Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2
Due: Mar 7th, 2008
Credit: Ashish Rastogi, Afshin Rostamizadeh
Ameet Talwalkar, and Eugene Weinstein.

1. Show that a finite concept class $\mathcal{C}$ has VC dimension at most $\log |\mathcal{C}|$.

2. Determine the VC dimension of the following concept classes:

   (a) The class of all polygons with k vertices in the plane.

   (b) The class of all circles in the plane.

   (c) The class of union of k intervals on the real line.

3. [VC Dimension] Let F be a finite-dimensional vector space of real functions on $\mathbb{R}^n$, $\dim(F) = r < \infty$. Let H be the set of hypotheses:

$$H = \{\{x : f(x) \geq 0\} : f \in F\}.$$

   Show that d, the VC dimension of H, is finite and that $d \leq r$ [*Hint:* select an arbitrary set of $m = r + 1$ points and consider the linear mapping $u : F \mapsto \mathbb{R}^m$ defined by: $u(f) = (f(x_1), \ldots, f(x_m))$.]

4. [Regression] Consider the problem of learning a real valued function $h : \mathbb{R}^n \mapsto \mathbb{R}$ based on a training sample $S = \{(x_i, y_i), 1 \leq i \leq m\}$, $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. Consider $h(x) = w \cdot x$, where the weight vector $w \in \mathbb{R}^n$ is determined according to the solution of the following optimization problem:

$$\min_{w \in \mathbb{R}^n} \frac{1}{2}\|w\|_2^2 + \gamma \sum_{i=1}^{m} (w \cdot x_i - y_i)^2.$$

   Let X be a $m \times n$ matrix where $\mathsf{X}_{i,j} = (x_i)_j$ and let Y be an m-dimensional column vector whose ith coordinate is y_i. Finally let W denote the n-dimensional column vector corresponding to the weight vector w.

   (a) Express the objective function above in terms of matrices X and the vectors W, Y, together with the tradeoff constant γ.

(b) Determine the closed-form solution for the optimal weight vector W^* in terms of $\mathsf{X}, \mathsf{Y}, \gamma$ (let I denote the identity matrix). [Hint: you may use $\frac{\partial \|\mathsf{A}\|_2^2}{\partial \mathsf{A}} = 2\mathsf{A}$ for a matrix A].

(c) What is the time complexity of computing the optimal weight vector W as a function of the number of features n and the number of training points m. What is the complexity of computing $h(x)$ for a new point $x \in \mathbb{R}^n$?

(d) The matrix $\mathsf{X}\mathsf{X}^T$ is called the Gram matrix K. Using the observation that

$$\mathsf{X}^T \left(K + \gamma \mathsf{I}\right)^{-1} = \left(\mathsf{X}^T\mathsf{X} + \gamma \mathsf{I}\right)^{-1} \mathsf{X}^T,$$

derive another expression for the optimal weight vector W. What is the complexity of computing using this alternate closed-form expression?