COS 511: Theoretical Machine Learning

Homework #2 — Due:
Sample size bounds, growth function, VC dimension — February 27, 2014

Problem 1

[10] As on Problem 1 on Homework #1, let $X = \mathbb{R}$, and let $\mathcal{C}_s$ be the class of concepts defined by unions of s intervals. Compute the VC-dimension of $\mathcal{C}_s$ exactly.

Problem 2

[15] For $i = 1, \ldots, n$, let $\mathcal{G}_i$ be a space of concepts ($\{0,1\}$-valued functions) defined on some domain X, and let $\mathcal{F}$ be a space of concepts defined on $\{0,1\}^n$. (That is, each $g_i \in \mathcal{G}_i$ maps X to $\{0,1\}$, and each $f \in \mathcal{F}$ maps $\{0,1\}^n$ to $\{0,1\}$.) Let $\mathcal{H}$ be the space of all concepts $h : X \to \{0,1\}$ of the form

$$h(x) = f(g_1(x), \ldots, g_n(x))$$

for some $f \in \mathcal{F}$, $g_1 \in \mathcal{G}_1, \ldots, g_n \in \mathcal{G}_n$.

Give a careful argument proving that

$$\Pi_{\mathcal{H}}(m) \leq \Pi_{\mathcal{F}}(m) \cdot \prod_{i=1}^{n} \Pi_{\mathcal{G}_i}(m).$$

Problem 3

[15] Show that Sauer's Lemma is tight. That is, for each $d = 0, 1, 2, \ldots$, give an example of a class $\mathcal{C}$ with VC-dimension equal to d such that for each m,

$$\Pi_{\mathcal{C}}(m) = \sum_{i=0}^{d} \binom{m}{i}.$$

Problem 4

This problem explores another general method for bounding the error when the hypothesis space is infinite.

Some algorithms output hypotheses that can be represented by a small number of examples from the training set. For instance, suppose the domain is $\mathbb{R}$ and we are learning a half-line of the form $x \geq a$ where a defines the half-line. A simple algorithm chooses the left most positive training example a and outputs the corresponding half-line, which is clearly consistent with the data. Thus, in this case, the hypothesis can be represented by a single training example.

More formally, let F be a function mapping labeled examples to concepts, and assume that algorithm A, when given training examples $(x_1, c(x_1)), \ldots, (x_m, c(x_m))$ labeled by some unknown $c \in \mathcal{C}$, chooses some $i_1, \ldots, i_k \in \{1, \ldots, m\}$ and outputs the consistent hypothesis $h = F((x_{i_1}, c(x_{i_1})), \ldots, (x_{i_k}, c(x_{i_k})))$. In a sense, the algorithm has "compressed" the sample down to a sequence of just k of the m training examples. (We assume throughout that $m > k$.)

a. [5] Give such an algorithm for axis-aligned hyper-rectangles in $\mathbb{R}^n$ with $k = O(n)$. (An axis-aligned hyper-rectangle is a set of the form $[a_1, b_1] \times \cdots \times [a_n, b_n]$, and the corresponding concept, as usual, is the binary function that is 1 for points inside the rectangle and 0 otherwise. For $n = 2$, this is the class of rectangles used repeatedly as an example in class.) Your algorithm should run in time polynomial in m and n.

b. [15] Returning to the general case, assume as usual that the examples are chosen at random from some distribution D. Also assume that the size k is fixed. Argue carefully that the error of the output hypothesis h, with probability at least $1 - \delta$, satisfies the bound:

$$\mathrm{err}_D(h) \leq O\left(\frac{\ln(1/\delta) + k \ln m}{m - k}\right).$$