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

# MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL 576104

(Constituent College of Manipal University)

**SIXTH SEMESTER B.TECH DEGREE END SEMESTER EXAMINATION-MAY 2016**
**SUBJECT:OPEN ELECIVE-II MACHINE LEARNING (ICT 364)**
**(REVISED CREDIT SYSTEM)**

**TIME: 3 HOURS** **16/05/2016** **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. What do you understand by the term *hypothesis function*? Derive normal equation for parameter $\theta$ as per the LMS algorithm.

1B. For logistic regression, derive the relation for parameter updation.

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

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

Show that the Bernoulli 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:

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

Here, the parameters of the model are $\phi$, $\Sigma$, $\mu_0$ and $\mu_1$. The log-likelihood of the data is given by

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

By maximizing $l$ with respect to the four parameters, derive the relation for $\phi, \mu_0, \mu_1$, and $\Sigma$.

2B. Frame the optimal margin classifier as an optimization problem.

2C. Bias and variance are the twin evils of machine learning. With appropriate diagrams explain the bias-variance trade off and behavior of the model.

[5+3+2]

3A. Suppose, you have a supervised learning problem where the number of features $n$ is very large ($n \gg m$), but you suspect that there is only a small number of features that are "relevant" to the learning task. Explain various techniques for feature selection.

3B. Suppose, 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 find the parameters of a model $p(x, z)$ to the data, where the likelihood is given by

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

But the explicit finding the maximum likelihood estimates of parameter $\theta$ may be hard. Also, here $z^{(i)}$'s are latent variable. For such a setting, the EM algorithm gives an efficient method for maximum likelihood estimation. Establish preliminary relation required for applying EM algorithm as per the Jensen's inequality.

3C. Given $\gamma$ 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 $\gamma$ of generalization error? Assume $\delta = 2k\exp(-2\gamma^2 m)$.

[5+3+2]

4A. Marginal distributions of Gaussians are themselves Gaussians, and as per the definition of the multivariate Gaussian distribution, it is known that $x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$, where

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma^{-1}\Sigma_{21} \end{aligned}$$

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

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

where $\mu \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{n \times k}$, and the diagonal matrix $\Psi \in \mathbb{R}^{n \times n}$, $(k < n)$. Workout the expression for the log likelihood of the parameters $l(\mu, \Lambda, \Psi)$.

4B. Let a sequence of examples $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$ be given. Suppose that $||x^{(i)}|| \leq D$ for all $i$, and further that there exists a unit-length vector $u$ such that $y^{(i)}.(u^T x^{(i)}) \geq \gamma$ for all examples in the sequence. Show that the total number of mistakes that the perceptron algorithm makes on this sequence is at most $(D/\gamma)^2$.

4C. What do you understand by the term *mixture of Gaussians*?

[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. Show that if uniform convergence occur, the generalization error of $\hat{h}$ is at most $2\gamma$ worse than the best possible hypothesis in $\mathcal{H}$.

5B. What do you mean by a convex function? Why is it so important in optimization theory?

5C. The following questions require a true/false or a short answer.

i) Let there be a binary classification problem with continuous-valued features. What will the decision boundary look like if we model the two classes using separate covariance matrices $\Sigma_0$ and $\Sigma_1$?

ii) Let any $x^{(1)}, x^{(2)}, x^{(3)} \in \mathbb{R}^p$ be given ($x^{(1)} \neq x^{(2)}, x^{(1)} \neq x^{(3)}, x^{(2)} \neq x^{(3)}$). Also let any $z^{(1)}, z^{(2)}, z^{(3)} \in \mathbb{R}^q$ be fixed. Then there exists a valid Mercer kernel $K : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ such that for all $i, j \in \{1, 2, 3\}$ we have $K(x^{(i)}, x^{(j)}) = (z^{(i)})^T z^{(j)}$. True or False?

[5+3+2]

6A. 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:

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

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.

6B. Describe the method of constructing GLMs.

6C. Suppose $x, z \in \mathbb{R}^n$, and consider $K(x, z) = (x^T z)^2$. You know that $K(x, z) = \phi(x)^T \phi(z)$. Write feature map $\phi(x)$ for the given kernel. Here assume that $n = 3$.

[5+3+2]