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

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL

A Constituent Institution of Manipal University

SIXTH SEMESTER B.TECH DEGREE END SEMESTER EXAMINATION-APRIL/MAY 2017
SUBJECT:OPEN ELECTIVE-II MACHINE LEARNING (ICT 364)
(REVISED CREDIT SYSTEM)

TIME: 3 HOURS 03/05/2017 MAX. MARKS: 50

Instructions to candidates

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

1A. For the data set given in Table Q.1A, design a polynomial learning machines whose inner product kernel is given by

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

Table: Q.1A

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

[5]

1B. Consider a unsupervised learning problem, where you are given a training set $\{x^{(1)}, \ldots, x^{(m)}\}$. Model the data with 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, the likelihood for the data is given as

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

Derive the expression for μ using EM algorithm.

[3]

1C. Write steps involved in the method of k-fold cross validation.

[2]

2A. Describe locally weighted linear regression, and give its comparison with linear regression.

[5]

2B. State the three assumptions made while constructing a Generalized Linear Model (GLM).

[3]

2C. Consider a classification problem in which the response variable y can take on one of k values, so $y \in \{1, 2, \ldots, k\}$. Which regression technique will solve this classification problem? Write the model for such technique. [2]

3A. 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 $y^{(i)} \in \{0, 1\}$. Model the joint distribution of (x, y) according to:

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

Assume that 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 $\phi, \Sigma, \mu_0, \mu_1$. [5]

3B. Explain LMS algorithm. [3]

3C. A generalized linear model assumes that the response variable y (conditioned on x) is distributed according to a member of the exponential family, that is

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

Show that the Bernoulli distribution, which is defined as

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

is an example of exponential distribution. [2]

4A. 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 $y^{(i)} \in \{0, 1\}$. 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 our model are ϕ, Σ, μ_0 and μ_1. Assume n (the dimension of x) is 1, so that $\Sigma = [\sigma^2]$ is just a real number, and likewise the determinant of Σ is given by

$|\Sigma| = \sigma^2$. Given the dataset, it is claimed that the maximum likelihood estimates of the parameters are given by

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\}$$
$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$
$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$
$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

The log-likelihood of the data is

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

By maximizing l with respect to any two parameters, prove that the maximum likelihood of ϕ, μ_0, μ_1, and Σ are indeed as given in the formula above. [5]

4B. Compare and contrast between the Gaussian Discriminant Analysis (GDA) model and logistic regression. [3]

4C. Explain the following:

i) Discriminative learning algorithm
ii) Generative learning algorithm. [2]

5A. With all the necessary mathematical formalism, explain the working of Naive Bayes classifier. [5]

5B. How will Gaussian distribution behave under the following conditions for a standard normal distribution?

i) covariance matrix (Σ) is scaled down
ii) covariance matrix (Σ) is scaled up
iii) off-diagonal entry in Σ is increased
iv) off-diagonal entry in Σ is decreased. [3]

5C. In any optimization problem, we prefer objective function to be convex. What is the advantage of having a convex function as an objective function? [2]

6A 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. [5]

6B State lemma for Hoeffding inequality (also known as Chernoff bound), and write its interpretation. [3]

6C State Mercer's theorem. [2]