International Institute of Information Technology, Hyderabad

(Deemed to be University)

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

Mid-semester Examination 1

Maximum Time : 90 Minutes | Total Marks : 100

Roll No. _____ Programme _____ Date _____

Room No. _____ Seat No. _____ Invigilator Sign. _____

Marks secured

Multiple Choice Questions

Question:	1	2	3	4	Total
Points:	2	2	2	3	9
Score:					

Long Questions-1

Question:	5	6	7	8	9	10	11	12	13	14	15	Total
Points:	9	15	12	2	2	6	3	3	6	2	2	62
Score:												

Long Questions-2

Question:	16	17	18	19	Total
Points:	4	14	2	9	29
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. **No questions will be answered during the exam. Make necessary reasonable assumptions, state them and proceed.**

Multiple Choice Questions

For the following questions, specify ALL the correct answers. (Note: Partial marks will not be given for partially correct answers.)

1. (2 points) Suppose the k-means algorithm is used to cluster n samples from a dataset. Each sample is l-dimensional. Suppose the number of clusters is K and the number of iterations for k-means to converge is m. What is the order of the run-time for the algorithm ?

   A. $O(nKm)$ B. $O(nKlm)$ C. $O(nlm)$ D. None of the above

2. (2 points) For the same settings above, i.e. K, n, l, m, what is the order of total storage space for the k-means algorithm ?

   A. $O((K+n)*l)$ B. $O(Kl)$ C. $O(nKlm)$ D. None of the above

3. (2 points) In Figure 1, consider the 2-D dataset whose labels are given as + and −. What is the smallest value of k for which a point at location marked will be classified with a − label ? Assume Eulclidean distance.

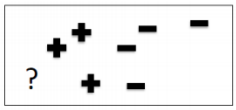


Figure 1:

   A. 1 B. 3 C. 4 D. 7

4. (3 points) Suppose X is a random variable with range $\{a_1, a_2, \ldots a_n\}$. What is the maximum possible value for $H(X)$ – the entropy of X ?

   A. $\frac{1}{n}$ B. $log_2(n)$ C. $\frac{n}{2}$ D. None of the above

Long Questions

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

Suppose we have a 2-class dataset with n samples. Each sample is labelled positive or negative class. Suppose the fraction of positive-labelled samples is x. It is also known that $x < 0.5$. We can define some simple non-machine learning classifiers that assign labels based simply on the proportions found in the training data as follows:

- Random Guess Classifier (RGC): Randomly assign half of the samples to positive class and the other half to negative.
- Weighted Guess Classifier (WGC): Randomly assign x fraction of the samples to positive class and the remaining $(1-x)$ fraction to negative class.
- Majority Class Classifier (MCC): assign the label of majority class to all the samples.

The baseline performance of these classifiers is determined by predicting labels on the n-sample dataset.

5. (9 points) Write down the confusion matrices for the three classifiers.

6. (15 points) Fill the following table (write your answer in the answer sheet only)

Classifier	Accuracy	Precision	Recall
RGC			
WGC			
MCC			

7. (12 points) Suppose we now have k classes and x_i represents the fraction of i-th class samples among the n samples. What is the accuracy for each of the classifiers specified above ?

For a k-means clustering setting, assume the following notation:

- K : The number of clusters
- n_k : Number of instances in k-th cluster
- μ_k : The center of k-th cluster
- x_{pq} : A data sample within the q-th cluster, i.e. $1 \leqslant p \leqslant n_q$

8. (2 points) Write down the expression J_k for the average Euclidean distance between members of k-th cluster.

9. (2 points) Write down the expression S_k for the sum of Euclidean distance of each cluster member from its center in k-th cluster.

10. (6 points) Let $J = \sum_k J_k$. Let $S = \sum_k S_k$. Derive the mathematical relationship between S and J.

Consider a labelled dataset with B binary input variables, $X_i \in \{0, 1\}, 1 \leqslant i \leqslant B$. The number of output classes is C.

11. (3 points) How many parameters (probabilities) must be estimated to train a Naive Bayes classifier on this data ?

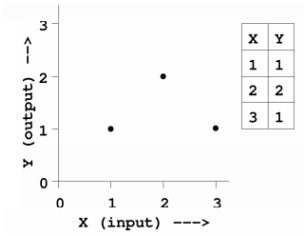
12. (3 points) How many parameters must be estimated if we do not make the Naive Bayes assumption ?

Suppose we have a set of observations $x_1, x_2, \ldots x_n$. It is assumed that the data has been generated by sampling from an exponential distribution

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geqslant 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{1}$$

13. (6 points) What is the maximum likelihood estimate of λ ?

For the set of points in Figure 2



X	Y
1	1
2	2
3	1

Figure 2:

14. (2 points) What is the equation of the least-squares-error linear regression line ?

15. (2 points) What is the value of the mean squared error for the estimated line ?

Suppose you are given a labelled dataset $\mathcal{D} = (x_i, y_i), 1 \leqslant i \leqslant N, x_i \in \mathbb{R}^d$ where the class labels are binary, i.e. $y_i \in \{0, 1\}$.

16. (4 points) Let $p(z) = \dfrac{e^z}{1+e^z}$. Show that its derivative $p'(z) = p(z)(1 - p(z))$.

17. (14 points) From the expression for the likelihood of the given data under the logistic regression model and from the equations used to obtain maximum likelihood estimate of the model parameters, show that $\sum_{i=1}^{N} y_i x_i = \sum_{i=1}^{N} P_i x_i$ where $P_i = p(y_i = 1 | x_i; \beta)$ and $\beta \in \mathbb{R}^{(d+1)}$ stands for the parameter vector of the logistic regression model. Hint: Use the result from the previous question.

Suppose we wish to predict the gender of a person based on two binary attributes: Leg-Cover (pants or skirts) and Facial-Hair (some or none). We have a dataset of 2,000 people, half male and half female. 75% of the males have no facial hair. Skirts are worn by 50% of the females. All females are fully bare-faced and no male wears a skirt.

$\log_2$ approx.

log2(1/8) = -3.0
log2(1/4) = -2.00
log2(1/3) = -1.58
log2(3/8) = -1.42
log2(3/7) = -1.22
log2(1/2) = -1.00
log2(4/7) = -0.81
log2(5/8) = -0.68
log2(2/3) = -0.58
log2(3/4) = -0.42
log2(7/8) = -0.19

Figure 3:

18. (2 points) What is the initial entropy of the dataset ?

19. (9 points) Suppose we wish to build a decision tree for our prediction task. Compute the Information Gain for each choice of 'Leg-Cover', 'Facial-Hair' as root node. Based on the gain values, which attribute is preferable as root node ? Use the values from Figure 3.