Assignment 2

Due 2014/10/22 23:59:59

Submit your work with the filename: "`ML_assignment-${number}_${your-student-id}_${your-name}`." Late submissions will not be accepted.

1. Suppose the input dimension is 2. Show that the VC dimension of the hypothesis class of axis-aligned rectangles that include positive instances inside and exclude negative instances outside is 4.

2. Read the proof of Hoeffding's inequality, and show that given a random variable Z and a real-valued function f with values $f(Z) \in [a, b]$, for any real number η, we have

$$E[e^{\eta f(Z)}] \leq \frac{b - E[f(Z)]}{b-a} e^{\eta a} + \frac{E[f(Z)] - a}{b-a} e^{\eta b}.$$

   (Hint: $\exp(\cdot)$ is a convex function)

3. Let $\mathcal{H}$ be a function class of VC dimension v. Based on Sauer's lemma, show that for $N > v$,

$$\mathcal{S}_{\mathcal{H}}(N) \leq \left(\frac{eN}{v}\right)^v.$$

   (Hint: recall that $e^v = \lim_{n\to\infty}(1 + \frac{v}{n})^n \geq (1 + \frac{v}{N})^N$ for any $0 \leq \frac{v}{N} < 1$)

4. Show that $Var[1(g^*(\boldsymbol{x}) \neq r)] \leq \frac{1}{4}$.
   (Hint: $1(g^*(\boldsymbol{x}) \neq r) \in \{0, 1\}$ so if $P(1(g^*(\boldsymbol{x}) \neq r)) = p$, then $P(1(g^*(\boldsymbol{x}) = r)) = 1 - p$)

5. Given three datasets and three hypothesis classes, plot their generalization errors and VC bounds (see coding assignment for details). Which dataset and hypothesis class combination results in the tightest VC bound?