Question Paper

Exam Date & Time: 19-May-2022 (10:00 AM - 01:00 PM)

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

SIXTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, MAY 2022

MACHINE LEARNING [ICT 4032]

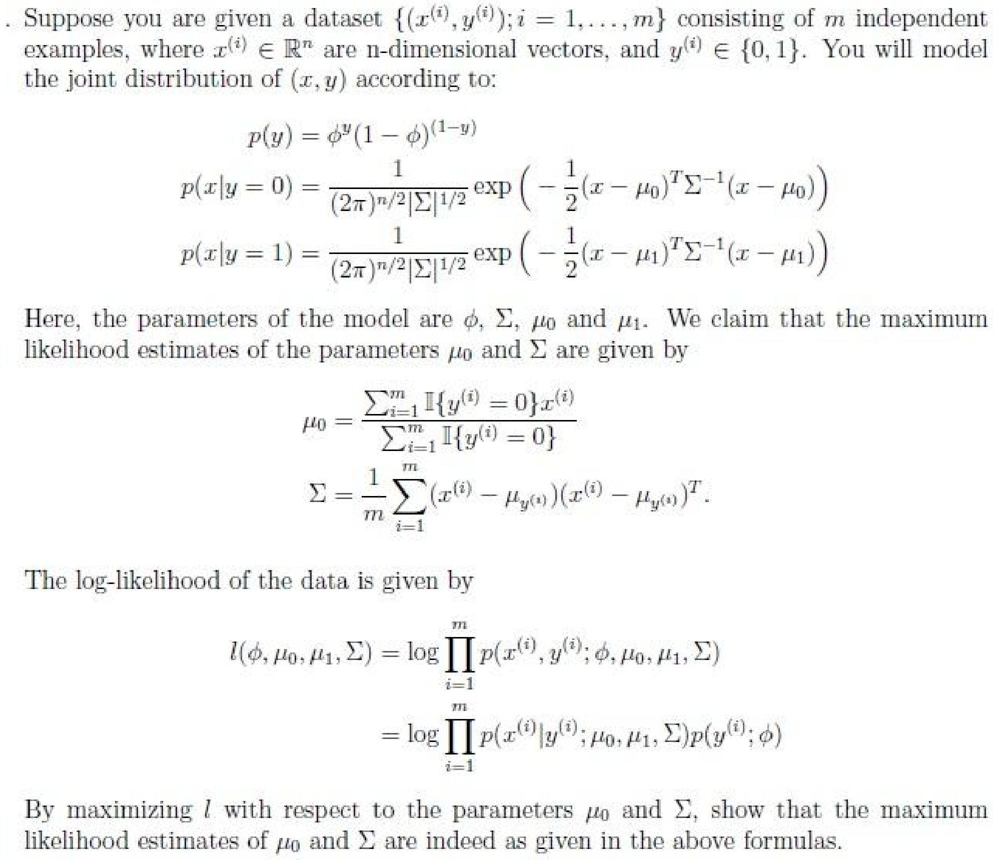
Marks: 50 **Duration: 180 mins.**

A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) A) (5)



. Suppose you are given a dataset $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ consisting of m independent examples, where $x^{(i)} \in \mathbb{R}^n$ are n-dimensional vectors, and $y^{(i)} \in \{0, 1\}$. You will model the joint distribution of (x, y) according to:

$$p(y) = \phi^y (1-\phi)^{(1-y)}$$
$$p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\Big)$$
$$p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\Big)$$

Here, the parameters of the model are ϕ, Σ, μ_0 and μ_1. We claim that the maximum likelihood estimates of the parameters μ_0 and Σ are given by

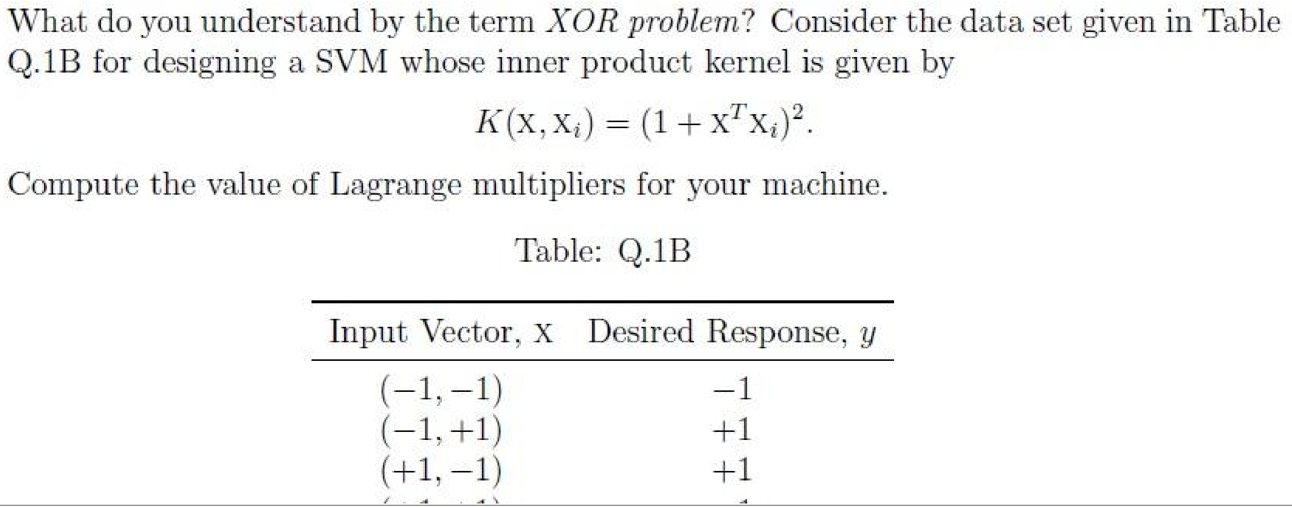
$$\mu_0 = \frac{\sum_{i=1}^m \mathbb{I}\{y^{(i)} = 0\} x^{(i)}}{\sum_{i=1}^m \mathbb{I}\{y^{(i)} = 0\}}$$
$$\Sigma = \frac{1}{m}\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

The log-likelihood of the data is given by

$$l(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$
$$= \log \prod_{i=1}^m p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi)$$

By maximizing l with respect to the parameters μ_0 and Σ, show that the maximum likelihood estimates of μ_0 and Σ are indeed as given in the above formulas.

B) (3)



What do you understand by the term *XOR problem*? Consider the data set given in Table Q.1B for designing a SVM whose inner product kernel is given by

$$K(\mathrm{x}, \mathrm{x}_i) = (1 + \mathrm{x}^T \mathrm{x}_i)^2.$$

Compute the value of Lagrange multipliers for your machine.

Table: Q.1B

Input Vector, x	Desired Response, y
$(-1,-1)$	-1
$(-1,+1)$	$+1$
$(+1,-1)$	$+1$

$(+1, +1)$	-1

C) Define a Markov Decision Process (MDP). (2)

2)

A) Consider Cocktail Party Problem (CPP), wherein sources are modeled by a random variable $s \in \mathbb{R}^n$, which is drawn according to some density $p_s(s)$. Now let another random variable be defined according to $x = As$, where $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. Here, matrix A is known as mixing matrix, and in order to find the sources we need to compute unmixing matrix $W = A^{-1}$, we can also write the observed variable as $x = W^{-1}s$. The density of observed variable x can be written as (5)

$$p(x) = \prod_{i=1}^{n} p_s(w_i^T x)|W|,$$

where $p(s) = g'(s)$ and g is a sigmodal function, which is defined as

$$g(s) = \frac{1}{1 + e^{-s}}.$$

The square matrix W is parameter in the model. Given a training set $\{x^{(i)}; i = 1, \ldots, m\}$ the likelihood function is given by

$$L(W) = \prod_{i=1}^{m} p(x^{(i)}).$$

Using maximum-likelihood estimate to derive the expression for W.

B) (3)

Consider a Markov model with given set of states $S = \{s_1, s_2, \ldots, s_{|S|}\}$, wherein we can choose a series over time $\vec{z} \in S^T$.

i) State two Markov assumptions that will allow you to tractably reason about time series.

ii) Derive a relation to compute the probability of a state sequence, $P(\vec{z})$.

iii) Assume that the transition matrix from a weather system is given by

$$A = \begin{array}{c|cccc} & s_0 & s_{sun} & s_{cloud} & s_{rain} \\ \hline s_0 & 0 & 0.4 & 0.5 & 0.1 \\ s_{sun} & 0 & 0.5 & 0.2 & 0.3 \\ s_{cloud} & 0 & 0.2 & 0.6 & 0.2 \\ s_{rain} & 0 & 0.1 & 0.7 & 0.2 \end{array}$$

Compute the probability for sequence of observation

$$\vec{z} = \{z_1 = s_{sun}, z_2 = s_{cloud}, z_3 = s_{cloud}, z_4 = s_{rain}, z_5 = s_{cloud}\}.$$

C) Consider the Poisson distribution parameterized by λ: (2)

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}.$$

Show that the Poisson distribution is in the exponential family, and clearly state what are $b(y), \eta, T(y)$ and $a(\eta)$.

3) (5)

A)

The EM algorithm is given by

Repeat until convergence{

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta)$$

(M-step) Set

$$\theta := \underset{\theta}{argmax} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

}

Now consider the Gaussian mixture model with $p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$, where $x^{(i)}|z^{(i)} \sim \mathcal{N}(\mu_j, \Sigma_j)$ and $p(z^{(i)} = j) = \phi_j$. Apply EM algorithm to fit the parameters ϕ and μ for the Gaussian mixture model.

B) Consider a learning problem in which you have a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of k hypothesis. Show that if uniform convergence occur, the generalization error of $\hat{h}$ is at most 2γ worse than the best possible hypothesis in $\mathcal{H}$. (3)

C) Given a training set $\{(x^{(i)}, y^{(i)})|i = 1, \ldots, m\}$, where $x^{(i)} \in \mathbb{R}^{n_i}$, and n_i is the number of words in the i-th training example. The likelihood of the data is given by (2)

$$\mathcal{L}(\phi, \phi_{k|y=0}, \phi_{k|y=1}) = \prod_{i=1}^{m} \left(\prod_{j=1}^{m} p(x_j^{(i)}|y;\phi_{k|y=0}, \phi_{k|y=1}) \right) p(y^{(i)};\phi_y).$$

Maximizing $\mathcal{L}(\phi, \phi_{k|y=0}, \phi_{k|y=1})$ yields the maximum likelihood estimates of the parameters:

$$\phi_{k|y=0} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}n_i}$$

$$\phi_{k|y=1} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}n_i}$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}.$$

Apply Laplace smoothing to the above parameters and re-write those parameters.

4) (5)

A) Consider a classification problem in which the response variable y can take on any one of the k values, so $y \in \{1, 2, \ldots, k\}$. Derive a Generalized Linear Model (GLM) for modeling this type of multinomial data.

B) Consider a binary classification problem with $y \in \{0, 1\}$. Is it advisable to use classical linear regression for this problem? Given a logistic regression model, derive least square regression using maximum likelihood estimate under the following set of probabilistic assumptions: (3)

$$p(y = 1|x; \theta) = h_\theta(x)$$
$$p(y = 0|x; \theta) = 1 - h_\theta(x).$$

Here θ and h_θ have their usual meaning.

C) Given an unlabeled set of examples $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ certain SVM algorithm tries to find a direction w that maximally separates the data from the origin. Precisely, it solves the primal optimization problem: (2)

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}||w||^2 \\ &\text{subject to} \\ &w^T x^{(i)} \geq 1 \quad i = 1, \ldots, m \end{aligned}$$

A query example x is labeled 1 if $w^T x \geq 1$, and 0 otherwise. Derive the corresponding dual optimization problem. Note that dual optimization problem should be free from w.

5) A) Suppose, there are a finite set of models $\mathcal{M} = \{M_1, \ldots, M_d\}$, and you are trying to select one among them, which describes the behavior of your data. How will you select your model so that the empirical error is minimal? Describe various techniques for model selection. (5)

B) Show that PCA is a variance maximizing problem. (3)

C) Starting with basic definition of a convex function derive the relation for Jensen's inequality. (2)

-----End-----