Indian Statistical Institute

Semester Examination : 2018 – 2019

Master of Mathematics, Semester IV

Statistical Learning Theory (Special Topic)

Date: 13/05/2019 Maximum Marks: 100 Duration: 3 hours

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

1. Principal component analysis (PCA) for data decomposition works best when the components are in mutually orthogonal or near orthogonal orientation – explain. In independent component analysis (ICA) the data must not be distributed normally – why? Would you suggest PCA instead if the data is normally distributed? Why?

   6 + 6 + 2 + 6 = 20

2. What is shattering by class of sets and shattering by class of functions? What is VC dimension? Without using any result show by geometric construction that the VC dimension of the class of two dimensional planes in three dimensional Euclidean space is 4. Show that $\sqrt{\frac{1}{N}\left[D\left(\log\left(\frac{2N}{D}\right)+1\right)-\log\left(\frac{\eta}{4}\right)\right]}$ is an increasing function in D, where D is VC dimension. N is a natural number and η is a positive real number. Now with the help of the following formula, define structural risk minimization.

   $$P\left(test\ error \le training\ error + \sqrt{\frac{1}{N}\left[D\left(\log\left(\frac{2N}{D}\right)+1\right)-\log\left(\frac{\eta}{4}\right)\right]}\right)=1-\eta$$ 2 + 2 + 2 + 6 + 4 + 4 = 20

3. (a) Let $C(X)$ be the convex set generated by set X in an Euclidean space. Let H be a hyperplane, such that, $C(X)\bigcap H=\phi$ (the null set). Let $a\in X$ be a point whose distance from H is minimum among all points in X. Show that there cannot be a point $b\in C(X)$, such that, distance of b from H is shorter than the distance of a from H.

   10

   (b) Hence or otherwise show that two finite sets A and B are linearly separable in an Euclidean space if and only if $C(A)$ and $C(B)$ are linearly separable. 10

4. How would you define kernel in statistical learning theory? Consider four points in two dimensional Euclidean plane (1,1), (5,1), (5,5) and (1,5). (1,1) and (5,5) belong to one type of data, say T_1, and (5,1) and (1,5) belong to another type of data, say T_2. Are T_1 and T_2 linearly separable in the two dimensional Euclidean plane? Justify your answer with a suitable diagram. Let $\mathbf{x}=(x_1, x_2)^T$, $\varphi:\Re^2\to\Re^3$, such that, $\varphi(\mathbf{x})=(x_1, x_2, x_1x_2)^T$. Show that T_1 and T_2

are linearly separable in $\Re^3$ after being mapped by φ. Drawing a diagram with approximately correct location of the four points in the 3D space with a separating hyperplane will be enough. Give the equation of one such hyperplane and show T_1 falls on side of the hyperplane and T_2 on the other. 2 + 3 + 10 + 5 = 20

5. Write short notes on (a) k-means clustering and (b) hierarchical clustering. Present a comparative study between these two types of clustering. 4 + 8 + 8 = 20