Indian Statistical Institute

Mid-Semester Examination : 2018 – 2019

Master of Mathematics, Semester III

Special Topic: Statistical Learning Theory

Date: 3 March 2019 Maximum Marks: 50 Duration: 2 hours

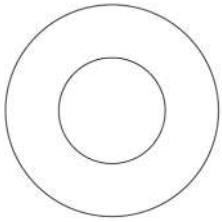
Attempt all the questions. Credit will be given for precise and brief answers.

1. What is supervised learning and what is unsupervised learning? Present one example of each with reason. What is over fitting and what is under fitting? Show that tradeoff between over fitting and under fitting is equivalent to tradeoff between bias and variance. 2 + 4 + 2 + 4 = 12

2. Let us consider a mixture of two different data consisting of n_1 member of type 1 and n_2 member of type 2, where n_1 and n_2 are finite. If they are linearly separable in a n dimensional linear space, show that they are separable by a pair of parallel hyper planes, such that, no data falls in between the parallel planes. Show the pair of parallel planes with maximum separation is a unique pair, or, if the assertion is wrong give a counter example (a convincing diagram will do). 6 + 6 = 12

3. (a) The number of input and output nodes in an artificial neural network (ANN) is fixed, but number of hidden nodes and connection weights are not. They are called free parameters of an ANN. “If there are too few free parameters the network will not be able to learn the training set well enough. If there are too many of them the network will not generalize.” Please explain the statement within the double quote. 6
   (b) Describe logistic regression classifier. 6

4. (a) Consider the following configuration in a two dimensional space. If a neural network is designed to identify this configuration at the least how many hidden layers the neural network must have and why? 5



(b) In Fisher's linear discriminant the expression $J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_\mathbf{w} \mathbf{w}}$ has to be maximized, where $\mathbf{w}$ is a vector, $\mathbf{S}_B$ and $\mathbf{S}_\mathbf{w}$ are nonsingular, square matrices of appropriate dimension. Show that when $J(\mathbf{w})$ is maximum, $\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_\mathbf{w} \mathbf{w}$ will have to hold for some scalar λ. Also show that for the optimally discriminating hyperplane $\mathbf{w}^T \mathbf{x} = c$, $\mathbf{w}$ is given by $\mathbf{w} = \mathbf{S}_\mathbf{w}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ (assume that $\mathbf{S}_B$ is in the direction of $\mathbf{m}_1 - \mathbf{m}_2$, only the direction of $\mathbf{w}$ matters not the magnitude). $\mathbf{m}_1$ and $\mathbf{m}_2$ are mean of the two data sets respectively which will have to be optimally discriminated (separated) from each other by a hyperplane. 2.5 + 2.5 = 5

(c) Describe k-nearest neighbor classification algorithm. If the size of the training set is n and the test set is bigger than the training set, show that the time complexity of the naïve algorithm on the test set is bounded below by $O(n^2.\log(n))$, given that the sorting is always done by an O(m.log(m)) algorithm for input size m. 4