Artificial Intelligence and Machine Learning - Unit II
June 23, 2022

First and last name, Student ID: ______________________________ Seat: ______

1. You have to implement some geometrical checks for a machine learning algorithm.

   (a) In Eq. (1) left, $\mathbf{X}$ is a design matrix where each column is a sample. Complete the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to compute the empirical average and the covariance matrix associated to $\mathbf{X}$. $\boxed{1/2}$

$$\mathbf{X}=\begin{bmatrix}1 & -5 & 10 & 2 & -3\\ 2 & 10 & -5 & -1 & 4\end{bmatrix}\quad \boldsymbol{\mu}=\begin{bmatrix}\qquad\qquad\end{bmatrix}\quad \boldsymbol{\Sigma}=\begin{bmatrix}\qquad\qquad\qquad\qquad\qquad\end{bmatrix}\tag{1}$$

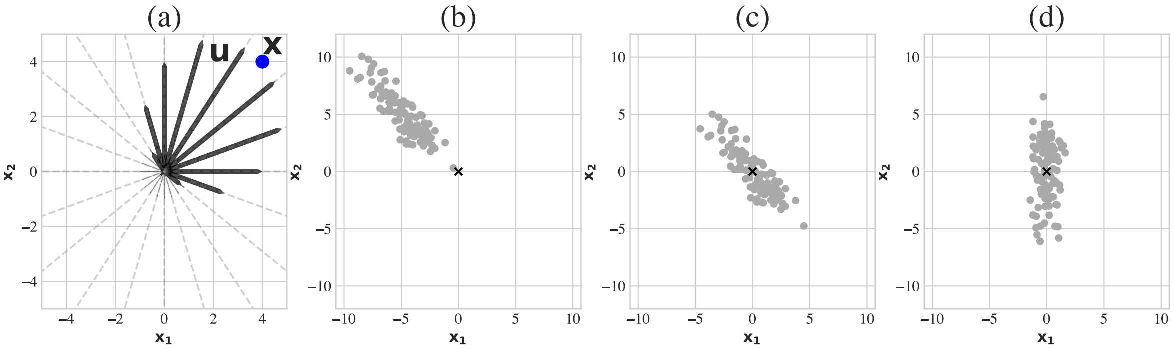

Figure 1: Projection and Point cloud

   (b) Referring to Fig. 1 (a), assume you have a training point: $\mathbf{x} \in \mathbb{R}^2$ and a unit vector $\mathbf{u}$—thus $||\mathbf{u}||_2 = 1$—that functions as a direction passing through the origin. Define with linear algebra the projection of $\mathbf{x}$ over $\mathbf{u}$. Now $\mathbf{x}$ is fixed and you can rotate $\mathbf{u}$: how can you set $\mathbf{u}$ to maximize the projection length? What is the maximum value of the projection length? Black segments in Fig. 1 (a) indicate the projection length over varying directions $\mathbf{u}$. $\boxed{3}$

   (c) A 2D point cloud $\mathbf{X} \doteq \{\mathbf{x}_i\}_{i=1}^N$ is shown in Fig. 1 (b). Fig. 1 (c) shows the same but centered $\bar{\mathbf{X}}$. How do we center the point cloud $\mathbf{X}$ to $\bar{\mathbf{X}}$? Assuming $\mathbf{X} \in \mathbb{R}^{N\times 2}$, which means is given to you as a matrix of N rows and 2 columns, write the one liner `numpy` code to perform the centering. What does `numpy` try to do when shape of matrices do not match? $\boxed{1}$

   (d) Given the centered point cloud $\bar{\mathbf{X}}$ in (c), which transformation you apply to make it as Fig. 1 (d)? How do you compute this transformation? After the transformation, what happens to the covariance matrix? $\boxed{3}$

Total for Question 1: 7½

2. We have to build a **Gaussian Mixture Model (GMM)**, from a training set of data. Each data point lives in a space such as $x_1 \in \mathbb{R}^1$. The assignment z of the GMM are known and given to you already as $z \in \{0, 1, 2\}$, for each training point—see Tab. 1. Assume the estimate for GMM is maximum likelihood (MLE).

x_1	11	3	-1	10	-5	-6	-4	2	4	1	-2	-3
z	2	1	0	2	2	1	2	0	1	0	0	2

Table 1: Training set of a GMM with assignments.

(a) How many modes does the GMM described above have? Please, motivate your answer. ½

(b) Give a definition and mathematically describe what is the probability density function used in GMMs. Write down the name of the distribution if your recall it. 2

(c) Compute the density function given the training data in Tab. 1 *(Hint: to compute it, you need just to specify which distribution you have in GMM, and compute the parameters of those distribution given the training set. It is OK to say: it distributes as distribution A with parameters indicated by the B and C and offer numerical values for B and C.)* 2½

(d) Each mode of the GMM models $p(\mathbf{x}|z)$. Given $\mathbf{x}'$ as new unseen input, how could you compute the probability $p(z = 0|\mathbf{x}')$? 2

(e) After you have fit the GMM, let's say that you want to sample a new point from the generative model behind the GMM. What are the steps necessary to sample from your estimated GMM? *(Describe how sampling works step by step, which distributions you sample from using which technique.)* 1

Total for Question 2: 8

3. Given the training points below for $y \in \{0, 1\}$ binary classification:

$$(x_1 = 1/4; y_1 = 1) \quad (x_2 = 1/2; y_2 = 0) \quad (x_3 = 3/4; y_3 = 1)$$

(a) Determine the output of a **K Nearest Neighbour (K-NN)** classifier for all points on the interval $0 \leq x \leq 1$ using: • 1-NN and • 3-NN. *(You have to write explicitly how the [0,1] interval is classified though you can augment your answer with drawing if you want)* $\boxed{3}$

(b) Assume you want to regress continuous values—thus $y \in \mathbb{R}$. The regressed output is the mean of the **K Nearest Neighbour (K-NN)** of a test point. Determine the ouput on the interval $0 \leq x \leq 1$ using the same training data above for $K = 2$. $\boxed{1\ ½}$

(c) Your mate tells you that he/she got perfect (100%) training accuracy by classifying the training data with K-NN when $K = 1$. Also, adds that increasing K is a bad idea because training accuracy drops. What is happening and how would you reply? What is the effect of increasing K? *(Explain clearly all details as much as possible)* $\boxed{1\ ½}$

Total for Question 3: 6

4. We are given a training set, where each attribute $\mathbf{x}$ describes if a patient 1) had cough 2) had soar throat 3) had fever; the $y \in \{\text{C, F, H}\}$ indicates having contracted COVID-19 (C); normal flu (F) or being healthy (H).

Patient	Coughing	Sore Troath	Fever	$y = \{\text{Covid, Flu, Healthy}\}$
$\mathbf{x}_1$	Y	Y	Y	C
$\mathbf{x}_2$	Y	N	Y	F
$\mathbf{x}_3$	Y	N	N	H
$\mathbf{x}_4$	Y	Y	N	H
$\mathbf{x}_5$	N	Y	Y	C
$\mathbf{x}_6$	N	N	Y	F

Table 2: Training set for disease classification.

(a) Using the training data above, construct a **decision tree** for the 3-class classification problem above. Use the Information Gain (IG) with **entropy** as impurity function, as the decision criterion to select which attribute to split on. Show your calculations for the IG for all possible attributes for just the first split. Draw the configuration of the resulting tree. $\boxed{3}$

(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 trainng set in Tab. 2. $\boxed{2}$

(c) Let us assume that you have pairs of points as $\{x_i, y_i\}_{i=1}^N$ where y is a continuous value in $\mathbb{R}$ from a unknown function $y = f(x)$. Is it possible to learn the function $f(\cdot)$ with a tree? Can you tell how the tree can approximate f? Which loss function are you going to minimize? $\boxed{1}$

Total for Question 4: 6

5. You work as a data scientist for `VisionGrad` an hot startup working with automatic differentiation. Your job is to analyze the computational graph shown in Fig. 2.

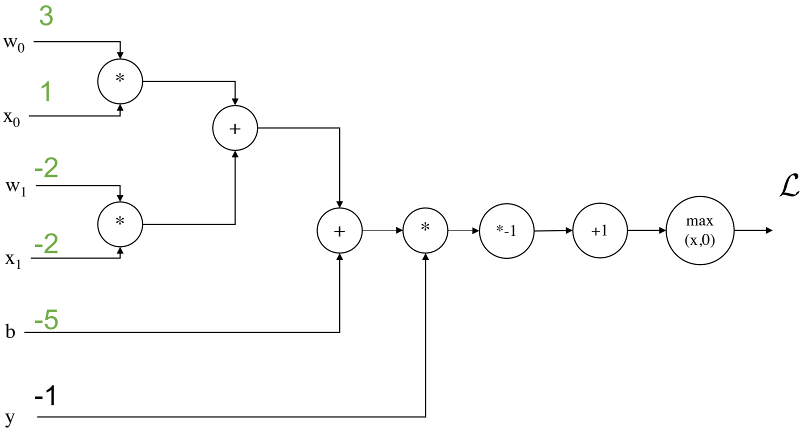

Figure 2: Computational Graph

(a) Given the graph, write down the function in a vectorized form, that implements the graph. Assume $y \in \{1, -1\}$. In which algorithm we have encountered a loss function similar to one in the graph above? Explain what the loss does with as much as details as possible for each single step. $\boxed{1}$

(b) Fill in the graph to compute the forward pass and backward pass: compute the derivatives over all the inputs (e.g. $\frac{\partial \mathcal{L}}{\partial w_0}, \frac{\partial \mathcal{L}}{\partial w_1}, \frac{\partial \mathcal{L}}{\partial x_0}$, etc.), showing also the intermediate values of those. Write the forward pass value $f(x)$ above each gate, write $\frac{\partial f(x)}{\partial x}$ below each gate. Skip computation on y. $\boxed{3}$

(c) Given the partial derivatives you computed on the inputs, let's assume that you can perturb only a single input with $+1$ or -1. Which input do you perturb to decrease the loss the most? Please, specify if you are adding or subtracting one. Given your perturbation, can you compute the value of the new loss after your perturbation yet **without forwarding the new input in the graph**? If yes, motivate how you can do it and write the new value of the loss; if no, explain why. $\boxed{3\frac{1}{2}}$

Total for Question 5: $7\frac{1}{2}$

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

Question:	1	2	3	4	5	Total
Points:	7½	8	6	6	7½	35
Score:						