The students are required to write name and entry no.

Name ..... **Entry No:**.....

CS503

INDIAN INSTITUTE OF TECHNOLOGY ROPAR

CS503 - Machine Learning

Second Semester of Academic Year 2023- 2024

End semester Examination

Duration: 3 Hours | Max. Marks: 70 | Date:29-04-2024

Instructions:

- There are a total of **09 questions**. Check the whole paper.
- Question Q.0. is a bonus question and it is not compulsory.
- Use of scientific calculator is allowed.
- No clarifications will be entertained during the examination.
- Make appropriate assumptions wherever necessary and explicitly mention the same.
- Be precise and concise in your answers. Partial marks can be awarded for steps/ explanations.
- Go through all the questions before you start answering as there might be some questions present later that you may attempt easily.
- Write legibly so that it could be understood what you want to convey in your answers.

Q.0. Mention the book and the authors for where we followed the following topics in the course:

(A) ANN

(B) GMM

(C) Clustering

(D) CNN **[2 Bonus Marks if all answers are correct]**

Q.1. Short Answer. Justify the following statements in 1-2 lines only. [2x10 = **20 marks**]

A. Generalization error measures how well an algorithm perform on unseen data. The test error obtained using cross-validation is an unbiased estimate of the generalization error.

B. The True Positive Rate (TPR) measures the fraction of positively predicted instances among True Positives and True Negatives while False Positive Rate measures the fraction of negatively predicted instances among the False Positives and False Negatives.

C. A Bayesian Network encodes all type of conditional independences among the Random Variables.

D. The Truncated BPTT that does a forward pass for every timestep and perform the gradient update for the entire sequence length (n) is essentially the BPTT algorithm.

E. Running K-means with a larger value of K always enables a lower possible final objective value than running K-means with a smaller K.

F. The state distribution at a particular time "T" in a Markov Model of length "N" (N>T) will change with different initialization of starting state.

G. EM algorithm is well suited for finding maximum likelihood solutions for problems where complete data description is available.

H. If we generate all possible trajectories of states sequences in HMM and assign a score to each once of them as the sum of all the edge weights in the state-trellis, then sequence with the maximum score is given by the forward algorithm.

I. In agglomerative clustering, single link focuses on the most similar points between clusters while complete linkage is affected by the complete structure of both the clusters.

J. If we multiply the forward and backward probabilities for a particular time t and a particular state j, it will result in a joint probability of the whole observation sequence O along with being in state j at time t.

Q.2. Derive the E-M update rules for a univariate Gaussian mixture model (GMM) with two mixture components. Unlike the GMMs we covered in the course, the mean **μ** will be shared between the two mixture components, but each component will have its own standard deviation σ_k (i.e. σ_0 & σ_1). The mixture model is defined as follows: [2+2+3+3 = **10 marks**]

$$z \sim \text{Bernoulli}(\theta)$$

$$x \mid z = k \sim \mathcal{N}(\mu, \sigma_k)$$

(A)Write the probability density defined by this model (i.e. the probability of x, with z marginalized out)

(B) E-Step: Compute the posterior probability $\gamma(z_{ik}) = Pr(z_i = k \mid x_i)$.

(C) M-Step: Derive the update rule for μ (keeping σ_k fixed)

(D) M-Step: Derive the update rule for σ_1 (keeping μ fixed)

Q.3. Show that if any elements of the parameters π (start probability) or **A** (transition probability) for a Hidden Markov model (HMM) are initially set to zero, then those elements will remain zero in all subsequence updates of running the EM based Baum-Welch algorithm for learning the parameters of the HMM. **[4 marks]**

Q.4. Given the two-dimensional points in Table, assume that k = 2, and that initially the points are assigned to clusters as follows: $C_1 = \{x_1, x_2, x_4\}$ and $C_2 = \{x_3, x_5\}$. Apply the K-means algorithm until convergence, that is, the clusters do not change, assuming

(A) The usual Euclidean distance or the L2-norm as the distance between points

(B) The Manhattan distance or the L1-norm [2+2 = **4 marks**]

	X_1	X_2
$\mathbf{x}_1$	0	2
$\mathbf{x}_2$	0	0
$\mathbf{x}_3$	1.5	0
$\mathbf{x}_4$	5	0
$\mathbf{x}_5$	5	2

Q.5. Consider the joint probability distribution over 3 boolean variables x_1, x_2, y given in Figure (a) below. Consider also the marginal probabilities for this same distribution, given in Figures (b), (c), and (d). [1+2+3+3+2 = **11 marks**]

x_1	x_2	y	$p_D(x_1, x_2, y)$
0	0	0	.15
0	0	1	.25
0	1	0	.05
0	1	1	.08
1	0	0	.1
1	0	1	.02
1	1	0	.2
1	1	1	.15

(a) Joint distribution

	$x_1 = 0$	$x_1 = 1$
$y = 0$	.4	.6
$y = 1$	.66	.34

(b) $P_D(x_1|y)$

	$x_2 = 0$	$x_2 = 1$
$y = 0$	.5	.5
$y = 1$	.54	.46

(c) $P_D(x_2|y)$

y	$P_D(y)$
$y = 0$	.5
$y = 1$	.5

(d) $p_D(y)$

(A) What is the rule to assign class labels as used by the Bayes optimal classifier?

(B) Express $P_D(y = 0 \mid x_1, x_2)$ in terms of $P_D(x_1, x_2, y = 0)$ and $P_D(x_1, x_2, y = 1)$.

(C) Find the value of $P(y = 1 | x_1 = 1, x_2 = 0)$ predicted by the Bayes optimal classifier. Show your work.

(D) Find the value of $P(y = 1 | x_1 = 1, x_2 = 0)$ predicted by the Naive Bayes classifier. Show your work.

(E) The expressions and values that you must have written for (C) and (D) should be unequal. Explain why in one sentence.

Q.6. Consider the following product reviews, each labeled with a customer sentiment as **Happy** or **Angry**:

- Good, Value, Worth, Worth | **Happy**
- Bad, Not-worth, Usable | **Angry**
- Value, Delivery, Bad, Good, Good | **Happy**
- Not-worth, Usable, Usable, Good | **Angry**
- Delivery, Bad, Usable, Worth | **Angry**

$P(H|X) = \frac{P(H|B)\,P(B)}{P(H)}$

You got a new review as **X: Bad, Value, Usable, Delivery.**
Compute the most likely class for X using a naive Bayes classifier and use add-1 smoothing for the likelihoods. **[3 marks]**

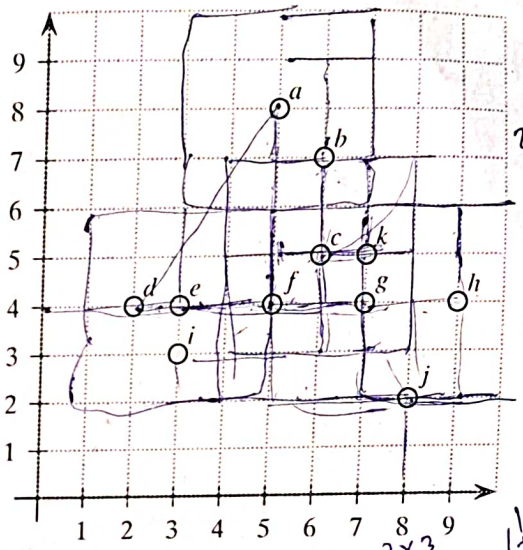
Q.7. For a multi-class classification problem with $k \geq 3$, derive the weight update equations for the connections between the last hidden layer and output layer. The output layer employs the softmax function and hidden layer employs sigmoid as activation functions. **[4 marks]**

Q.8. Consider the following distance measures for a d-dimensional space: [2+3 = **5 marks**]

$$L_{\infty}(x, y) = \max_{i=1}^{d} \{ | x_i - y_i | \}$$

$$L_{pow}(x, y) = \left(\sum_{i=1}^{d} 2^{i-1} (x_i - y_i)^2 \right)^{1/2}$$

For the points shown in the plot,



(A) Using $\epsilon = 2$, $minpoints = 5$ & L_{∞} distance, find all the core, border and noise points.

(B) Using $\epsilon = 4$, $minpoints = 3$ & L_{pow}, show the cluster found by DBSCAN.

Q.9. Briefly describe the following: [3x3 = **9 marks**]

(A) What are the clustering objectives functions in Spectral Clustering? Discuss the graph laplacian and undirected mutual kNN graph.

(B) What is a consistent hypothesis? Discuss that every consistent hypothesis is a MAP hypothesis.

(C) What are the three major advantages of CNN over ANN? Discuss each one in brief.

*****End*****