CHENNAI MATHEMATICAL INSTITUTE

Quiz 1 : FML

Date: Sep 18, 2024

(1) Let χ be a domain consisting of three points x_1, x_2, x_3. Let $\mathcal{D}$ be the distribution on χ with probability of x_1, x_2, x_3 being $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{6}$ respectively. Let f be a labelling function which assigns x_1,x_2, x_3 the binary values $1, 0, 1$ respectively. Suppose h is a hypothesis which assigns x_1, x_2, x_3 the values $0, 0, 1$ respectively. What is $L_{\mathcal{D}}(h)$ with respect to the labelling function f. Suppose we pick up sample $S = x_1, x_2$, and our learning algorithm returns h. What is $L_S(h)$? Compute $\mathbb{E}_{S' \sim \mathcal{D}^2} L_{S'}(h)$. Recall this means we are picking two elements independently from χ under the distribution $\mathcal{D}$ and computing the expected empirical risk. 5 marks

(2) Let $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_r$ be hypothesis classes of functions from a domain χ to $\{-1, 1\}$. Suppose the VC dimension of these classes are $d_1, \ldots, d_r$ respectively. Let $d = max_i d_i$ and $\mathcal{H} = \cup_{i=1}^{i=r} \mathcal{H}_i$. Given a subset of size k give an upper bound on the number of different functions you can get by using elements from $\mathcal{H}$. Using the inequality $x \leq a \log x + b \implies x \leq 4a \log(2a) + 2b$, give an upper bound on the VC dimension of $\mathcal{H}$. 5 marks