Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2
Due: February 21, 2006

Problem 1: VC dimension

(1) Show that the VC dimension of the class C of halfspaces over $\mathbb{R}^n$ is $n+1$. To do that, proceed as follows.

   (a) Show that $\text{VCdim}(C) \geq n+1$.

   (b) Use Radon's theorem: *Any set of $n+2$ points $X \subset \mathbb{R}^n$ can be partitioned in two subsets X_1 and X_2 such that the convex hulls of X_1 and X_2 intersect.*
   to show that $\text{VCdim}(C) \leq n+1$.

   (c) Prove Radon's theorem.

(2) Consider now the class C_k of convex intersections of k halfspaces. Give lower and upper bounds estimates for $\text{VCdim}(C_k)$.

(3) Let A and B be two sets of functions mapping from X into $\{0,1\}$, and assume that both A and B have finite VC dimension, with $\text{VCdim}(A) = d_A$ and $\text{VCdim}(B) = d_B$. Let $C = A \cup B$ be the union of A and B.

   (a) Prove that for all m, $\Pi_C(m) \leq \Pi_A(m) + \Pi_B(m)$.

   (b) Use Sauer's lemma to show that for $m \geq d_A + d_B + 2$, $\Pi_C(m) < 2^m$, and give a bound on the VC dimension of C.

Problem 2: Sample complexity

A function $h : \{0,1\}^n \to \{0,1\}$ is *symmetric* if its value is uniquely determined by the number of 1's in the input. Let C denote the set of all symmetric functions.

(a) Determine the VC dimension of C.

(b) Give lower and upper bounds on the sample complexity of any consistent PAC learning algorithm for C.

(c) Note that any hypothesis $h \in C$ can be represented by a vector $(y_0, y_1, ..., y_n) \in \{0, 1\}^{n+1}$, where y_i is the value of h on examples having precisely i 1's. Devise a consistent learning algorithm for C based on that representation.