Artificial Intelligence and Machine Learning - Unit II
16 September 2022

First and last name, Student ID: _____ Seat: _____

1. Answer the following questions by reporting also the procedure that leads you to the numerical values, not just the values themselves.

   (a) In Eq. (1) left, $\mathbf{X}$ is a design matrix where each row is a sample. What is the dimensionality of the samples in $\mathbf{X}$ and how many samples do we have? Complete the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to compute the empirical average and the covariance matrix associated with $\mathbf{X}$. [2]

   $$\mathbf{X} = \begin{bmatrix} 2 & 5 \\ -10 & -7 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \quad\quad \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \quad\quad \end{bmatrix} \tag{1}$$

   (b) Referring to transformation $\mathbf{T} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, compute the area of the ellipsoid that gets generated once you map the unit circle through $\mathbf{T}$ as shown in Fig. 1 *(Hint: start from the area of the unit disk which is πr^2 and remember how to compute the change in volume induced by a transformation).* Explain how your approach can work also in higher dimensions, not only in 2D. Given a generic transformation $\mathbf{T}$, when it may happen that the ellipsoid is squished down to a line? [3]

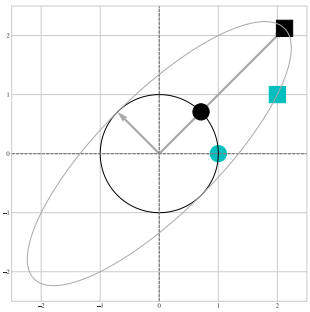

Figure 1: The black round point is sent to the black squared point. The gray round point to the gray squared point.

   (c) You work as a computer scientist for a medical company and you interact with a doctor. He wants to visualize the most prominent variation in 3D scans of skulls. The doctor can give you a set of 3D point cloud of skulls. The point clouds are densely registered and aligned across all samples and given to you as matrix $\mathbf{S} \in \mathbb{R}^{3P \times N}$, where the P is the number of points in each cloud and N is the number of individuals. Describe the technique that helps the doctor in the visualization, describing even with math what you need to implement, how you find the most prominent variation and how you morph the data along the prominent variation. [2]

Total for Question 1: 7

2. We are given a set of points $\mathbf{X}$ in 2D with no associated labels shown in Tab. 1. We wish to find the main 2 clusters identified by the centers $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$. We hypothesize that the clusters distribute as Gaussian blobs with the same standard deviation across clusters, that is, clusters more or less will distribute as spheres all of the same size.

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
[2.44, 0.96]	[2.28, 1.06]	[1.45, 4.23]	[1.91, 3.82]	[2.13, 1.62]	[0.92, 4.51]

Table 1: Training set

(a) Define the objective function that can solve the clustering problem mentioned above and describe the necessary steps to minimize the function. [2]

(b) Assume that the two starting cluster centers are $\boldsymbol{\mu}_1 = [1.4, 3.0]$ and $\boldsymbol{\mu}_2 = [1.8, 2.0]$. Given that now we know the cluster centers, compute the **assignment step**. Fill in the blanks alongside each point to indicate the assignment for that point. Show the procedure that was used to get the assignment just for point $\mathbf{x}_3$. [1]

$$\left[\mathbf{x}_1 \_\_\_\_; \quad \mathbf{x}_2 \_\_\_\_; \quad \mathbf{x}_3 \_\_\_\_; \quad \mathbf{x}_4 \_\_\_\_; \quad \mathbf{x}_5 \_\_\_\_; \quad \mathbf{x}_6 \_\_\_\_;\right]. \tag{4}$$

(c) Given the assignments you have found in Eq. (4), now compute the **update step** and explaining the procedure for your computation. [1]

(d) Different clustering methods that we reviewed in our course can handle different shapes of the data density. Fig. 2 offers four different configurations of the data density. Each color indicates a cluster. For each Case: (a) state which clustering algorithm is the more appropriate and (b) report if you need to model the covariance matrix (c) if so, how do you model it, specifying how many parameters you need in the covariance matrix, if you need off-diagonal values (d) if you need a specific covariance matrix for each blob. [3]

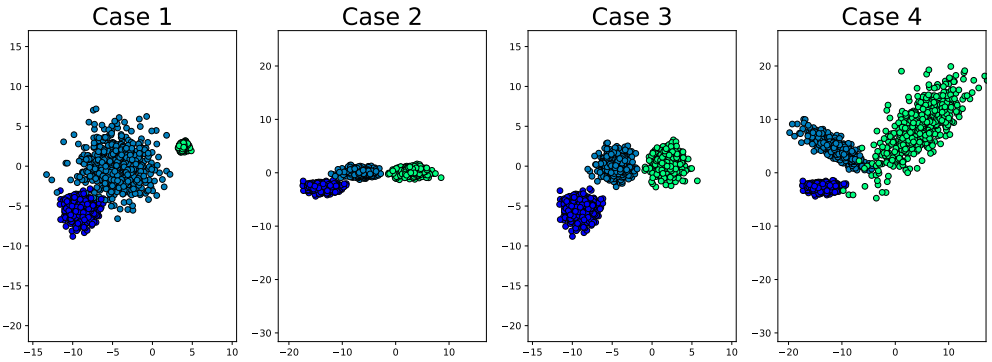

Figure 2: Different shapes of the data density.

Total for Question 2: 7

3. We are given an already learned decision tree for binary classification shown in Fig. 3 below. Each square represents a training sample and each circle is a node. Negative points are white squares while positive points are gray squares.

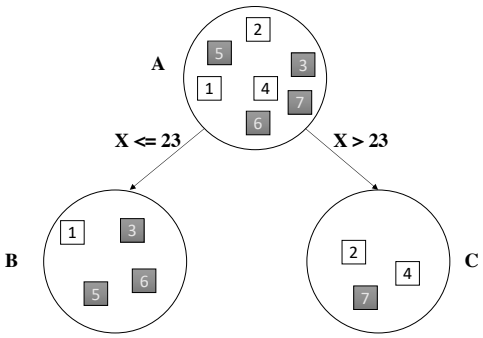

Figure 3: Decision Tree

(a) Compute the Impurity using the Entropy for nodes B and C. What is the Entropy of the Tree? What is the Entropy of the entire training set? [3]

(b) Define the Gini Impurity function over a set $\mathcal{S}$ for a generic k-class classification problem and compute the Gini Impurity for the entire training set in Fig. 3. [2]

(c) Given the splitting attribute of the decision tree in Fig. 3, assume we have to classify a test sample which feature $\mathbf{x}$ is 25. What is the y' label predicted by the above decision tree, given $\mathbf{x}'$? How much is the probability returned for that prediction, $p(y'|\mathbf{x}')$? [1]

Total for Question 3: 6

4. We want to perform some evaluation of a binary classifier that has positive labels +1 and negative labels −1 and the scores are reported as s below the labels in Tab. 2.

y	+1	-1	-1	-1	-1	+1	-1
s	100	-100	-99.5	-99	99	0	-1

Table 2: Labels and unnormalized scores for a binary classifier.

(a) Give a definition of True Positive Rate (TPR) and False Positive Rate (FPR). Given a binary classifier with unnormalized scores s—the higher the score, the more is likely that $y = +1$—compute the ROC curve for the values in Tab. 2 by showing the TPR and FPR in a table. 2

(b) Compute the Area Under the Curve (AUC) of the above ROC. 2

(c) Let's assume that in Tab. 2, we replace each score as $s' = \text{sign}(s)\sqrt{|s|}$. Will the ROC change? What if we use $s' = \cos(s)$? Motivate why will/will not change for both cases. 2

(d) Setting the classifier threshold to -0.5, compute the confusion matrix for the values in Tab. 2. For each cell of the confusion matrix, indicate what is the metric computed. 2

Total for Question 4: 8

5. We have to solve a linear regression problem, given a design matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ where N is the number of samples and D is the number of features. We want to regress a value $\mathbf{y} \in \mathbb{R}^N$ by learning parameters $\boldsymbol{\theta}$. Assume no bias. We thus want to solve the following optimization problem.

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 = \arg\min_{\boldsymbol{\theta}} \frac{1}{2} (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) \tag{6}$$

(a) Describe with words a generic problem in which you need to use **regression** instead of **classification**. What is the difference between the two? Referring to Eq. (6), how many parameters do we have to learn? (*Hint: use dimensionality check to see if you answer is correct for the #params.*) [1]

(b) Derive the closed form solution to minimize the objective function Eq. (6). Show all the steps of how you derive the solution and justify all the steps. [3]

(c) Now assume that it is forbidden to use the closed form solution to minimize Eq. (6). Is there any way to find an approximate solution using another method? If so, explain the algorithm you are going to use from a mathematical and computer science perspective. [3]

Total for Question 5: 7

You can use this space for writing. Summary for points is at the bottom.

Question:	1	2	3	4	5	Total
Points:	7	7	6	8	7	35
Score:						