Reg. No. | | | | | | | | | | |

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

SEVENTH SEMESTER B.TECH (INFORMATION TECHNOLOGY / COMPUTER AND COMMUNICATION ENGINEERING) DEGREE END SEMESTER EXAMINATION-NOVEMBER 2017

SUBJECT:PROGRAM ELECTIVE-V MACHINE LEARNING (ICT 4007)

(REVISED CREDIT SYSTEM)

TIME: 3 HOURS 25/11/2017 MAX. MARKS: 50

Instructions to candidates

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

1A. Explain the following terminologies in reference to Machine Learning:

i) Examples
ii) Labels
iii) Training sample
iv) Validation sample
v) Test sample
vi) Loss function
vii) Hypothesis set

[5]

1B. 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. [3]

1C. Consider the univariate Gaussian distribution parameterized by μ, i.e $y \sim \mathcal{N}(\mu, 1)$. Show that the univariate Gaussian distribution is in exponential family, and clearly state what are $b(y), \eta, T(y)$, and $a(\eta)$. [2]

2A. Given a dataset $\{(x^{(i)}, y^{(i)}; i = 1, \ldots, m)\}$ consisting of m independent examples, where $x^{(i)} \in \mathbb{R}^n$, and $y^{(i)} \in \{0, 1\}$. 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. The log-likelihood of the data is given by

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

Using MLE find the relation for ϕ, and μ_0.

[5]

2B. Consider the data set given in Table Q.2B, 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 optimum value of the dual objective function.

Table: Q.2B

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

[3]

2C. The Gaussian kernel is given by the function

$$K(x, z) = e^{-\frac{\|x-z\|^2}{\sigma^2}},$$

where $\sigma^2 > 0$ is some fixed positive constant. Prove that the Gaussian kernel is indeed a valid kernel. [Hint: $\|x - z\|^2 = \|x\|^2 - 2x^T z + \|z\|^2$.] [2]

3A. Describe various techniques for feature selection. [5]

3B. Consider a binary classification problem with labels $y \in \{0, 1\}$, and let $\mathcal{D}$ be a distribution over (x, y). Let $\mathcal{H} = \{h_1, \ldots, h_k\}$ be a finite hypothesis class, and suppose our training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ is obtained by drawing m examples IID from $\mathcal{D}$. Suppose we pick $h \in \mathcal{H}$ using empirical risk minimization: $\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{\varepsilon}(h)$. Also let $h^* = \arg\min_{h\in\mathcal{H}} \varepsilon(h)$. Let any $\delta, \gamma > 0$ be given. Show that for $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$ to hold with probability $1 - \delta$, it suffice that $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$. [3]

3C. What do you understand by the term *online learning*? How is it different from *batch learning*? [2]

4A. In a factor analysis model, assume a joint distribution on (x, z) as follows

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

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.

[5]

4B. 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 $(\hat{\theta}_A^{(2)}, \hat{\theta}_B^{(2)})$.

[3]

4C. Briefly discuss various types of inherent ambiguities associated with Independent Component Analysis (ICA).

[2]

5A. 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

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

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 derive the expression for W. [5]

5B. Consider a generic convex optimization problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \ i = 1, \ldots, m \\ & h_i(x) = 0, \ i = 1, \ldots, p \end{aligned}$$

where f, g_i are convex function, and h_i are affine functions, and x is optimizable variable. Write the primal and dual problem for the given constrain optimization problem. [3]

5C. Why do you need to pre-process the data before applying Principal Component Analysis? List those pre-processing steps. [2]