Artificial Intelligence and Machine Learning - Unit II

02 February 2023

1. Answer the following questions by reporting the mathematical procedure. If you have to compute the actual value, please write the procedure that leads you to the numerical values. **Read well the text before proceeding!**

   (a) In Eq. (1) left, $\mathbf{X}$ is a design matrix where each column is an attribute (or feature), for a 2-D feature vector. How many samples are present in the design matrix $\mathbf{X}$? Complete the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to compute the empirical average and the covariance matrix associated with $\mathbf{X}$. $\boxed{2}$

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ -1 & -1 \\ 1 & 1 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \qquad\qquad\qquad\qquad \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \qquad\qquad\qquad\qquad \\ \\ \\ \end{bmatrix} \tag{1}$$

   (b) Our machine learning algorithm has just learned a 2×2 linear transformation $\mathbf{T}$ in the form of a matrix. We would like to analyze the transformation $\mathbf{T}$. So we sample a set of points $\mathbf{X} \doteq \{\mathbf{x}_i\}_{i=1}^{N}$ on a 2D unit circle centered in the origin; we then map $\mathbf{X}$ through $\mathbf{T}$ and we may end up with an ellipsoid. Define mathematically how you can recover: $\boxed{3}$

   ◇ the direction of the principal axis of the ellipsoid

   ◇ describe also how you can recover the length of those directions/axes

   ◇ assume you want to project the ellipsoid points onto the major axis, write how you implement the projection and how you select the "major" axis.

   Note: you do not have to compute it numerically, just write the "math" and explain it.

   An example in 2D is shown in Fig. 1 for the transformation $\mathbf{T} = \begin{bmatrix} 1.5 & 2. \\ 2. & 0.5 \end{bmatrix}$.

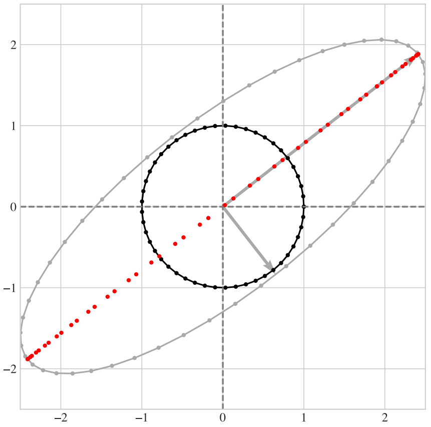

Figure 1: The transformation $\mathbf{T}$ maps the points on the circle as an ellipsoid.

   (c) Write a proof sketch that shows that in order to find the unit direction $\mathbf{u}$ that maximizes the variance of the projection on $\mathbf{u}$ of the centered data $\bar{\mathbf{X}}$, you need to solve an eigenvalue/vector problem. Please motivate each step in the proof. Remember that the centered covariance matrix is $\frac{1}{N}\bar{\mathbf{X}}\bar{\mathbf{X}}^T$. $\mathbf{x}_i$ is the i-th centered sample. Start from: $\boxed{2\frac{1}{2}}$

$$\arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} ||(\mathbf{x}_i^T\mathbf{u})\mathbf{u}||_2^2 \quad \text{s.t.} \quad ||\mathbf{u}||_2 = 1 \tag{3}$$

   Which algorithm do we have encountered in the course that uses this optimization criterion as loss function?

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

2. We are in the i-th step of the Expectation-Maximization (EM) for learning the parameters of a GMM. Let us assume the Expectation part just finished. The responsibilities γ for each training point x are given in Tab. 1 along with the training points x in 1D. Assume the estimate for GMM is maximum likelihood.

x	1	0	10	-5
γ	[.15, .15, .6, 0.1]	[.25, .25, .25, .25]	[.9, 0.034, 0.033, 0.033]	[1, 0, 0., 0.]

Table 1: Training set of a GMM with responsibilities.

(a) How many modes does the GMM described above have? Please, motivate your answer. $\boxed{1}$

(b) Define the responsibilities γ from a mathematical point of view and explain which kind of information they provide. For an arbitrary point $\mathbf{x}$ associated to a responsibility γ what is the meaning of $\gamma[3] = 0.01$, where 3 indexes the third value of the γ vector? $\boxed{1\frac{1}{2}}$

(c) Given the responsibilities and the training point defined in Tab. 1, compute numerically the Maximization Step, that is, estimate the probability density function (pdf) of the GMM at that step. Please, write the equations you are using for the computation. *(Hint: to compute the pdf you just have to find the parameters of the GMM and then say it distributes according to e.g.,* $\frac{1}{5} \cdot \mathcal{N}(\frac{1}{2}, 29) + \ldots$ *etc.)* $\boxed{2}$

(d) Let's assume that you have an image $\mathbf{x}$ of dimension H×W. Each pixel is expressed as a linear combination of red, green and blue, where each color component is quantized with 8 bits. $\boxed{2\frac{1}{2}}$

◇ How many bits you have to transmit in total if you want to send this image to a friend over the internet? See Fig. 2 for a sketch.

Now assume that you cluster all the pixels of $\mathbf{x}$ in the RGB color space with k-means and you impose 4 clusters in total.

◇ Given the clustering result, describe what you need to send to your friend so that you can save bandwidth and he is able to approximately reconstruct the image.

◇ How many bits do you have to send now?

◇ Explain if you are you really saving bandwidth or not.

Assume your friend knows a) the dimension H×W b) the order that you are using to send the pixels—row by row top to bottom, for each row left to right c) he knows how to interpret additional information of the clustering.

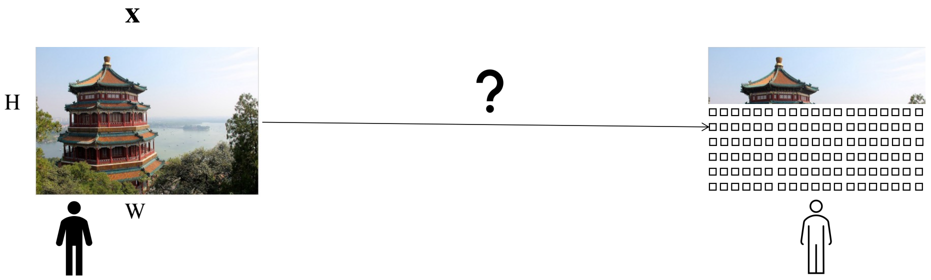

Figure 2: Image compression in colorspace with clustering.

Total for Question 2: 7

3. We are given an already learned decision tree for multi-class classification shown in Fig. 3 below. Each shape represents a training sample where the shape identifies the class. The feature of each point are two-dimensional $\mathbf{x} = [x_1, x_2]$.

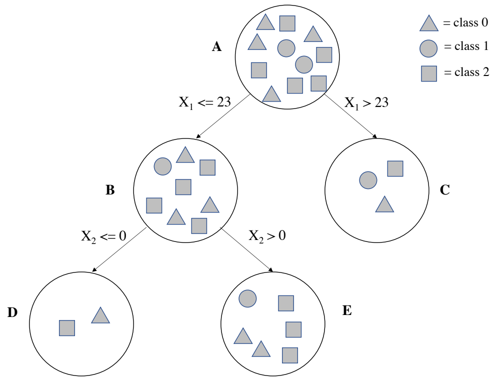

Figure 3: Decision Tree for multi-class classification.

(a) Compute the Impurity using the entropy **for the entire TREE.** $\boxed{2}$

(b) You have a test sample $\mathbf{x} = [x_1, x_2] = [-1, -1000]$, which class you assign to $\mathbf{x}$—triangle, square or circle or maybe it is better to return that the algorithm is unsure? $\boxed{1}$

(c) Let us assume that you have enough computational power that you can afford to learn and do inference on multiple decision trees. At the same time you are troubled because the single decision tree that you learned is over-fitting badly. Describe with more details you can which technique you can use to mitigate overfitting in this case. $\boxed{1\frac{1}{2}}$

(d) Let's assume that you are in a leaf of the tree, e.g. the node E, and you want to make the decision tree a randomized algorithm for prediction. Describe how you can make the decision tree prediction random with probabilities that are given by the labels you find in the leaf. $\boxed{1}$

Total for Question 3: $5\frac{1}{2}$

4. We want to perform some evaluation of a binary classifier using logistic regression that reports probability for an input $\mathbf{x}$ being of class $+1$ as $p(y = +1|\mathbf{x})$. The ground-truth labels y_{gt} are show in Tab. 2 as positive labels $+1$ and negative labels -1.

y_{gt}	+1	-1	-1	-1	-1	+1	-1	+1
$p(y = +1\|\mathbf{x})$	0.25	0.01	0.3	0.01	0.165	0.15	0.02	0.5

Table 2: Labels and probabilities for a binary classifier.

(a) Give a definition of True Positive Rate (TPR) and False Positive Rate (FPR). Given a binary classifier with probability of $\mathbf{x}$ being positive $p(y = +1|\mathbf{x})$—compute the ROC curve for the values in Tab. 2 by showing the TPR and FPR in a table. $\boxed{2}$

(b) Compute the Area Under the Curve (AUC) of the above ROC. $\boxed{2}$

(c) Setting the classifier threshold to 0.1, compute the confusion matrix for the values in Tab. 2. For each cell of the confusion matrix, indicate what is the metric computed. $\boxed{1\frac{1}{2}}$

(d) Two interns are working on a machine learning approach to spam detection. Each student has their own set of 100 labeled emails, 90% of which are used for training and 10% for validating the model. Student A runs a k-NN classification algorithm and reports 80% accuracy on her validation set. Student B experiments with over 100 different learning algorithms and variations of them, training each one on his training set, and recording the accuracy on the validation set. His best formulation achieves 95% accuracy. Whose algorithm would you pick for protecting a corporate network from spam? Motivate your answer. $\boxed{1\frac{1}{2}}$

Total for Question 4: 7

5. You are given to study the following formulation for binary classification $\{-1, +1\}$:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \doteq \frac{1}{1 + \exp^{-\boldsymbol{\theta}^{\top}\mathbf{x}}} \tag{4}$$

where $p(y = +1|\mathbf{x}) \doteq f_{\boldsymbol{\theta}}(\mathbf{x})$ and $p(y = -1|\mathbf{x}) = 1 - p(y = +1|\mathbf{x})$

(a) Assuming that the input feature $\mathbf{x}$ has 100 dimensions, what is the dimensionality of the vector $\boldsymbol{\theta}$ that you have to learn without a bias? $\boxed{1}$

(b) What is the role of the function $\frac{1}{1+\exp^{-z}}$ and what is usually called? $\boxed{1}$

(c) Now considering Eq. (4), the learned parameters $\boldsymbol{\theta} = [0.82, -0.46, 0.1]$ and each sample in Tab. 3. $\boxed{1\frac{1}{2}}$

◇ Compute $\boldsymbol{\theta}^{\top}\mathbf{x}$ for each sample.

◇ Compute the probability for each sample of being positive.

◇ Classify $\hat{y}$ the training samples as either positive (+1) or negative (-1) thresholding the probability at 50%.

$\mathbf{x}$	[-2, -2, 1]	[-1, -1, 1]	[0, 0, 1]	[1, 1, 1]	[2, 2, 1]	[3, 3, 1]
$\boldsymbol{\theta}^{\top}\mathbf{x}$	______	______	______	______	______	______
$p(y = +1\|\mathbf{x})$	______	______	______	______	______	______
$\hat{y}$	______	______	______	______	______	______

Table 3: Testing point for the probabilistic classifier in Eq. (4).

(d) Describe how you can extend the classifier in Eq. (4) from binary classification to multi-class classification: explain the new activation function and how you need to change $\boldsymbol{\theta}$, assuming that now you have 4 classes. $\boxed{2}$

(e) 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 $f(x)$ <u>above</u> each gate, write $\frac{\partial f(x)}{\partial x}$ <u>below</u> each gate. $\boxed{2\frac{1}{2}}$

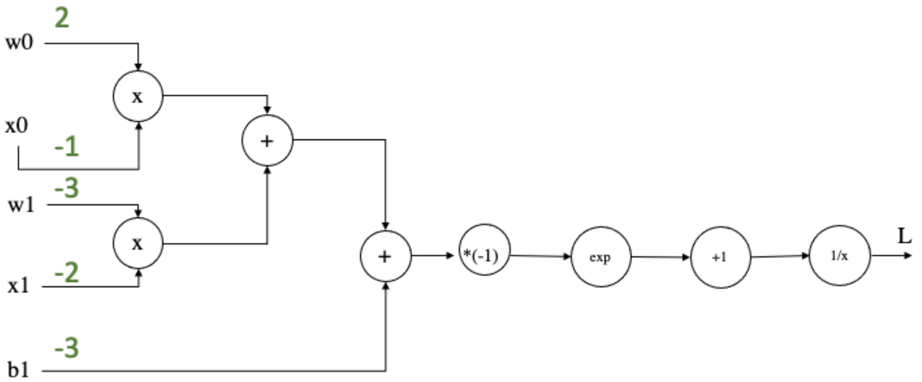

Figure 4: Computational Graph

Total for Question 5: 8

You can use this space for writing. The summary of points is at the bottom.

Question:	1	2	3	4	5	Total
Points:	$7\frac{1}{2}$	7	$5\frac{1}{2}$	7	8	35
Score:						