May 7, 2024
IIT Jodhpur

End Semester Exam
CSL2050 - Pattern Recognition and Machine Learning

NOTE:

This is Question-cum-Answer sheet. Maximum Points: 105, Total Questions: 9, Total Pages: 12 (6 sides), Total Time: 2 Hours. **If there is anything not clear in the problems, go ahead with your own assumptions but state them clearly. No doubts will be entertained during the exam.** Be precise and write the answer in the box provided. Verbosity will be penalized. Use the other answer sheet for rough work and submit both.

Name: **Roll Number:** **Signature:**

Question Number	**Topic**	**Total Marks**	**Marks Obtained**
1	Short Answers	20	
2	Artificial Neural Networks	15	
3	SVM	10	
4	Bayes Theorem	10	
5	GMM	10	
6	Kmeans and Agglo. Clustering	10	
7	PCA	10	
8	Decision Tree	10	
9	True/False	10	
	Total	**105**	

1. **Short Answers. Please be to the point.**(**10** $\times$ **2 pts**)

(a) I want to split my 10,000 data points into training and held-out validation sets. Ananya recommends using 9,900 points for training and 100 for validation. Sunita suggests a split of 7,000 training points and 3,000 validation points, while Priya thinks using 100 training points and 9,900 validation points is optimal. Whose advice should I follow, and why? Why should I not go along with the other two suggestions?

(b) Suppose 10,000 2D-data points are uniformly distributed in the peripheral of a unit radius origin-centered circle. What will be the variance(x) and variance(y)?

(c) Swati is developing an automatic assessment tool for T/F questions. She begins with a model that always predicts 5 points out of 10 for all answersheet. Suppose the dataset has a uniform points distribution between 1 and 10. Then, what will Swati's Model perform in terms of Mean Absolute Error (assume discrete points and 100 samples)?

(d) Write down the formula for the cross entropy loss function. What is the expected random binary classifier loss when cross-entropy loss is computed?

(e) Suppose you have to design an image classifier that classifies landmark places at the IIT Jodhpur campus (specifically, Brain Tree, LHC, Clock Tower, Shamiyana, Jodhpur Club, and Indoor Sports Complex). The goal is to make it usable for non-IITJ residents. How will you approach this problem from scratch? Write three important phases (chronologically) toward solving and finally deploying this solution.

(f) Suppose that in a population of 100 students, RTPCR tested students s_1 and s_2 as positive and the remaining as negative. But, students $\{s_1, s_2, s_5, s_9, s_{10}, s_{11}, s_{100}\}$ were actually positive. Compute the F1-score of the RTPCR test.

(g) What is the VC dimension of $f(x, b) = sign(x.x - b)$. This classifier classifies samples with non-negative f as positive class and remaining as negative.

(h) Neatly draw bias and variance curves with respect to increasing model complexity.

(i) I have randomly initialized an MLP that is giving me 70% accuracy for binary classification over a test set of 10000 samples. Is it possible? If yes, list out the scenarios. If no, give the reason.

(j) Which among the following is more prone to overfitting: (i) Increasing the number of layers in a neural network or (ii) performing data augmentation during training?

2. **(Artificial Neural Networks)** Consider a three-layer neural network. Assume the input features are 2-dimensional. Further, hidden and output layers have two and one neuron, respectively. Input and hidden layers are fully connected with weights $[w_1, w_2, w_3, w_4]$. The hidden and output layers are fully connected with weights $[w_5, w_6]$. The neurons in the hidden layer have bias b_1 and b_2, respectively, and the output layer neuron has bias b_3. Each neuron in the network has ReLU activation. Note $ReLU(x) = max(0, x)$.

(a) Draw the neural network neatly. **(3 pts)**

(b) Assume $x_1 = 1, x_2 = 1$ is a training sample with label $y = 1$ and weights and bias are randomly initialized as follows: $[w_1, w_2, w_3, w_4, w_5, w_6]$=$[0.1, 0.1, 0.1, 0.1, 0.2, 0.2]$ and $[b_1, b_2, b_3] = [0.2, 0.2, 0.5]$. Using the feed-forward pass, compute the output of the network $\hat{y}$. Write each neuron's output too. Only write the final answers here. **(2 pts)**

(c) For the above sample, compute loss value $\mathcal{L}$, assume loss as $\frac{(\hat{y}-y)^2}{2}$. **(1 pts)**

(d) Now, compute numerical gradients of the loss function with respect to all the weight and bias terms, i.e $\frac{\partial \mathcal{L}}{\partial w_i}$ and $\frac{\partial \mathcal{L}}{\partial b_j}$ for $i \in \{1, 2, 3, 4, 5, 6\}$ and $j \in \{1, 2\}$. **(Do the computation in the rough sheet and only write final answer here). (7 pts)**

(e) Using gradient descent, update the weights and bias. Assume learning rate = 0.01. **(2 pts)**

3. **(SVM)** (a) Find out a closed form non-linear transformation $\phi(x)$ for the following 1-D data points so that they become linearly separable: $(1,+),(2,+),(3,-),(4,+),(5,+),(6,-),(7,+),(8,+),$ $(9,-)$. **(2 pts)**

(b) In class, we learned that SVM can be used to classify linearly inseparable data by transforming it to a higher dimensional space with a kernel $K(x,z)=\phi(x)^T\phi(z)$, where $\phi(x)$ is a feature mapping. Let K_1 and K_2 be $R^n \times R^n$ kernels, K_3 be a $R^d \times R^d$ kernel and $c \in R^+$ be a positive constant. $\phi_1 : R^n \rightarrow R^d$, $\phi_2 : R^n \rightarrow R^d$, and $\phi_3 : R^d \rightarrow R^d$ are feature mappings of K_1, K_2 and K_3 respectively. Explain how to use ϕ_1 and ϕ_2 to obtain the following kernels: **(2 + 2 pts)**

(i) $K(x,z)=cK_1(x,z)$

(ii) $K(x,z)=K_1(x,z)K_2(x,z)$

(c) Write down the soft-margin SVM objective function. Introduce each variable. What is the role of C parameter in it. **(2+2 pts)**

4. **(Bayes Theorem)** (a) Suppose Ajay has to take one of the three paths, namely A_1, A_2, and A_3. The chances of encountering a monster in these paths are 0.3, 0.6, and 0.75, respectively. Suppose there is no prior knowledge of encountering any existing monsters along the path. If Ajay does not come out alive, what is the probability that he took path A_3? (Also, note that there is no way to come alive if one encounters the monster). **(6 pts)**

(b) For data D and hypothesis H, Choose the most specific relations among the following that always holds: (i) $\leq$ (ii) $\geq$ (iii) $=$ (iv) depends **(Note: "depends" is the least specific. Assume all probabilities are non-zero.) (2+2 pts)**

(A) $P(H = h|D = d)$ □ $P(H = h)$

(B) $P(H = h|D = d)$ □ $P(D = d|H = h)P(H = h)$

5. **(GMM)** (a) Given $p_1(x)$, $p_2(x)$, $\cdots$ $p_k(x)$ as Gaussian Distribution, prove that p(x), which is defined as follows is also a valid PDF:

$p(x) = \sum_{i=1}^{k} \pi_i p_i(x)$ where $\sum_i \pi_i = 1$ and all π_is are non-zero. (Note that Gaussian are valid probability density function). **(3 pts)**

(b) Suppose two clusters of images, namely apple and orange, follow Gaussian Distribution $\mathcal{N}(\mu_a, \Sigma_a)$ and $\mathcal{N}(\mu_o, \Sigma_o)$ respectively. The prior probabilities of apple and orange are π_a and π_o, respectively. Compute the probability that an image x_i belongs to cluster apple using the E-step of the EM-algorithm. **(3 pts)**

(c) **(4 points)** Re-Estimate the $\mu_a, \Sigma_a, \mu_o, \Sigma_o$ using M-Step of EM-algorithm. **(4 pts)**

6. (a) Given the following points: 1, 5, 20, 25, 26, 27, 28, 40, 44. Show agglomerative clustering. (Hint: distance between points a and b is abs(a-b).) **(5 pts)**

(b) If we use K-means clustering on the above data and initialize cluster centers as 5, 25, and 45. After the first iteration, update the cluster means and show cluster assignments for each point. Use the same distance as in Question 6(a). **(5 pts)**

7. **(PCA)** Given the following 2D points: (-1,1), (-2,2), (-3,3).

(a) Compute zero mean data points: **(2 pts)**

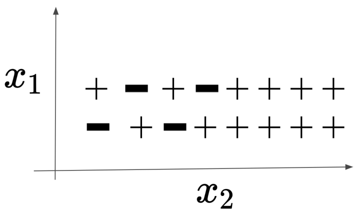
(b) Compute covariance matrix of zero mean data points: **(2 pts)**

(c) Find out the Eigen values and Eigen vectors of the covariance matrix. **(2 pts)**

(d) Project all the zero mean data points to the first principle component. **(2 pts)**

(e) Reconstruct the projected data points and compute reconstruction error. Use mean absolute error. **(2 pts)**

8. **(Decision Tree)** Consider the data points shown in the Figure:



Now, consider the following two extreme decision tree algorithms: (i) The ABC algorithm constructs a decision tree using the conventional approach but refrains from pruning at any stage. (ii) Conversely, the XYZ algorithm avoids the risk of splitting altogether, resulting in the entire decision tree being a single leaf node.

(a) What is the precise count of leaf nodes in the ABC decision tree generated on this dataset? **(3 pts)**

(b) Could you provide the leave-one-out classification error of applying ABC to this dataset? Please report the total number of misclassifications. **(3 pts)**

(c) What is the leave-one-out classification error when utilizing XYZ on our dataset? Please report the total number of misclassifications. **(3 pts)**

(d) Which of ABC and XYZ will overfit to the training data? **(1 pts)**

(Please Turn Over) →

9. **(10 pts)** Write TRUE or FALSE in capital letters and clear handwriting.

Write your roll number here:	

SN	**Statement**	**TRUE/FALSE**
1	Typically, the storage requirement for KNN is much larger than SVM for deployment.	
2	KNN is a non-linear classifier.	
3	$1 - cosineSimalarity(x, y)$ is a valid distance metric for any d-dimensional vectors x and y.	
4	Backpropogation is used to update the weights, while gradient descent is used to compute the gradients in the neural network.	
5	The Receiver Operating Characteristic (ROC) curve is a plot between Precision and Recall.	
6	A mixture of two Gaussian distributions is always a Gaussian distribution.	
7	The Time required for making a prediction in the Decision Tree is directly proportional to the number of nodes.	
8	The decision boundary of an SVM can be non-linear in the original space.	
9	For multiclass problems with a large number of classes, making a prediction using one-vs-one SVM is slower than one-vs-rest SVM.	
10	With a high value of the C parameter in SVM, the risk of overfitting increases, as the model may learn to capture specific noise or details in the training data that do not generalize well to new, unseen data.	