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

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

SEVENTH SEMESTER B.TECH (INFORMATION TECHNOLOGY / COMPUTER AND COMMUNICATION ENGINEERING) DEGREE END SEMESTER EXAMINATION-NOVEMBER 2019
SUBJECT:PROGRAM ELECTIVE-V MACHINE LEARNING (ICT 4007)
(REVISED CREDIT SYSTEM)

TIME: 3 HOURS **26/11/2019** **MAX. MARKS: 50**

Instructions to candidates
- Answer **ALL** 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 θ as per the LMS algorithm. [5]

1B. Suppose you are given a well fitted machine learning model $\hat{f}$ that you want to apply to some test dataset. For each point x in your dataset, you want to predict the associated target $y \in \mathbb{R}$, and compute the mean square error (MSE)

$$E_{(x,y)\sim \text{test set}} |\hat{f}(x) - y|.$$

You realize that this MSE is too high, and try to find an explanation for this result, which could be either of the following

- overfitting
- underfittng,or
- data is noisy.

Derive a relation, that formalizes this intuition between *bias* and *variance*. [3]

1C. Briefly describe filter feature selection method. [2]

2A. Consider a Markov model with given set of states $S = \{s_1, s_2, \ldots, s_{|S|}\}$, wherein we can choose a series over time $\vec{z} \in S^T$.

i) State two Markov assumptions that will allow you to tractably reason about time series.

ii) Derive a relation to compute the probability of a state sequence, $P(\vec{z})$.

iii) Assume that the transition matrix from a weather system is given by

$$A = \begin{array}{c} s_0 \\ s_{sun} \\ s_{cloud} \\ s_{rain} \end{array} \overset{\begin{array}{cccc} s_0 & s_{sun} & s_{cloud} & s_{rain} \end{array}}{\begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0.8 & 0.1 & 0.1 \\ 0 & 0.2 & 0.6 & 0.2 \\ 0 & 0.1 & 0.2 & 0.7 \end{bmatrix}}$$

Compute the probability for sequence of observation

$$\vec{z} = \{z_1 = s_{sun}, z_2 = s_{cloud}, z_3 = s_{cloud}, z_4 = s_{rain}, z_5 = s_{cloud}\}.$$ [5]

2B. Consider a Markov model with a state transition matrix $A \in \mathbb{R}^{(|S|+1)\times(|S|+1)}$. For learning purpose, find the parameters A that maximize the log-likelihood for sequence of observations $\vec{z}$. The log-likelihood of a Markov model is defined as

$$l(A) = \sum_{i=1}^{|S|}\sum_{j=1}^{|S|}\sum_{t=1}^{T} 1\{z_{t-1} = s_i \wedge z_t = s_j\} \log A_{ij}.$$

[3]

2C. Consider the Poisson distribution parameterized by λ:

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

Show that the Poisson distribution is in the exponential family, and clearly state what are $b(y), \eta, T(y)$ and $a(\eta)$.

[2]

3A. Suppose you are given a dataset $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ consisting of m independent examples, where $x^{(i)} \in \mathbb{R}^n$, 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 the model are ϕ, Σ, μ_0 and μ_1. The log-likelihood of the data is given by

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

Apply MLE and derive the relation for ϕ, and μ_0. [5]

3B. Describe the following event models for text classification

i) Multi-variate Bernoulli event model

ii) Multinomial event model.

[3]

3C. Show that for a valid kernel, the kernel matrix K has to be positive-semi definite. [2]

4A. 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. Also, for each technique write its suitability condition. [5]

4B. 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 2γ worse than the best possible hypothesis in $\mathcal{H}$. [3]

4C. Show that PCA is a variance maximizing problem. [2]

5A. 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

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

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

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

where $\mu \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{n\times k}$, and the diagonal matrix $\Psi \in \mathbb{R}^{n\times n}$, $(k < n)$. Derive the expression for the log likelihood of the parameters $l(\mu, \Lambda, \Psi)$. [5]

5B. Consider a coin-flipping experiment in which you are given a pair of coins A and B of unknown biases θ_A and θ_B respectively (i.e., on any given flip, coin A will land on heads with probability θ_A and on tail with probability $(1 - \theta_A)$, similarly for coin B). Consider the dataset collected using following procedure five times: labels of the coins are removed. Now, randomly choose one of the two coin and perform ten independent coin tosses with the selected coin. Let $x^i = j$ denotes j number of heads obtained during i-th set of experiment. The dataset obtained from this experiment are

$$\{x^{(1)} = 5, x^{(2)} = 9, x^{(3)} = 8, x^{(4)} = 4, x^{(5)} = 7\}.$$

With initial estimate of biases $\hat{\theta}_A^{(0)} = 0.4$ and $\hat{\theta}_B^{(0)} = 0.7$, apply EM algorithm to compute $(\hat{\theta}_A^{(1)}, \hat{\theta}_B^{(1)})$. [3]

5C. Write the algorithms for *value iteration* and *policy iteration*. [2]