Max. marks: 70   Max. time: 3 Hours

Note1: Attempt all questions. All questions carry equal marks.

Note2: For full credit use proper notation, follow systematic approach towards the problem and show all intermediate steps neatly.

1. (a) Given the distribution $f(x) = \lambda e^{-\lambda x}$, find the MLE for the parameter λ. (6)

   (b) Suppose we obtain data from a cubic function and try to fit linear model ($M1$) and 12th degree polynomial ($M2$) to it. Assuming, standard Gaussian noise in the data, what can you say about the *bias* and *variance* of $M1$ and $M2$. State the reason(s) for your answer. (2)

   (c) Consider a decision tree for a 2-class problem, trained on n training points, each having d real-valued features. At each node of the tree, every possible splitting value for each feature is examined and the split that maximizes the information gain is chosen. Find the time complexity of the algorithm when h is the depth of the tree. Assume that linear time sorting algorithm is available. (6)

2. (a) Derive the expression for cross-entropy loss for K-class classification problem. (8)

   (b) Answers any *three* of the following questions in *exactly* two sentences each. (6)

      i. List two strategies that can help reduce overfitting in decision trees.

      ii. How can neural networks be used for regression as well as classification?

      iii. What is the difference between LASSO and Ridge regularization?

      iv. Let λ_{ik} be the loss incurred for taking action α_i, when the correct class is C_k. What is the expected risk?

3. (a) Explain (using relevant notation) why EM algorithm delivers soft clustering, while K-Means outputs hard clustering. (4)

   (b) How do the sampling methods differ in Bagging and AdaBoost? Explain with clear notation, how AdaBoost algorithm computes the weights of dependent classifiers? (6)

   (c) Neural networks can learn nonlinear decision boundaries by ———. (2)

   (d) For autonomous vehicle driving, a neural network is to be trained using grayscale 64×64 pixel images. The training labels include the human driver's steering wheel angle in degrees and the human driver's speed in miles per hour. The architecture includes a hidden layer with 2,048 units and an output layer with 2 units (one for steering angle, one for speed). An activation function is used for the hidden units. How many weights are learned by the network?. (2)

4. (a) How is LASSO useful for feature selection? (3)

   (b) What is color quantization? Explain how is K-Means clustering used for color quantization (5)

(c) Consider 4-dimensional training set with N instances, categorized in 3 classes (Appearance), with details of attributes given below.

Age [Young, Mid-age, elderly]

Weight [heavy, light]

Weight [heavy, light]

Dress [Western, Ethnic]

Appearance [Smart, Dashing, Unimpressive]

List the conditional probabilities that need to be computed for NB classifier. What is the size of the model? (6)

5. (a) Given the bi-variate normal data partitioned into two (practically) linearly separable classes, draw iso-probability contours when

i. the two classes have diagonal covariance matrix with equal variances on both dimensions

ii. the two classes have diagonal covariance matrix with unequal variances on dimensions

iii. the two classes have arbitrary but shared covariance matrix with equal variances

The figures should be labeled neatly. (6)

(b) Show that for two Gaussian classes with shared covariance matrix Σ, and respective means μ_1, μ_2, the log odds is linear. (5)

(c) How many binary classifiers are required to solve a k-class classification problem? Explain your answer. (3)