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

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

SEVENTH SEMESTER B.TECH (INFORMATION TECHNOLOGY / COMPUTER AND COMMUNICATION ENGINEERING) DEGREE MAKE UP EXAMINATION-DECEMBER 2017

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

(REVISED CREDIT SYSTEM)

TIME: 3 HOURS 30/12/2017 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. Multi-variate Bernoulli and Multinomial event models are the popular model for text classification. Consider x_i as the identity of the i-th word in the text (here, email), and $x_i \in \{1, \ldots, |V|\}$, where $|V|$ is the size of vocabulary. Overall probability of the model is given by

$$p(y)\prod_{i=1}^{n} p(x_i|y).$$

Model is parameterized by:

$$\phi_y = p(y),\ \phi_{k|y=1} = p(x_j = k|y = 1)(\text{for any } j) \text{ and } \phi_{k|y=0} = p(x_j = k|y = 0).$$

You are given a training set $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, where $x^{(i)} = \{x_1^{(i)}, x_2^{(i)} \ldots, x_{n_i}^{(i)}\}$, n_i is the number of words in the ith training example. The likelihood of the data is given by

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

Use MLE to find the expression for model parameters. [5]

1B. 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:

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

Here θ and h_θ have their usual meaning. [3]

1C. Consider the Poisson distribution parameterized by λ:

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

Show that the Poisson 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:

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

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 Σ. [5]

2B. What do you understand by the term *XOR problem*? 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 value of Lagrange multipliers for your machine.

Table: Q.2B

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

[3]

2C. Consider a valid kernel $K(x, z) = (x^T z + c)^2$. Assuming that $x, z \in \mathbb{R}^3$, write the feature mapping vector $\Phi(z)$ for the given kernel function. [2]

3A. Explain various techniques of cross-validation. [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 drwaing 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 $\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$. Let any $\delta, \gamma > 0$ be given. Show that to hold with probability $1 - \delta$, it suffice that $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$. [3]

3C. Write K-means clustering algorithm. [2]

4A. Suppose you have an estimation problem in which you are given a training set $\{x^{(i)}; i = 1, \ldots, m\}$ consisting of m independent examples. It is required to fit the parameters of a model $p(x, z)$, where z is a latent variable. The likelihood is given by

$$l(\theta) = \sum_{i=1}^{m} \log \sum_{z} p(x, z; \theta).$$

Explicit finidng of maximum likelihood estimate of the parameter θ may be difficult. With all the necessary steps, show how Expecation-Maximization algorithm can be applied for the given problem. [5]

4B. Show that the Principle Component Analysis (PCA) is basically a variance maximization algorithm.

[3]

4C. Consider the scenario of Cocktail Party Problem. What kind of ambiguity arise, if you assume that the sources are Gaussian distributed, and then if you apply Independent Component Analysis (ICA) to recover the individual source from the observed data?

[2]

5A. Starting with the following optimization problem for SVM

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

Solve for w and b using primal and dual problem approach. [5]

5B. Radio Mirchi partitions radio station listeners into two groups-the 'young' and 'old'. They assume that, given the knowledge that a customer is either 'young' or 'old', is sufficient to determine whether or not a customer will like a particular radio station, independent of their likes or dislikes for any other stations. Given that a customer is young, she has 95% chance to like Radio1, a 5% chance to like Radio2, a 2% chance to like Radio3 and a 20% chance to like Radio4. Similarly, an old listener has a 3% chance to like Radio1, an 82% chance to like Radio2, a 34% chance to like Radio3 and a 92% chance to like Radio4. It is known that 90% of the listeners are old.

Given this model, and the fact that a new customer likes Radio1 and Radio3, but dislikes Radio2 and Radio4, what is the probability that the new customer is young? [3]

5C. What do you understand by the term *Gaussian Mixture Model*(GMM)? Give an example of GMM.

[2]