Question Paper

Exam Date & Time: 30-May-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SIXTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, MAY 2023

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) Consider a linear regression problem in which we want to *weight* different training examples differently. Specifically, suppose we want to minimize (5)

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}\left(\theta^T x^{(i)} - y^{(i)}\right)^2.$$

i) The given cost function in vectorial notation can be written as

$$J(\theta) = (X\theta - \vec{y})^T W (X\theta - \vec{y}).$$

Solve $\underset{\theta}{\text{argmin}}\ J(\theta)$ and write the results in closed form as a function of X, W and $\vec{y}$.

ii) Suppose we have a training set $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ of m independent examples, but $y^{(i)}$'s were observed with differing variances. Suppose that

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}}\exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2(\sigma^{(i)})^2}\right).$$

Show that finding the maximum likelihood estimate of θ reduces to solving a weighted linear regression problem. Clearly state what the $w^{(i)}$'s are in terms of the $\sigma^{(i)}$. [Hint: Solve $\underset{\theta}{\text{argmax}} \log \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)$ as much as possible.]

B) (3)

Assume that the target variable and the inputs are related via $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$, where $\epsilon^{(i)}$ is an error term that captures either unmodeled effects or random noise. Further, assume that $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$, and the density of $\epsilon^{(i)}$ is given by

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\epsilon^{(i)})^2}{2\sigma^2}}.$$

Using these probabilitic assumption on the data show that the least-square regression corresponds to finding the maximum likelihood estimate of θ.

C) A generalized linear model assume that the response variable y is distributed according to a member of the exponential family: (2)

$$p(y;\eta) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right).$$

Show that the Bernoulli distribution, $p(y, \phi) = \phi^y(1-\phi)^{1-y}$ is an example of exponential distribution.

2) A) In a factor analysis model, assume a joint distribution on (x, z) as follows (5)

$$z \sim \mathcal{N}(0, I)$$
$$x|z \sim \mathcal{N}(\mu + \Lambda z, \Psi)$$

where $\mu \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{n\times k}$, and the diagonal matrix $\Psi \in \mathbb{R}^{n\times n}$, $(k < n)$. Equivalently factor analysis model can also be defined according to

$$z \sim \mathcal{N}(0, I)$$
$$\epsilon \sim \mathcal{N}(0, \Psi)$$
$$x = \mu + \Lambda z + \epsilon$$

Also we have

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right).$$

Consider a training set $\{x^{(i)}; i = 1, \ldots, m\}$, the log-likelihood of the parameter is given by

$$l(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2}|\Lambda\Lambda^T + \Psi|^{1/2}} exp\left(-\frac{1}{2}(x^{(i)} - \mu)^T(\Lambda\Lambda^T + \Psi)^{-1}(x^{(i)} - \mu)\right).$$

Apply EM algorithm to estimate Λ.

B) Describe following methods for feature selection: (3)

i) Forward search

ii) Backward search

iii) Filter method.

C) Given γ and some $\delta > 0$, how large must m be before you can guarantee that with probability at least $1 - \delta$, training error will be within γ of generalization error? Assume $\delta = 2k\exp(-2\gamma^2 m)$. (2)

3) (5)

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

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

A new test example x is labeled 1 if $w^T x \geq 1$, and 0 otherwise. For the given primal optimization problem, write down the corresponding dual optimization problem. Simplify your answer as much as possible.

B) Show that for Principal Component Analysis (PCA), maximizing variance corresponds to finding the eigen vectors of the covariance matrix. (3)

C) The Markov Decision Process (MDP) provides the formalism in which the reinforcement problems are posed. Define MDP. (2)

4) (5)

A)

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) Consider a coin-flipping experiment in which you are given a pair of coins A and B of unknown biases θ_A and θ_B respectively (i.e., on any given flip, coin A will land on heads with probability θ_A and on tail with probability $(1-\theta_A)$, similarly for coin B). Consider the dataset collected using following procedure five times: labels of the coins are removed, now randomly choose one of the two coin and perform ten independent coin tosses with the selected coin. Let $x^{(i)} = j$ denotes j number of heads obtained during i-th set of experiment. The dataset obtained from this experiment are $\{x^{(1)} = 5, x^{(2)} = 9, x^{(3)} = 8, x^{(4)} = 4, x^{(5)} = 7, \}$. With initial estimate of biases $\hat{\theta}_A^{(0)} = 0.6$ and $\hat{\theta}_B^{(0)} = 0.5$, apply EM algorithm to compute $(\hat{\theta}_A^{(1)}, \hat{\theta}_B^{(1)})$. (3)

C) Show that a valid kernel matrix, K must be positive semi-definite. (2)

5)

A) Consider a classification problem in which the response variable y can take any one of k values, so $y \in \{1, 2, \ldots, k\}$. Assume that the response variable is discrete. Model this classification scenario as distributed according to a multinomial distribution, and derive the result for hypothesis function, $h_\theta(x)$ using GLM approach. (5)

B) (3)

The log-likelihood of a Markov model is defined as

$$l(A) = \log P(\vec{z}; A)$$
$$= \log \prod_{t=1}^T A_{z_{t-1}z_t}$$

where $\vec{z}$ is an observed sequence. Find the maximum likelihood estimate for the parameter, A.

C) What do you understand by the term *Gaussian Mixture Model (GMM)*? Give an example of GMM. (2)

-----End-----