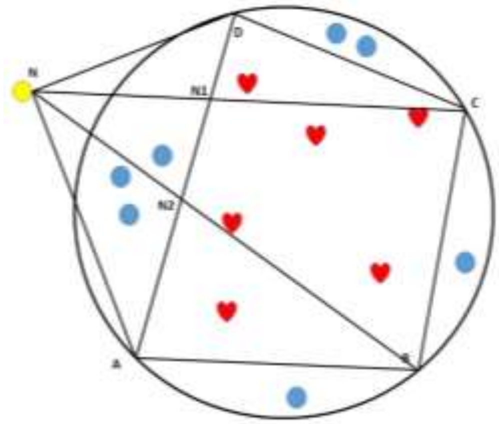
Ans Q1).

STEP 1: Determine ***sides*** of the quadrilateral.
A side is formed by a pair of points (P,Q) such that all other points are only on one side of the line joining P and Q. ***Diagonals will be excluded*** by this logic.

STEP 2: For every ***side*** determine the ***direction of the interior*** of the quadrilateral. The direction of the interior is the ***direction of all other points***.

STEP 3: Given a new point (x,y) determine if it's on the ***interior direction*** of all sides. If YES, then interior, otherwise exterior.

An alternative approach is to calculate sum of areas formed by triangles > area of the quadrilateral if the point lies outside.



Alternative answer can be based on angle subtended at the point by the adjacent vertices.

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.

Ans Q2).

STEP 1: Randomly select a floating point value r between 0 and R
STEP 2: Randomly select an orientation \theta between -pi to +pi
STEP 3: Generate a polar coordinate centred about (cx, cy)
STEP 4: Generate the coordinate as (cx + r cos(theta), cy + r sin(theta))
STEP 5: Provide label +1 or -1 based on whether it lies inside the quadrilateral or outside.

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

Ans Q3). loss(A, B, C, D) = $\sum_{i=1}^{i=N}(1 - y_i \times inside(A, B, C, D, x_i))$

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?

Ans Q4.
Agglomerative is prone to error as it joins unrelated blobs of points into a cluster.
K Means is not prone to error since the noisy points are less in number, for optimal K.

DBSCAN is also not prone to error provided radius and neighbourhood is defined properly.

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?

Reviewed, the answer below is correct.

Ans Q5. 1.9, 2.1

```
grad(l,w) = 2*(max(w*x,0) - y)*x (if w*x > 0)
or
grad(l,w) = 2*(max(w*x,0) - y)*0 (if w*x <= 0)

x=1, y=2, w=2.1
w*x = 2.1
grad(l,w) = 2*(2.1-2)*1 = 2*0.1 = 0.2
w = 2.1 - 1 * 0.2 = 1.9
```

~~w*x = 1.9~~
~~grad(l,w) = 2*(2.08-2)*0 = 0~~
~~w = 1.9 - 1*0 = 1.9~~

```
w*x = 1.9
grad(l,w) = 2*(1.9-2)*1 = -0.2
w = 1.9 - 1*(-0.2) = 1.9 + 0.2 = 2.1
```

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

Ans Q6. A qualitative comparison table to be provided.

Method	**Full data As a rule of thumb**	**Subset of rows As a rule of thumb**	**Subset of columns As a rule of thumb**	**Underfitting general observation**	**Overfitting general observation**
Bagging	No	Yes	No	If individual estimators are too simple	No as averaging takes place

				Overall estimator is Underfitting	
Boosting	Yes	No	No	No Already individual estimators are weak	Yes
Random Forests	Yes	No	Yes	Yes if individual estimators are weak No if individual estimators are complex enough	No

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?**

Ans Q7A.

```
Ta = subset(T, 1, "=", a), Tb = subset(T, 1, "=", b),
Tc = subset(T, 1, "=", c)
S = size(T), Sa = size(Ta), Sb = size(Tb), Sc = size(Tc)
Va = Var(Ta, 4) + Var(Ta, 5), Vb = Var(Tb, 4) + Var(Tb, 5)
Vc = Var(Tc, 4) + Var(Tc, 5)
VX1 = (Sa/S * Va) + (Sb/S * Vb) + (Sc/S * Vc)
```

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

Ans Q7B.

```
m = mean(T, 2)
Tlt = subset(T, 2, "<", m), Tgte = subset(T, 2, ">=", m)
Vlt = Var(Tlt, 4) + Var(Tlt, 5)
Vgte = Var(Tgte, 4) + Var(Tgte, 5)
S = size(T), Slt = size(Tlt), Sgte = size(Tgte)
VX2 = (Slt/S * Vlt) + (Sgte/S * Vgte)
```

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.

Ans Q8.

gini impurity index = $\sum_{i=1}^{i=K} p_i \times (1 - p_i) = 1 - \sum_{i=1}^{i=K} p_i^2$

log loss or entropy = $-\sum_{i=1}^{i=K} p_i \times \log(p_i)$

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

Ans Q9. (a) matrix multiplication or dot products, (b) subset selection, (c) stacking, (d) convolution, (e) de-convolution

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

Ans Q10. True: B, D, E, H, I

All The Best