International Institute of Information Technology, Hyderabad
(Deemed to be University)

Statistical Methods in AI (CSE/ECE 471) - Spring-2019

Mid-semester Examination 2

Maximum Time : 90 Minutes Total Marks : 75

Roll No. _______________ Programme __________ Date __________

Room No. __________ Seat No. __________ Invigilator Sign. _______________

Marks secured

Multiple Choice Questions

Question:	1	2	3	4	5	6	7	8	9	10	11	Total
Points:	2	2	2	2	2	2	2	2	2	2	3	23
Score:												

Long Questions-1

Question:	12	13	14	15	16	17	18	19	20	Total
Points:	5	5	15	6	4	5	5	3	4	52
Score:										

General Instructions to the students

1. **QUESTION BOOKLET NEEDS TO BE RETURNED ALONG WITH ANSWER SHEETS. PLEASE TIE TOGETHER YOUR ANSWER SHEETS AND QUESTION BOOKLET, WITH THE BOOKLET ON TOP.**
2. Multiple-choice and True/False questions MUST be answered clearly within the question booklet itself. **NO MARKS FOR WRITING THE CHOICES IN ANSWER SHEET.**
3. **No questions will be answered during the exam. Make necessary <u>reasonable</u> assumptions, state them and proceed.**

True or False

Circle True or False. **NOTE: This section (True or False) has negative marking for incorrect answers.** (2 points each)

1. (2 points) True False Two random variables A, B are independent if $p(A, B) = p(A|B)p(B)$.
2. (2 points) True False By minimizing its loss function, k-means clustering always reaches the global minimum.
3. (2 points) True False Naive Bayes classifier finds a Maximum Aposteriori Probability (MAP) estimate of its parameters.
4. (2 points) True False Any boolean function can be learnt by a linear classifier (perceptron).
5. (2 points) True False Suppose x_1, x_2 are two data points with the same class label $\mathcal{A}$ and $x_1 \neq x_2$. Suppose $x_3 = \frac{x_1+x_2}{2}$ is a datapoint that belongs to a different class $\mathcal{B}$. No perceptron exists that classifies x_1, x_2 into $\mathcal{A}$ and classifies x_3 into class $\mathcal{B}$.
6. (2 points) True False Suppose we have a model from a fixed hypothesis set. As the amount of training data decreases, the possibility of overfitting the model increases.
7. (2 points) True False For a given dataset, a random forest classifier tends to have a lower bias than a decision tree.

Multiple Choice

Mark all answers you think are correct. No marks for partially correct answers.

8. (2 points) Consider the following regression model : $\arg\min_\theta \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2$. What does increasing λ do ?

   A. Bias of the model increases, Variance decreases

   B. Bias of the model increases, Variance stays the same

   C. Bias of the model decreases, Variance increases

   D. Bias of the model decreases, Variance stays the same

9. (2 points) Which of the following activation functions has an unbounded range ?

   A. ReLU ($max(x, 0)$) B. Linear C. Sigmoid D. Tanh

10. (2 points) For which of the following machine learning approaches can we have a kernel-ized version (similar to SVM) ?

    A. k-NN B. k-means C. PCA D. None of the above

11. (3 points) A 1-nearest neighbor classifier has than a ______________ than a 5-nearest neighbor classifier.

    A. larger variance B. larger bias C. smaller variance D. smaller bias

Long Questions

Write detailed answers. Adequately explain your assumptions and thought process.

12. (5 points) Figure 1 shows two plots, corresponding to the 2-D distribution of two different datasets. Suppose PCA is performed on the given data. Clearly draw the directions of the first and second principal component vectors in each plot. **NOTE: Draw directly on the plots in the question paper.**

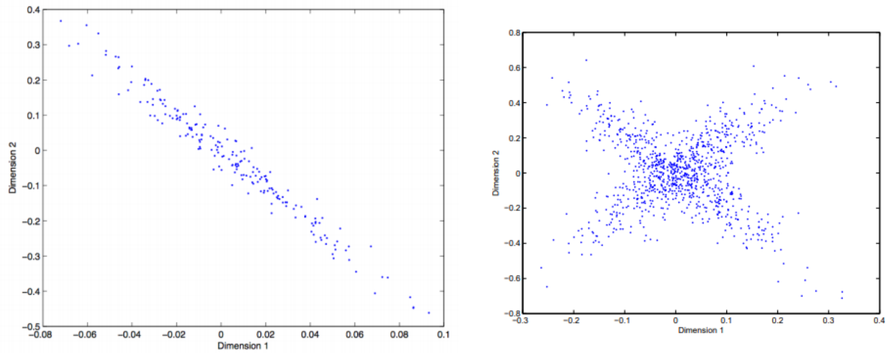

Figure 1:

13. (5 points) Suppose the month of the year is one of the attributes in your dataset. Currently, each month is represented by an integer $k, 0 \leqslant k \leqslant 11$ and let's say $k = 0$ corresponds to December,$k = 1$ to January etc. Come up with a feature representation $f(k)$ such the representation for December is at equal Euclidean distance from representations of January and November, i.e. $||f(0) - f(1)||_2 = ||f(0) - f(11)||_2$. Hint: $f(k)$ can be a vector.

14. (15 points) Figure 2 shows a 2-D dataset (circles). Suppose the k-means algorithm is run with $k = 2$ and the squares represent the initial locations of the estimated means. Indicate the new locations of the cluster means after 1 iteration of the k-means algorithm. Draw a triangle at the location of each cluster mean. Also write 1, 2 alongside each data point and the new cluster mean to show which data points belong to cluster 1 and which datapoints belong to cluster 2. Assume that datapoints whose locations do not align with integer axes coordinates have coordinates of 0.5. For e.g. the coordinates of top-left datapoint are $(0, 7)$. The coordinates of datapoint immediately to its right are $(0.5, 7)$

15. (6 points) The loss function for k-means clustering with $k > 1$ clusters, data-points $x_1, x_2 \dots x_n$, centers $\mu_1, \mu_2, \dots \mu_k$ and Euclidean distance is given by

$$L = \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - \mu_j\|_2^2$$

where S_j refers to points with cluster center μ_j. Suppose **stochastic** gradient descent with a learning rate of η is used. Derive the update rule for parameter μ_1 for a given data-point x_p. NOTE: x_p may or may not be a sample in S_1.

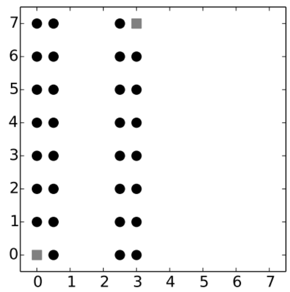

Figure 2:

Consider the following dataset (row is a data sample, each sample has two dimensions)

$$X = \begin{bmatrix} 6 & -4 \\ -3 & 5 \\ -2 & 6 \\ 7 & -3 \end{bmatrix}$$

Suppose PCA is used to determine the principal components.

16. (4 points) What are the unit vectors in the directions corresponding to the principal components ? HINT: There might be a faster way to guess the vectors instead of computing the covariance matrix.

17. (5 points) What is sum of eigenvalues corresponding to the principal components ?

---

18. (5 points) Figure 3 shows the truth table for a NAND logic gate. Implement the NAND function via a neural network architecture with a single neuron and an appropriate choice of weights, bias and activation function.

x_1	x_2	y
0	0	1
0	1	1
1	0	1
1	1	0

Figure 3:

In the lecture on SVM, we saw that one could use a mapping function $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^d$ to transform points from the original $\mathbb{R}^n$ space to another space $\mathbb{R}^d$. We also saw that one could define a kernel function $K(x, z)$ such that $K(x, z) = \phi(x)^T \phi(z)$. Suppose α is a positive real constant value. Suppose $\phi_1 : \mathbb{R}^n \longrightarrow \mathbb{R}^d, \phi_2 : \mathbb{R}^n \longrightarrow \mathbb{R}^d$ are feature mappings of K_1 and K_2 respectively. In terms of ϕ_1, ϕ_2

19. (3 points) Write the formula for the feature mapping ϕ_3 corresponding to $K(x, z) = \alpha K_1(x, z)$

20. (4 points) Write the formula for the feature mapping ϕ_3 corresponding to $K(x, z) = K_1(x, z) K_2(x, z)$