NITTE EDUCATION TRUST

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

An Autonomous Institution Approved by UGC/AICTE/Govt. of Karnataka
Accredited by NBA (Tier – I) and NAAC 'A+' Grade
Affiliated to Visveswaraya Technological University, Belagavi
Post Box No. 6429, Yelahanka, Bengaluru – 560 064, Karnataka, India

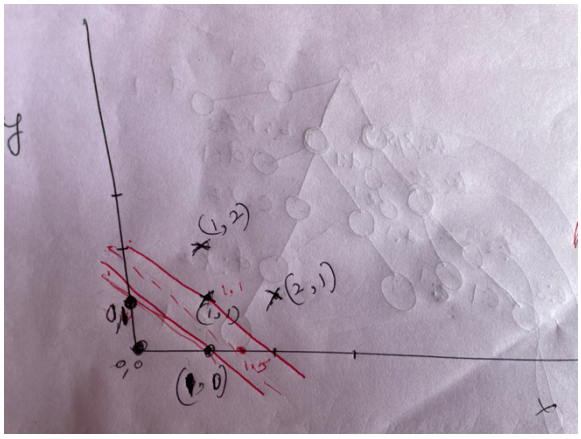
KNOWLEDGE • CHARACTER • UNITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MID SEMESTER EXAMINATION-III – Scheme & Solution

Course Title with code	Introduction to Machine Learning, 18CSE751	**Maximum Marks**	**30 Marks**
Date and Time	02-02-22, 2 pm to 3 pm	**No. of Hours**	1.0
Course Instructor(s)	Dr. Vani V		
Instructions to Students			
1. Answer any **two full questions.** 2. Any missing data may assume suitably.			

Q. No	Question	MAX MARKS	CO	BL	PO/PSO
1. a	Compare parametric approach with non-parametric approach with necessary examples **Solution:** **Parametric approach :** The sample comes from a known distribution. – The advantage of any parametric approach is that given a model, the problem reduces to the estimation of a small number of parameters, which, in the case of density estimation, are the sufficient statistics of the density. – For example: the mean and covariance in the case of Gaussian densities. – used quite frequently – Assume rigid parametric model may be a source of bias in many applications where this assumption does not hold. **Examples:** Logistic Regression,Linear Discriminant Analysis, Perceptron, Naive Bayes, Simple Neural Networks **Non-Parametric Approach:** Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data. Non-parametric methods seek to best fit the training data in constructing the mapping function, whilst maintaining some ability to generalize to unseen data. As such, they are able to fit a large number of functional forms. An easy to understand nonparametric model is the k-nearest neighbours algorithm that makes predictions based on the k most similar training patterns for a new data instance. The method does not assume anything about the form of the mapping function other than patterns that are close are likely to have a similar output variable. **Examples:** k-Nearest Neighbors, Decision Trees like CART and C4.5, Support Vector Machines	4	4	2	1,2,3/2

1. b	Consider the following training data and answer the questions (a) to (c). class 1: (1,1), (1,2), (2,1) class 2: (0,0), (1,0), (0,1) **Solution:** (a) Plot these six training points. Are the classes {class1, class2} linearly separable?  (b) Construct the weight vector of the maximum margin hyperplane by inspection and identify the support vectors. The maximum margin hyperplane should have a slope of −1 and should satisfy x1 = 3/2, x2 = 0. Therefore it's equation is x1 + x2 = 3/2, and the weight vector is $(1, 1)^T$ (c) If you remove one of the support vectors does the size of the optimal margin decrease, stay the same, or increase? In this specific dataset the optimal margin increases when we remove the support vectors **Scheme:** Each answer carries 2 marks 3 x 2 = 6 Marks	6	4	3	1,2,3/2
1.c	Discuss the need for linear, polynomial and RBF kernel functions in Support Vector Machine (SVM) and show how to use SVM for multi-class classification. Kernel Function generally transforms the training set of data so that a non-linear decision surface can be transformed to a linear equation in a higher number of dimension spaces. Linear Kernel: used when data is linearly separable. Polynomial Kernel: It represents the similarity of vectors in training set of data in a feature space over polynomials of the original variables used in kernel. RBF: It is used to perform transformation when there is no prior knowledge about data • polynomials up to some degree s in the elements x_k of the input vector (e.g., x_3^3 or $x_1 \times x_4$) with kernel: $$\mathbf{K}(\mathbf{x},\mathbf{y}) = (1+\mathbf{x}^T\mathbf{y})^s \quad (8.19)$$ For $s = 1$ this gives a linear kernel • radial basis function expansions of the x_ks with parameter σ and kernel: $$\mathbf{K}(\mathbf{x},\mathbf{y}) = \exp\left(-(\mathbf{x}-\mathbf{y})^2/2\sigma^2\right)$$ ■ The SVM only works for two classes. ■ For the problem of N-class classification, – *train an SVM that learns to classify class one from all other classes,* – *then another that classifies class two from all the others.* – *So, for N-classes, we have N SVMs.* ■ This still leaves one problem: How do we decide which of these SVMs	5	4	2	1,2,3/2

is the one that recognises the input?

- *The answer is just to choose the one that makes the strongest prediction*
- *Classifier return either the class label (as the sign of y) or the value of y*
- *The value of y is telling us how far away from the decision boundary it is, and clearly it will be negative if it is a misclassification.*

Therefore, ***use the maximum value of this soft boundary as the best classifier***

Different kernels produce different decision boundaries

Scheme: 3 marks kernel, 2marks for Multiclass SVM

2. a — 10 | 5 | 3 | 1,2,3/2

For the given intial set of three clusters (k=3)

C1 = {(1,3),(3,6),(3,5)}
C2 = {(5,3),(6,7),(2,2)}
C3 = {(6,5),(3,1),(2,3)}

Apply k-means algorithm and show when the clusters converge.

Note: Use Manhattan Distance

Solution :

Let points be considered as 1 to 9

C1 = {(1,3),(3,6),(3,5)}
C2 = {(5,3),(6,7),(2,2)}
C3 = {(6,5),(3,1),(2,3)}

Initial				M1		M2		M3	
		Data Points		2.33	4.6667	4.333	4	3.667	3
	1	1	3	3		4.333		2.667	
	2	3	6	2		3.333		3.667	
	3	3	5	1		2.333		2.667	
	4	5	3	4.33		1.667		1.333	
	5	6	7	6		4.667		6.333	
	6	2	2	3		4.333		2.667	
	7	6	5	4		2.667		4.333	
	8	3	1	4.33		4.333		2.667	
	9	2	3	2		3.333		1.667	
C1(2,3)									
C2(5,7)									
C3(1,4,6,8,9)									
Iteration 1:				M1		M2		M3	
	Data Points			3	5.5	6	6	2.6	2.4
	1	1	3	4.5		8		2.2	
	2	3	6	0.5		3		4	
	3	3	5	0.5		4		3	
	4	5	3	4.5		4		3	
	5	6	7	4.5		1		8	
	6	2	2	4.5		8		1	
	7	6	5	3.5		1		6	
	8	3	1	4.5		8		1.8	
	9	2	3	3.5		7		1.2	
Initial and Iteration 1 are same the algorithm converges									
C1(2,3)									
C2(5,7)									
C3(1,4,6,8,9)									

Scheme : Each iteration : 5 Marks , 5 x 2 = 10 Marks

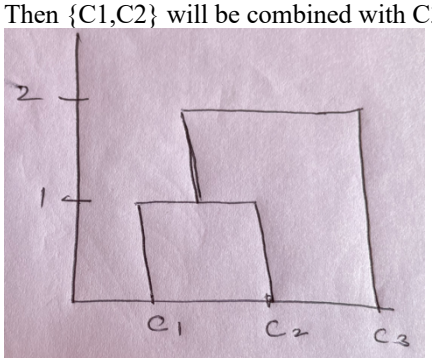
2. b — 5 | 5 | 2 | 1,2,3/2

Discuss the use of clustering in supervised learning with suitable examples

Solution:

- Dimensionality reduction methods find correlations between features and group features
- Clustering methods find similarities between instances and group instances
- Allows knowledge extraction through

	number of clusters, *prior probabilities,* *cluster parameters, i.e., center, range of features.* Example: CRM, customer segmentation ■ Estimated group labels h_j (soft) or b_j (hard) may be seen as the dimensions of a new k dimensional space, where we can then learn our discriminant or regressor. ■ Local representation (only one b_j is 1, all others are 0; only few h_j are nonzero) vs Distributed representation (After PCA; all z_j are nonzero) **Scheme:** 5 x 1 = 5 Marks				
3. a	Identify a suitable algorithm to estimate the gender of the students studying in a college given their height information. Discuss the identified algorithm with its iterative steps. **Solution:** ■ The Expectation-Maximization (EM) algorithm is used in maximum likelihood estimation where the problem involves two sets of random variables of which one, **X, is observable and the other, Z, is hidden**. ■ The goal of the algorithm is to find the parameter vector Φ that maximizes the likelihood of the observed values of X, L(Φ|X). But in cases where this is not feasible, we associate the extra hidden variables Z and express the underlying model using both, to maximize the likelihood of the joint distribution of X and Z, the complete likelihood Lc(Φ|X,Z). ■ Since the Z values are not observed, we cannot work directly with the complete data likelihood Lc ; ■ Instead, we work with its expectation, $\mathcal{Q}$ given X and the current parameter values Φ^l, where l indexes iteration. This is the expectation (E) step of the algorithm. ■ Then in the maximization (M) step, we look for the new parameter values, Φ^{l+1}, that maximize this. ■ Iterate the two steps 1. E-step: Estimate z given X and current Φ 2. M-step: Find new Φ' given z, X, and old Φ. $\text{E-step}: \mathcal{Q}(\Phi \mid \Phi') = E\left[\mathcal{L}_C(\Phi \mid \mathcal{X}, \mathcal{Z}) \mid \mathcal{X}, \Phi'\right]$ $\text{M-step}: \Phi^{l+1} = \underset{\Phi}{\arg\max}\, \mathcal{Q}(\Phi \mid \Phi')$ An increase in Q increases incomplete likelihood $\mathcal{L}(\Phi^{l+1} \mid \mathcal{X}) \geq \mathcal{L}(\Phi^{l} \mid \mathcal{X})$ **Scheme:** identifying and explaining EM algorithm : 2 Marks E-Step : 1.5 Mark M-Step: 1.5 Mark	5	5	2	1,2,3/2
3. b	For the given intial set of three clusters **C1 = {(1,3),(3,6),(3,5)}** **C2 = {(5,3),(6,7),(2,2)}** **C3 = {(6,5),(3,1),(2,3)}** Apply the agglomerative algorithm with complete cluster distance measures to produce a dendrogram tree Note : Use Manhattan distance to compute distance matrix Solution: C1 -> C2 : max distance : (5+4) = 9 C3->C1 : max distance (5+2) = 7 C3-> C2 : max distance (3 + 6) = 9	10	5	3	1,2,3/2

	C1 and C2 will be combined after calculating max distances Then {C1,C2} will be combined with C3  Calculation of max distances between clusters : 3 x 3 = 9 Marks Dendrogram : 1 Mark				

Faculty Signature	Course Co-Ordinator/Mentor Signature	HoD Signature