**Artificial Intelligence and Machine Learning - Unit II**
14 September 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). 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]

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \quad\quad\quad\quad\quad\quad \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \quad\quad\quad\quad\quad\quad \end{bmatrix} \tag{1}$$

   (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 of the 2D point cloud. The same point cloud undergoes some rotation and a translation that are unknown. They "move" the point cloud to Fig. 1(b). Describe an algorithm that you can use to align the point cloud so that they overlap perfectly—each point overlaps with itself. [3]

   Note: Remember that PCA returns to you the direction of the principal components (PC) but with up to an arbitrary sign (orientation is unknown).

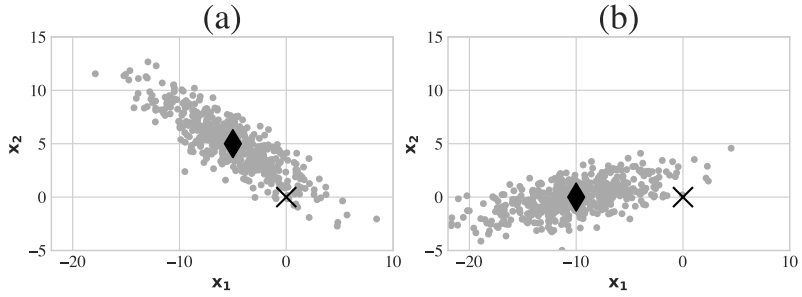


**Figure 1:** (a) A 2D point cloud on the left is translated and rotated to (b). How can you align them?

   (c) You are hired by a famous company that has to compute $\ell_2^2$ 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]

$$d^2 = ||\mathbf{x} - \mathbf{y}||_2^2$$

   ◇ Write the definition of $\ell_2^2$ euclidean distance for two vectors $\mathbf{x}$ and $\mathbf{y}$.

   ◇ 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 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\boldsymbol{\gamma}$ | [1, 0, 0,0] | [.15, .15, .35, .35] | [.25, .25, .25,.25,] | [.25, .5, .25,0] |

**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 $\boldsymbol{\gamma}[2] = 0.50$, where 2 indexes the third value of the $\boldsymbol{\gamma}$ vector? [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.)* [2]

Total for Question 2: 5

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 $\ell_2$ norm as distance function. 2

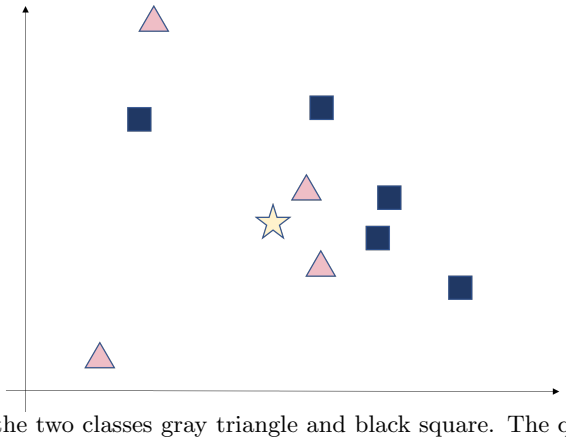

**Figure 2:** Feature vector for the two classes gray triangle and black square. The query point is indicated with the star $\star$.

(b) Assume that you have feature vectors where each feature has its own scale and also the values across features vary a lot. Describe if this can be a problem for $k$-NN and, if so, how to improve the $k$-NN classifier. 2

(c)
- Between a k-NN classifier and the perceptron, which can model more complex decision boundaries? 3
- If you notice that a classifier has very fragmented decision boundary, does the classifier suffer from underfitting or overfitting?
- Let's say that you want to make the decision boundary of a k-NN classifier less fragmented and more smooth: are you going to increase or decrease K? Explain your answer.

Total for Question 3: 7

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

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $y$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| s | 0.2 | 1 | 0.3 | 0.1 | -1 | -0.2 | 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 multiply all the score by $-0.5$. Is the ROC going to change? Do we have to recompute it? Explain what happens to the ROC. [1]

(d) Bob works for `IseekU`, a biometric company using AI, and she is happy since she developed a "perfect" classifier: it achieves 99.9% AUC in the validation set over $100K$ 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? [1]

Total for Question 4: 7

5. You are given data vectors that lives in a two dimensional space and you observe that they distribute like Fig. 3. You have two classes to be classified, indicated by two different colors.

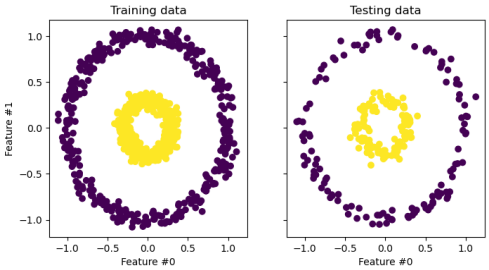


**Figure 3:** Data distribution of the features

(a)
- State if the distribution of the data in figure Fig. 3 can be classified with a linear classifier.
- Now let assume that you are able only to use a neural network with just a single, classification layer with biases. What is the total number of parameters you have to learn?

2

(b) The vectors in Fig. 3 have only two features (feature #0 $x_1$ and feature #1 $x_2$). Describe how you can augment the feature (i.e. by adding another one) so that the distribution becomes linearly separable.

2½

$$\text{feature} = [x_1,\ x_2,\ \_\_\_\_\_]$$

.

(c) Now let us say that the neural network mentioned before, with a single classification layer, is used to train a classifier for image classification with 10 classes. The images are gray scale of size $32 \times 32$ and are made to be a flatten vector. Describe a procedure that allows you to visualize in the input space what the network have learned to separate the data. At the end, the procedure should produce 10 images, one for each class.

3

Total for Question 5: 7½

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

| Question: | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Points: | 7½ | 5 | 7 | 7 | 7½ | 34 |
| Score: | | | | | | |