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

MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL 576104

(Constituent College of Manipal University)

SIXTH SEMESTER B.TECH DEGREE MAKE UP EXAMINATION-JULY 2016
SUBJECT:OPEN ELECIVE-II MACHINE LEARNING (ICT 364)
(REVISED CREDIT SYSTEM)

TIME: 3 HOURS **-/07/2016** **MAX. MARKS: 50**

Instructions to candidates

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

1A. With probabilistic assumption and interpretation, derive the required relation for least-square regression.

1B. What are the limitations of linear regression? How does locally weighted linear regression overcome those limitations?

1C. A generalized linear model assumes that the response variable y (conditioned on x) is distributed according to a member of the exponential family:

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

Show that the Gaussian distribution is an example of exponential distribution.

[5+3+2]

2A. 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:

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

Suppose you have already fit ϕ, μ_0, μ_1, and Σ, and now want to make a prediction at some new query point x. Show that the posterior distribution of the label at x takes the form of a logistic function, and can be written as

$$p(y = 1|x; \phi, \Sigma, \mu_0, \mu_1) = \frac{1}{1 + \exp(-\theta^T x)},$$

where θ is some appropriate function of $\phi, \Sigma, \mu_0, \mu_1$.

2B. What is the intuition behind *margins* in SVMs? Explain the following margins:

i) Functional margin, and

ii) Geometric margin.

2C. State *Hoeffding inequality* and give its interpretation.

[5+3+2]

3A. 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.

3B. Consider that you have an estimation problem in which you have a training set $\{x^{(1)}, \ldots, x^{(m)}\}$ consisting of m independent variables. You wish to fit the parameters of a model $p(x, z)$ to data, where the likelihood is given by

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

Due to latent variable z, explicit finding of maximum likelihood estimate of the parameter θ may be hard. You use EM algorithm, which is as follows

```
Repeat until convergence
{
(E-step) For each $i$, Set
        $Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta)$
(M-step) Set
        $\theta := arg\ \max_{\theta} \sum_{i} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}.$
}
```

Show that this algorithm converges.

3C. Define the following

i) Empirical error

ii) Empirical risk minimization.

[5+3+2]

4A. When we have data $x^{(i)} \in \mathbb{R}^n$ that comes from a mixture of several Gaussians, the EM algorithm can be applied to fit a mixture model. In this setting, we have sufficient data to be able to discern the multiple-Gaussian structure in the data. Now, consider a setting in which $n \gg m$. In such a problem, it might be difficult to model the data even with a single Gaussian, much less a mixture of Gaussian. For such a scenario, what are the problems and preliminary solutions? How the limitations of preliminary solutions can be overcome?

4B. Consider a unsupervised learning problem, where you are given a training set $\{x^{(1)}, \ldots, x^{(m)}\}$. You wish to model the data by specifying a joint distribution $p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$, where $z^{(i)} \sim$ multinomial(ϕ) ($\phi_j \geq 0, \sum_{j=1}^{k} \phi_j = 1$ and the parameter ϕ_j gives $p(z^{(i)} = j)$), and $x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$. Assume that k denotes the number of values that the $z^{(i)}$'s can take on. The parameters of the model are ϕ, μ, and Σ. To estimate them, you can write the likelihood for your data as

$$l(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma)p(z^{(i)}; \phi).$$

Use EM algorithm to derive the expression for μ.

4C. Write k-means clustering algorithm.

[5+3+2]

5A. Consider a learning problem in which you have a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of k hypothesis. Derive *uniform convergence* result.

5B. State Jensen's inequality and graphically depict its behavior.

5C. State Mercer's theorem on valid kernels.

[5+3+2]

6A. For the data set given in Table Q.6A, design a polynomial learning machines whose inner product kernel is given by

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

Table: Q.6A

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

6B. Consider a classification problem in which the response variable y can take on any one of k values, so $y \in \{1, 2, \ldots, k\}$. You can model such distribution as a multinomial distribution. Use GLM for modeling this multinomial distribution and derive the relation for parameters of exponential family distribution.

6C. Briefly explain *Fisher scoring* for maximizing $l(\theta)$.

[5+3+2]