Artificial Intelligence and Machine Learning - Unit II
12 July 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 itself.

   (a) In Eq. (1) left, $\mathbf{X}$ is a design matrix where each row is a sample. Each row indicates RGB pixel triplet. What is the dimensionality of the samples in $\mathbf{X}$? Complete the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to compute the empirical RGB average and the covariance matrix associated to $\mathbf{X}$. [1]

$$\mathbf{X} = \begin{bmatrix} 172 & 47 & 117 \\ 192 & 67 & 251 \\ 195 & 103 & 9 \\ 211 & 21 & 242 \\ 36 & 87 & 70 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \quad \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \quad \end{bmatrix} \tag{1}$$

   (b) Let's say that you have a RGB image $\mathbf{I}$ and each color channel is quantized over 8 bit. Colors range from 0 to $2^8 - 1$ with height 640 pixels and width 480 pixels. $\mathbf{I}$ shape is thus $640 \times 480 \times 3$. [3]
   ◇ How many images are possible to be sampled in total in the space where the image lives in?
   ◇ You want to transform the image into grayscale. Assume that you can do that by computing the mean across color channel. Write the `one-liner numpy` code that implements that. Please, be specific on the axis. ◇ What is the size of the tensor after you have computed the mean?

   (c) A point cloud $\mathbf{X} \doteq \{\mathbf{x}_i\}_{i=1}^N$ is sampled from a multi-variate Gaussian distribution in 2D and shown in Fig. 1 (a), the black diamond indicate the mean of the distribution. Fig. 1 (b) shows the same point cloud after "sphering" the data. Describe the mathematical process of sphering in the data. What happens to the covariance matrix of the point cloud in Fig. 1 (b)? *Though the example is in 2D, the solution should work in N dimensions too.* [3]

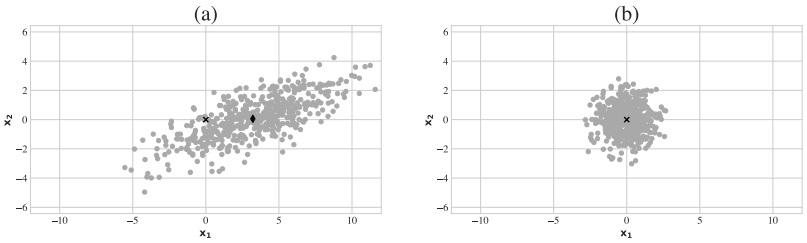

Figure 1: Point cloud and "Sphering"

Total for Question 1: 7

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 $\boldsymbol{\gamma}$ for each training point x are given in Tab. 1 along with the training points x. Assume the estimate for GMM is maximum likelihood.

x	18	8	-6	-13	0
$\boldsymbol{\gamma}$	[0.8, 0.1, 0.1]	[0.3, 0.2, 0.5]	[0.1, 0.7, 0.2]	[1, 0, 0]	[0.99, 0.005, 0.005]

Table 1: Training set of a GMM with responsibilities.

(a) How many modes does the GMM described above have? Please, motivate your answer. [1]

(b) Define the responsibilities $\boldsymbol{\gamma}$ from a mathematical point of view and explain which kind of information they provide. For an arbitrary point $\mathbf{x}$ associated to a responsibility $\boldsymbol{\gamma}$ what is the meaning of $\boldsymbol{\gamma}[0] = 0.8$, where 0 indexes the first value of the $\boldsymbol{\gamma}$ vector? [2]

(c) Given the responsibilities and the training point defined in Tab. 1, compute 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.* [3]

(d) You have written your own implementation of EM to fit GMM and you are fitting some data, but at some point, you suddenly see that the log-likelihood goes to infinity—you see `np.nan` or `np.inf` appearing in the log-likelihood. What may be happened? *(Hint: division by zero is a good starting point but needs to be motivated in to the context of GMM.)* [1]

Total for Question 2: 7

3. We have 2D training points below to be used for binary classification; each point is paired with its label $y \in \{0, 1\}$:

$$\mathbf{a} = [1/4, \sqrt{2}]; y_a = 1 \qquad \mathbf{b} = [4, -3\sqrt{2}]; y_b = 0 \qquad \mathbf{c} = [-2, 2\sqrt{2}]; y_c = 1 \tag{2}$$

(a) Irrespective of Eq. (2), give a generic definition of the distance between two points in D dimensions $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{z} \in \mathbb{R}^D$ using $\diamond\ \ell_2$ (Euclidean), $\diamond\ \ell_1$ (Manhattan) and $\diamond\ \ell_\infty$ norms. 2

(b) Using the training set specified above, classify a new point $\mathbf{x}' = [0, 0]$ using **K Nearest Neighbor (K-NN)** with k=2 using: $\diamond\ \ell_2\ \diamond\ \ell_1$ and $\diamond\ \ell_\infty$ norms to measure the distance between points. Show your computation of the distances and the way you classify $\mathbf{x}'$. 2

(c) Let's assume you have a k-NN algorithm and you want to cross-validate $k = \{7, 11, 21\}$ and three types of distanced ℓ_1, ℓ_2 and ℓ_∞. How many models do you have to train in total, assuming you do 10-fold cross-validation? Moreover, what is the complexity for training a k-NN classifier? 2

Total for Question 3: 6

4. We want to perform some evaluation of a binary classifier.

y	1	0	1	1	0	0	0
s	0.7	-5	0.3	0.1	-1	5	0

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 correlates with y—compute the ROC curve for the values in Tab. 2 by showing the TPR and FPR in a table. [3]

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

(c) Let's assume that in Tab. 2 we replace the score $s = -5$ with $s = -2.2$; we also replace $s = 0.3$ with $s = 0.22$. Is the ROC going to change? Do we have to recompute it? Motivate your answer. [1]

(d) Alice works for `IseekU`, a biometric company using AI, and she is happy since she developed a "perfect" classifier: it achieves 99.17% AUC in the validation set over $10K$ samples. Alice says "it is ready to be employed in practice since it will never generate false alarm". What would you tell Alice? What Alice should measure if the company wants a quota "X" on the false alarms? [2]

Total for Question 4: 8

5. We have to analyze a neural network in the form of a multi-layer perceptron (MLP). The neural network details follow in the sub questions below.

   (a) Deduce and write how many trainable parameters you have with a MLP with input feature vectors with dimension equal to 1024, a first layer with 512 units/neurons, a second layer with 256 units/neurons, and a final multi-class classification layer with 3 units/neurons. Assume all layers have the bias term. The network uses ReLU activation function after each layer except the classification that uses softmax. Write down the equation for the computation, not just the final value. Moreover, how many classes does the network classify? [1]

   (b) Give a definition of the softmax function and cross-entropy loss used for training neural nets for classification. [3]

   (c) Let us suppose that in the previous network we arrive at the last classification layer and define $\mathbf{z} \doteq \mathbf{W}\mathbf{x} + \mathbf{b}$ as the response of the linear layer for classification. $\mathbf{z}$, as you may know, has to go through the softmax function. Given the vector $\mathbf{z}$ below and a $\mathbf{y}$ as one-hot encoding for class label $\diamond$ compute the probability values after softmax, by filling in the values in $\mathbf{p}$ $\diamond$ compute the value of the cross-entropy loss $\mathcal{L}$, given $\mathbf{p}$ and $\mathbf{y}$. When computing cross-entropy use natural logarithm. [3]

$$\mathbf{z} = \begin{bmatrix} -0.5 \\ 0.5 \\ -1 \end{bmatrix} \qquad \mathbf{p} = \begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \mathcal{L} = \_\_\_\_\_$$

   What is the value of the loss instead if the network perform random guess?

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:						