**CS5691: Pattern recognition and machine learning**

**Final exam**

**Course Instructor** : Prashanth L. A.

**Date** : May-3, 2019 **Duration** : 180 minutes **Max Marks**: 30

**Name of the student** :

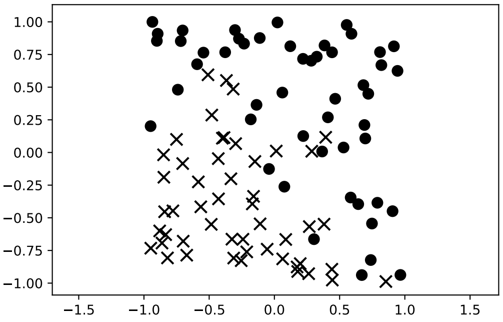
**Roll No** :

**INSTRUCTIONS**: For short answer questions, you do not have to justify the answer. For the rest, provide proper justification for the answers. Please use rough sheets for any calculations *if necessary.* Please **DO NOT** submit the rough sheets. DO NOT use pencil for writing the answers.

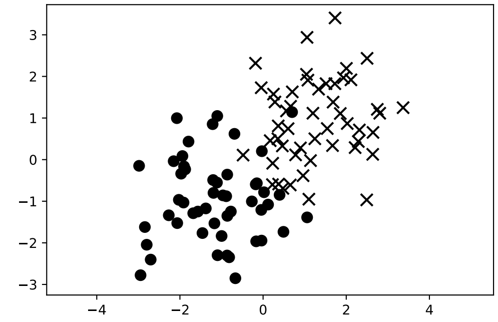
# I. Short Answer Questions

*Note: 1 mark for the correct answer. For some of the MCQs, more than one answer is correct. Please write the choice code(s). For other questions, provide the appropriate number or expression or a yes/no.*

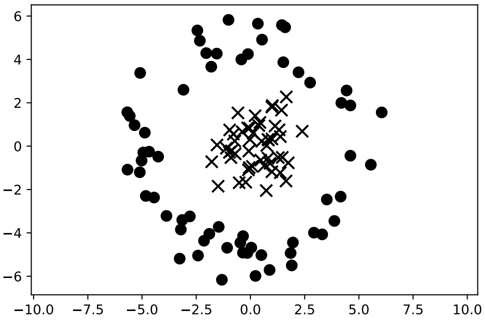
1. Consider a two-class classification problem. Recall that we assume $\mathbb{P}\left(y=+1 \mid \mathbf{x}\right) = \sigma(\mathbf{w}^\intercal \mathbf{x})$ in logistic regression. For each of the datasets below, predict whether the aforementioned assumption holds. (1 mark)



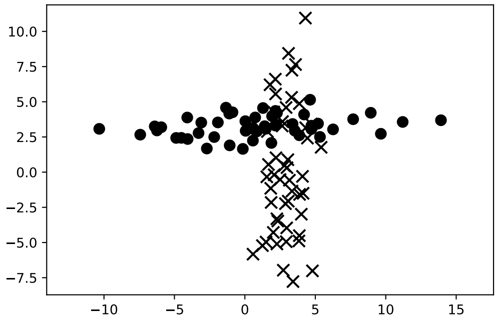
(a)



(b)



(c)



(d)

*Answer:*

| Dataset | Assumption satisfied? |
|---|---|
| (a) | |
| (b) | |
| (c) | |
| (d) | |

2. Let $\mathbf{N}(\mu, \sigma)$ be the univariate normal density, with $\sigma$=10000. Suppose we perform ML estimation assuming $\mathbf{N}(\mu, 1)$. Let the data $D_n = \{x_1, ...x_n\}$ be such that the number of samples $n$ is large. Is the ML estimate, say $\hat{\mu}_{ML}$, close to the true $\mu$? (1 mark)

   *Answer:*

3. Find the line of the form $wx + b$ that minimizes the squared error for the following $1d$-regression problem: (1 mark)

   | $x$ | $y$ |
   |---|---|
   | 1 | 1 |
   | 2 | 2 |
   | 4 | 3 |
   | 5 | 4 |
   | 6 | 4 |

   *Answer:*

4. Let $X|Y = -1$ be distributed according to a mixture of two Gaussians given by $\frac{1}{2}\mathcal{N}(3, 1) + \frac{1}{2}\mathcal{N}(9, 1)$. Let $X|Y = +1$ be distributed as $\mathcal{N}(6, 1)$. Give the Bayes classifier for the zero-one loss function and Bayes error (or the probability of mis-classification). (2 marks)

   *Answer:*

5. Let $W \sim \mathcal{N}(\mathbf{0}, \rho^2 I)$ be a vector in $\mathbb{R}^d$. Let $X$ be distributed according to some distribution over $\mathbb{R}^d$. Let $Y = W^\top X + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is an independent scalar random variable. Given i.i.d. samples $(x_i, y_i)$, it is known that the MAP estimate of $W$ is the solution of the optimisation problem:

   $$\min_{\mathbf{w}} \sum_{i=1}^{m} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\mathbf{w}||^2.$$

   Give $\lambda$ in terms of $\rho$ and $\sigma$. (1 mark)

   *Answer:*

6. Let $x_1, \ldots, x_n$ be i.i.d. samples from $\mathcal{N}(3, \sigma^2)$ and $y_1, \ldots, y_m$ be i.i.d. samples from $\mathcal{N}(7, \sigma^2)$. Then which option below is the ML estimate of $\sigma^2$? (1 mark)

   (a) $\frac{1}{n} \sum_{i=1}^{n} (x_i - 3)^2$

   (b) $\frac{1}{m} \sum_{i=1}^{m} (y_i - 7)^2$

   (c) $\frac{1}{m+n} \left( \sum_{i=1}^{n} (x_i - 3)^2 + \sum_{i=1}^{m} (y_i - 7)^2 \right)$

(d) $\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-3)^2+\frac{1}{m}\sum_{i=1}^{m}(y_i-7)^2\right)$

*Answer:*

7. Consider the binary classification dataset below with 4 points.

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 0 | 0 | $+1$ |
| 0 | 1 | $-1$ |
| 1 | 0 | $-1$ |
| 1 | 1 | $+1$ |

We use a neural network with one hidden layer (of size 2) to fit this dataset, i.e.

$$\begin{aligned} z_1 &= \phi(w_{11}x_1+w_{12}x_2+b_1)\\ z_2 &= \phi(w_{21}x_1+w_{22}x_2+b_2)\\ \widehat{y} &= 2z_1+2z_2-1 \end{aligned}$$

where $\phi(t)=\max(0,t)$ is the ReLU function. The variables $z_1, z_2$ are the hidden nodes, and the variable $\widehat{y}$ is the prediction of the network. The $w, b$ values are the parameters of the network. We fix the last layer parameters for simplicity.

Which of the following parameter settings achieves $\operatorname{sign}(\widehat{y})=y_i$ for all the four data points? (1 marks)

(a) $w_{11}=w_{12}=1$, $b_1=0.5$, $w_{21}=w_{22}=-1$ and $b_2=-1.5$.

(b) $w_{11}=w_{12}=1$, $b_1=0.5$, $w_{21}=w_{22}=1$ and $b_2=-0.5$.

(c) $w_{11}=w_{12}=1$, $b_1=-1.5$, $w_{21}=w_{22}=-1$ and $b_2=0.5$.

(d) $w_{11}=w_{12}=-1$, $b_1=-0.5$, $w_{21}=w_{22}=-1$ and $b_2=0.5$.

*Answer:*

# II. Problems that require a detailed solution

1. **Support vector machines**

Consider the optimization problem for finding the maximum margin separating hyperplane, assuming that the classes are linearly separable:

$$\min_{\mathbf{w}\in\mathbb{R}^d}\frac{1}{2}\mathbf{w}^\intercal\mathbf{w} \quad \text{subject to } y_i\left(\mathbf{w}^\intercal\mathbf{x}_i+\mathbf{b}\right)\geq 1, i=1,\ldots,n, \tag{1}$$

where $\{(\mathbf{x}_i, y_i), i=1,\ldots,n\}$ is the training dataset.

Show that the value $\rho$ of the margin of the maximum-margin hyperplane is given by

$$\frac{1}{\rho^2}=\sum_{i=1}^{n}\alpha_i^*,$$

where $\alpha_i^*, i=1,\ldots,n$ are obtaining by solving the dual problem. (2.5 marks)

*Hint:* Use $\sum_{i=1}^{n}\alpha_i^* y_i=0$ and $y_i\left(\mathbf{w}^{*\intercal}\mathbf{x}_i+\mathbf{b}^*\right)=1, \forall i$.

2. **PAC-learning**

Answer the following: (1+4 marks)

(a) Given n $i.i.d$ samples from Unif$[0, c]$, $0 < c < \infty$, calculate an ML estimate for $c$.

(b) Consider the following 'guess-the-number' game in a PAC-learning setting: There is a $c^* \in [0, 1]$, and the learning algorithm is given a set of examples of the form $\{(x_i, y_i), i = 1, ..., n\}$ where $x_i \in [0, 1]$ and $y_i = 1$ if $x_i \leq c^*$ and 0 otherwise. Assume that $x_i$, $\forall i$ are picked from a uniform distribution over $[0, 1]$.

Consider a learning algorithm that picks a suitable subset of the data and employs a ML estimate, say $c_n$, from part (a). Show that the resulting algorithm is PAC, i.e., show that for any $\epsilon, \delta > 0$, $\exists N < \infty$ such that

$$\mathbb{P}\left(|c_n - c^*| > \epsilon\right) < \delta, \text{ for all } n > N.$$

3. **Principal component analysis**

(a) Consider the data points $x_1 = (-1, -1)$, $x_2 = (0, 0)$ and $x_3 = (1, 1)$.

Answer the following questions using this data: (1+1+0.5 marks)

(i) Project the data onto the one-dimensional space using PCA, and identify the approximations $\tilde{x}_1, \ldots, \tilde{x}_3$ corresponding to the given data points.

(ii) What is the variance of the projected data?

(iii) Recall that the approximation error $J = \sum_{i=1}^{3} \|x_i - \tilde{x}_i\|_2^2$. Calculate $J$ after PCA on the given dataset.

(b) Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be $n$ i.i.d. samples from the bivariate normal distribution $\mathcal{N}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 & -3 \\ -3 & 9 \end{bmatrix}\right)$, where $n$ is a large number.

Let $S_{a,b,c} = \{\mathbf{x} \in \mathbb{R}^2 : ax_1 + bx_2 + c = 0\}$ be a line in $\mathbb{R}^2$. The approximation error of this line is given by

$$R(a, b, c) = \sum_{i=1}^{n} \left[ \min_{\mathbf{y} \in S_{a,b,c}} \|\mathbf{x}_i - \mathbf{y}\|_2^2 \right].$$

Find the values for $a, b, c$ that minimize the error defined above. (2.5 marks)

4. **EM algorithm**

Consider a special case of a Gaussian mixture model in which the covariance matrices $\Sigma_k$ of the components are all constrained to have a common value $\Sigma$. Derive the EM equations for maximizing the likelihood function under such a model. (2 marks)

5. **Logistic regression**

Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\mathbf{w}$ whose decision boundary $\mathbf{w}^\intercal\mathbf{x} = 0$ separates the classes, and then taking the magnitude of $\mathbf{w}$ to infinity. (3 marks)

6. **Decision trees**

Consider the following two-class classification dataset:

| **Class 1** | |
|---|---|
| $x$ | $y$ |
| 1 | 5 |
| 1 | 8 |
| 4 | 1 |
| 2 | 3 |
| 3 | 7 |
| 5 | 4 |
| 5 | 7 |
| 7 | 7 |

| **Class 2** | |
|---|---|
| $x$ | $y$ |
| 1 | 2 |
| 2 | 1 |
| 3 | 2 |
| 6 | 4 |
| 6 | 7 |
| 7 | 1 |
| 8 | 4 |
| 8 | 2 |

Answer the following: (2+1 marks)

(a) Provide a decision tree that has zero training error on the dataset provided above. Note that the non-leaf nodes of the decision tree split data using the following format: "Is $x_i < c$?".

(b) Calculate the cross-entropy and Gini index for the decision tree from the part above.