Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

VII SEMESTER B.TECH. (INFORMATION TECHNOLOGY / COMPUTER AND COMMUNICATION ENGINEERING)

MAKE UP EXAMINATION, DECEMBER 2018

SUBJECT: PROGRAM ELECTIVE V MACHINE LEARNING [ICT 4007]

REVISED CREDIT SYSTEM

(5/12/2018)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data if any may be suitably assumed.

1A. Given the training set of size $m \times n$ and the output vector y, derive the value of the parameter θ that minimizes $J(\theta)$ for linear regression. Compare the two ways used to obtain θ. **5**

1B. The assumptions made in Naïve Bayes help in reducing the complexity of the problem. Explain and obtain the probability distribution of predicting if the new mail received is spam or not. **3**

1C. How does Newton's method help in maximizing the value of $\ell(\theta)$ **2**

2A. Given the primal optimization problem for finding the optimal margin classifier as **5**

$$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m$$

Get the expression for w (width) and b(bias) using the optimization problem.

2B. Write an algorithm for forward search during feature selection. Compare it with backward search algorithm. **3**

2C. Distinguish between batch learning and online learning. Specify the predictions and update rules that are used in the perceptron algorithm. **2**

3A. Given the training error $\mathcal{E}(\hat{h}) = \sum_{j=1}^{m} Zj$, $\hat{h}$ hypothesis with smallest training error, m the number of training examples, Z a Bernoulli random variable with sample distribution (x, y) ~ D and Z = 1{hi(x) ≠ y}, obtain the expression $\mathcal{E}(\hat{h}) \leq (\min_{h \in \mathcal{H}} \mathcal{E}(h)) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$ in the case of finite hypothesis class $\mathcal{H}$ with k hypotheses. **5**

3B. Given that x represents a training example of size m and z the hidden variable having different (mixture) Gaussians distribution from which x is obtained. The joint distribution of $p(x^{(i)}, z^{(i)})$ with log likelihood is given as **3**

$\ell(\emptyset, \mu, \Sigma) = \sum_{i=1}^{m} log\, p\left(x^{(i)} \middle| z^{(i)};\, \mu, \Sigma\right) + \log p(z^{(i)}; \emptyset)$. Write the EM algorithm to compute the parameters and compare it with Gaussian where the values of $z^{(i)}$ are known.

3C. Does the k-means algorithm guarantee to converge? Justify. **2**

4A. In order to fit the parameters of the model p(x, z) to the data, the log likelihood is given by $\ell(\theta) = \sum_{i=1}^{m} \log p(x; \theta)$. With z as a latent variable and x as the training set with m training examples, construct a lower-bound on ℓ, and then optimize that lower-bound . Write the steps involved in the EM algorithm to find the maximum likelihood estimate of the parameter θ. **5**

4B. Map the parameters b(y), η, T(y) and a(η) of the generalized expression of the exponential family with that of i) Bernoulli's ii) Gaussian iii) Poisson with distribution $p(y|\lambda) = \frac{1}{y!} exp(ylog\lambda - \lambda)$ **3**

4C. Given the unit vector u and $x^{(i)}$ the points in the dataset, how do you maximize the variance of the projection of $x^{(i)}$ on to u. Obtain the relation for the principal eigen vector of the covariance matrix. Specify why is it a dimensionality reduction algorithm. **2**

5A. Obtain an expression for optimal value function and optimal policy given the tuple (S,A,{Psa},γ,R) as defined in the Markov decision process. **5**

5B. Given the joint distribution of (x, z) where z is the latent random variable such that $z \sim N(0, I)$, $\varepsilon \sim N(0, \psi)$, $x = \mu + \Lambda z + \varepsilon$ where the vector $\mu \in \mathbb{R}^n$, the matrix $\Lambda \in \mathbb{R}^{nxk}$, the diagonal matrix $\psi \in \mathbb{R}^{nxn}$, ε and z are independent and k<n. Obtain the expression for $\ell(\mu, \Lambda, \psi)$. **3**

5C. In ICA, "There is some inherent ambiguities in the mixing matrix because of which it is impossible to recover the original sources $s^{(i)}$, given only the combined dataset $x^{(i)}$". Give reasons. **2**