Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

SEVENTH SEMESTER B.TECH (INFORMATION TECHNOLOGY / COMPUTER AND COMMUNICATION ENGINEERING) DEGREE END SEMESTER EXAMINATION-NOVEMBER 2018
SUBJECT:PROGRAM ELECTIVE-V MACHINE LEARNING (ICT 4007)
(REVISED CREDIT SYSTEM)

TIME: 3 HOURS **29/11/2018** **MAX. MARKS: 50**

Instructions to candidates

- Answer **ALL** questions. All questions carry equal marks.
- Missing data if any, may be suitably assumed.

1A. Consider a linear regression problem in which we want to weight different training examples differently. Specifically, suppose you want to minimize

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

i) Show that $J(\theta)$ can be written as

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

for an appropriate diagonal matrix W, and where X and $\vec{y}$ are as defined in the class.

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

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

Show that finding the maximum likelihood estimate of θ reduces to solving a weighted linear regression problem. [5]

1B. Consider the geometric distribution, which is parametrized by ϕ given by

$$p(y;\phi) = (1-\phi)^{y-1}\phi.$$

Show that the geometric distribution is an exponential family distribution. Explicitly specify $b(y), \eta, T(y)$, and $a(\eta)$. Also write ϕ in terms of η. [3]

1C. Write the algorithm for value iteration and policy iteration for finite state MDP. [2]

2A. 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. [5]

2B. Consider a classification or regression problem where we would like to predict the value of some random variable y as a function of x.

i) Write the general assumptions made to derive GLM for such problems.

ii) Using GLM construction, show that ordinary least square is a special case of GLM family of models. [3]

2C. Assume that the input feature x_j, $j = 1, \ldots, n$ are discrete binary-valued variables such that $x_j \in \{0, 1\}$ and $x = [x_1 x_2, \ldots, x_n]$. For each training example $x^{(i)}$, assume that the output target variable $y^{(i)} \in \{0, 1\}$. Now, consider the Naive Bayes model, given the above context. This model can be parametrized by $\phi_{j|y=0} = p(x_j = 1|y = 0)$, $\phi_{j|y=1} = p(x_j = 1|y = 1)$, and $\phi = p(y = 1)$. Write the expression for $p(y = 1|x)$ in terms of $\phi_{j|y=0}, \phi_{j|y=1}$, and ϕ_y. [2]

3A. Explain the concept of functional and geometric margin in reference to Support Vector Machine (SVM). Pose an optimization problem in terms of geometric margin such that its solution gives the optimal margin classifier. [5]

3B. In class we have seen how SVM can be used for classification. In this problem, we will develop a modified algorithm, called the Support Vector Regression algorithm, which can be used for regression with continuous valued labels $y \in \mathbb{R}$.
Suppose we are given a training set $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, where $x^{(i)} \in \mathbb{R}^{n+1}$ and $y \in \mathbb{R}$. We would like to find a hypothesis of the form $h_{w,b}(x) = w^T x + b$ with a small value of w. Our optimization problem is

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)} - w^T x^{(i)} - b \leq \epsilon, \; i = 1, \ldots, m \\ & w^T x^{(i)} + b - y^{(i)} \leq \epsilon, \; i = 1, \ldots, m \end{aligned}$$

where $\epsilon > 0$ is a given, fixed value.

i) Write the Lagrangian for the given optimization problem. Use two sets of Lagrange multipliers α_i and β_i, corresponding to the two inequality constraints, so that the Lagrangian would be written as $\mathcal{L}(w, b, \alpha, \beta)$.

ii) Derive the dual optimization problem. [3]

3C. Suppose you are given a hypothesis $h_0 \in \mathcal{H}$, and your goal is to determine whether h_0 has generalization error within $\eta > 0$ of the best hypothesis, $h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$. Specifically, we say that a hypothesis h is η-optimal if $\varepsilon(h) \leq \varepsilon(h^*) + \eta$. Here, we wish to answer the question: Given a hypothesis h_0, is h_0 η-optimal?
Let $\delta > 0$ be some fixed constant, and consider a finite hypothesis class $\mathcal{H}$ of size $|\mathcal{H}| = k$. For each $h \in \mathcal{H}$, let $\hat{\varepsilon}(h)$ denote the training error of h with respect to some training set of m IID examples, and let $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$ denote the hypothesis that minimizes training error. Now, consider the following algorithm

1. Set $\gamma := \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$.
2. If $\hat{\varepsilon} > \hat{\varepsilon}(\hat{h}) + \eta + 2\gamma$, then return NO.

3. If $\hat{\varepsilon} < \hat{\varepsilon}(\hat{h}) + \eta - 2\gamma$, then return YES.

Show that if $\varepsilon(h_0) \leq \varepsilon(h^*) + \eta$, then the probability that the algorithm returns NO is at most δ.

[2]

4A. Describe the following cross-validation techniques

i) Hold-out cross validation

ii) K-fold cross validation, and

iii) Leave-one out cross validation.

Also, for each technique write its suitability condition. [5]

4B. Consider a factor analysis model defined according to

$$z \sim \mathcal{N}(0, I)$$

$$\epsilon \sim \mathcal{N}(0, \Psi)$$

$$x = \mu + \Lambda z + \epsilon$$

where $z \in \mathbb{R}^k$ is a latent random variable, $\mu \in \mathbb{R}^n$, the matrix $\Lambda \in \mathbb{R}^{n \times k}$, the diagonal matrix $\Psi \in \mathbb{R}^{n \times n}$, and ϵ and z are independent. The random variable z and x have a joint Gaussian distribution

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$$

Compute μ_{zx}, and Σ.

[3]

4C. 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, suppose $\theta^{(t)}$ and $\theta^{(t)}$ are the parameters from two successive iterations of EM. Prove that $l(\theta^{(t)}) \leq l(\theta^{(t+1)})$, which shows EM always monotonically improves the log-likelihood.

[2]

5A. Consider a n dimensional feature vector $x \in \mathbb{R}^n$. Derive the required relation to find the top k principal components, and express x in terms of those principal components. [5]

5B. Describe different types of ambiguities associated with ICA. [3]

5C. Define a Markov Decision Process (MDP). [2]