Artificial Intelligence and Machine Learning - Unit II
12 January 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 $\mathbf{X}$. 2

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

   (b) We have a point cloud $\{\mathbf{x}_i\}_{i=1}^{N}$ of data points that live in a two dimensional space and are generated from some unknown multivariate normal distribution $\mathcal{N}$. See Fig. 1(a) for an example a 2D point cloud. ⋄ Define mathematically how you can recover an estimate of the parameters of the unknown distribution $\mathcal{N}$. ⋄ Define how many parameters you have to estimate and what is their dimensionality in case of points that live in D dimensions. ⋄ Now consider that you have a query vector $\mathbf{q}$ that lives in the same space. We have to compute the distance between the point $\mathbf{q}$ against your recovered estimate of $\mathcal{N}$: mathematically define how to compute the distance between the point $\mathbf{q}$ and the distribution induced by the point cloud $\{\mathbf{x}_i\}_{i=1}^{N}$. The distances for all the points in the plane are shown in Fig. 1(b) using an heatmap: brighter means smaller distance and dark means larger distances. Points on the plane with the same color have the same distance. 3

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

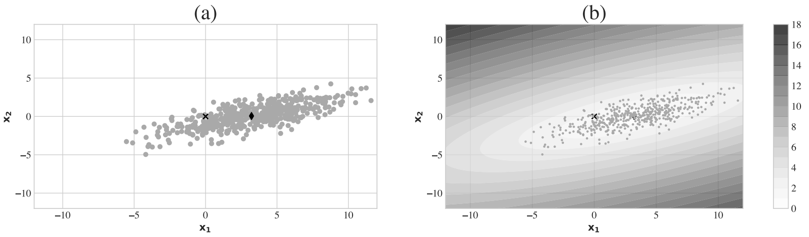

Figure 1: (a) A 2D point cloud (b) distances on the plane wrt to the point cloud as an heatmap. The range of distance value is shown in the right with a colorbar.

   (c) You are hired by a famous company that has to compute ℓ_2^2 distance between two feature vectors $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^D$. The company works with a machine learning algorithm that garantees that the features $\mathbf{x}$ and $\mathbf{y}$ lie on a unit hyper-sphere *(Hint: same as $||x||_2^2 = 1$)*. You have to compute: 2½

$$d^2 = ||\mathbf{x} - \mathbf{y}||_2^2 = \sum_{i=1}^{D} (x_i - y_i)^2$$

   ⋄ Can you get away computing this without computing the squared difference, i.e. avoiding computing $(x_i - y_i)^2$?

   ⋄ What is the minimum and maximum value that d can take?

   ⋄ Now assume that you have to compute the distance between a matrix of features $\mathbf{X}$ against $\mathbf{Y}$. Is the approach that you developed above still usable? If no explain why; if yes, explain how you can implement it and if you can do it with a few lines of code.

Total for Question 1: 7½

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 in 1D. Assume the estimate for GMM is maximum likelihood.

x	-1	0	1	7
$\boldsymbol{\gamma}$	[.15, .15, .7]	[.33, .33, .34]	[.25, .5, .25]	[1, 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 $\boldsymbol{\gamma}$ 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 $\boldsymbol{\gamma}[1] = 0.15$, where 1 indexes the second value of the $\boldsymbol{\gamma}$ vector? $\boxed{2}$

(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) Explain both with your own words and with math, what is the meaning of recovering an estimate of the parameters $\boldsymbol{\theta}$ given data points $\mathbf{X}$ using a maximum likelihood (MLE) approach. $\boxed{2}$

Total for Question 2: 7

3. Considering the k-NN algorithm, answer the following questions.

(a) Fig. 2 shows the 2D feature space for two sets of datapoints belonging to the gray triangle label or to the black square label. Taking into consideration the query point star $\star$, list all possible values of k that you can set in the k-NN algorithm so that the query star$\star$ will be classified as black square, using the ℓ_2 norm as distance function. $\boxed{2}$

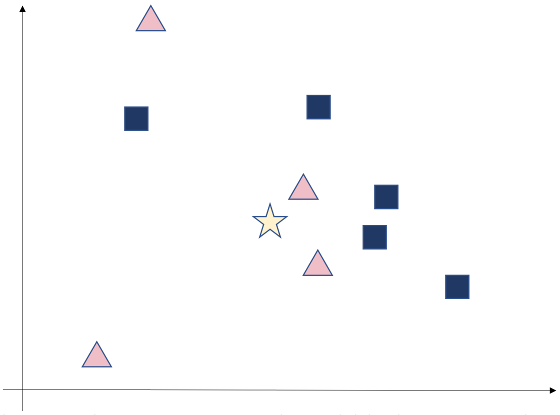

Figure 2: Feature vector for the two classes gray triangle and black square. The query point is indicated with the star $\star$.

(b) Define the set $S_{\mathbf{x}}$ of the k nearest neighbors of the input feature vector $\mathbf{x}$ given a dataset $\mathbf{D}$. $\boxed{2}$

(c) Describe the type of decision boundary we can model with k-NN. After you have described them, answer the following: $\diamond$ Between a k-NN classifier and a linear Support Vector Machine (SVM) which can model more complex decision boundaries? $\diamond$ If you notice that a classifier has very fragmented decision boundary, does the classifer suffer from underfitting or overfitting? $\boxed{2\ \frac{1}{2}}$

Total for Question 3: $6\frac{1}{2}$

4. We want to perform some evaluation of a face detector on an image. The image is shown in Fig. 3. The image contains 7 ground-truth faces from seven subjects. We run a face detector that outputs gray bounding boxes were it believes there is a face. A face is considered to be detected correctly if the bounding box is centered on the ground-truth head.

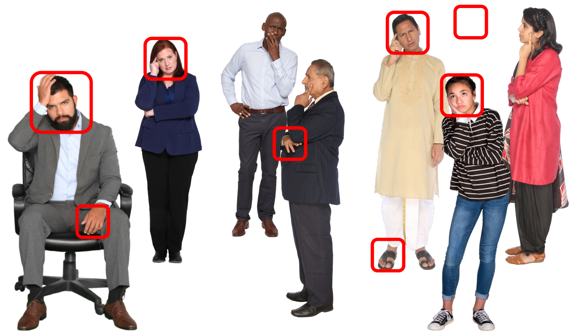

Figure 3: Image with faces and the detected faces with gray bounding boxes.

(a) Give a definition of Precision (P) and Recall (R) and then compute them numerically on the example given in Fig. 3. $\boxed{2}$

(b) Let us assume that we tune the face detection algorithm so that it places a gray bounding box everywhere in the image so to cover it fully and densely, for a total of 10 millions bounding boxes, as a brute force approach. Described what happens to the precision and the recall. $\boxed{2}$

(c) Using the previous definition of precision and recall, consider a PR curve such as that in Fig. 4. Explain why and how the precision-recall curve can have the zig-zag non monotonic trend showed in the figure. In other words explain the existence of 1) The curve climbs up slowly 2) or vertical drops. $\boxed{2}$

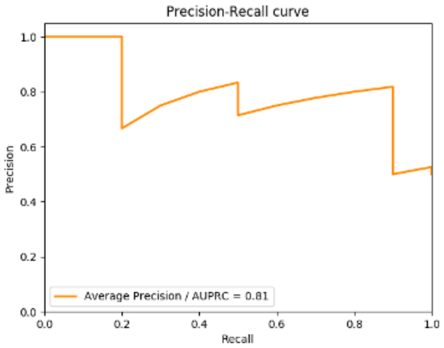

Figure 4: A Precision Recall curve

Total for Question 4: 6

5. You are given a linear SVM for binary classification.

(a) Write down and explain the equation that a linear SVM uses for binary classification and the equation used compute the geometric margin. $\boxed{2}$

(b) Now considering an SVM working in 2D with hyper-plane $\mathbf{w} = [0.82, 0.46]$ and bias $b = -0.17$. ⋄ Using the equation that you wrote in point (a) and the SVM provided values, classify the training samples in Tab. 2 as either positive (+1) or negative (-1). ⋄ Compute the geometric margin for all points. $\boxed{2}$

x	[-2, -2]	[-1, -1]	[0, 0]	[1, 1]	[2, 2]	[3, 3]
y	______	______	______	______	______	______
γ	______	______	______	______	______	______

Table 2: Testing point for the SVM, we have to find the labels and the geometric margins.

(c) It is given the computational graph in Fig. 5 that implements Hinge Loss for SVM. The forward pass to compute the loss and backward pass to compute the gradients over the input are already computed. $\boxed{3}$
⋄ Compute the new gradients in case we scale the loss by 5 and then sum 10 to it.
⋄ In case you have already solved the original graph, can you predict how the gradient will change in case the loss is subject to a modification of the type $\alpha\mathcal{L}(w_0; w_1; x_0; x_1; y) + \beta$ and explain why you can do it without using the computational graph? *(e.g. compute $\nabla_{x_0}\alpha\mathcal{L}(w_0; w_1; x_0; x_1; y) + \beta$)*

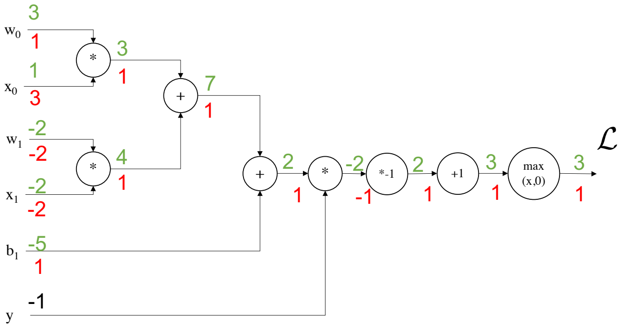

Figure 5: Computational graph.

Total for Question 5: 7

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	$6\frac{1}{2}$	6	7	34
Score:						