Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2 - solution
Credit: Ashish Rastogi, Afshin Rostamizadeh
Ameet Talwalkar, and Eugene Weinstein.

1. [10 points] Show that a finite concept class $\mathcal{C}$ has VC dimension at most $\log |\mathcal{C}|$.

   Solution: Proof by contraposition. Suppose that the VC dimension is $d > \log |\mathcal{C}|$. There must exist a set of d points such that every dichotomy on those d points is realized. The number of possible dichotomies is $2^d > 2^{\log |\mathcal{C}|} = |\mathcal{C}|$. But there are only $|\mathcal{C}|$ distinct concept classes.

2. [30 points] Determine the VC dimension of the following concept classes:

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

   Solution: In order to show that the VC-dimension of a class of concepts is d, we need to show that there exists a set of size d on which all dichotomies are realized, and that on all sets of size $d + 1$, there is some dichotomy that is not realized by the concept class. We know from class that the class of convex polygons with k vertices has VC dimension of $2k + 1$. Thus, we know that the VC dimension of our class is at least $2k + 1$. It can be shown that for all sets of size $2k + 2$ points, there is a labeling of these points that cannot be captured by (even) non-convex polygons with k vertices.

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

   Solution: It is clear that any two points can be shattered by a circle. Any three non-colinear points can also be shattered. Now, given any four points, there are two cases. The first, that the convex hull of these four points is a triangle. If so, labelling the points on the triangle as positive and the point inside as negative is a dichotomy that cannot be realized by a circle. If the convex hull of the four points is a quadrilateral, then choosing the further of the two diagonally opposite points as positive and the other two as negative is a dichotomy that cannot be realizeda lso. Finally, if the four points are colinear, there is a trivial dichotomy of alternate positive and negatives that cannot be realized. Thus, VC dimension of all circles in the plane is 3.

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

   Solution: Easy to check that a sequence of $2k + 1$ points on a line cannot be shattered, if successive points are labeled with alternate labels, starting with a positive label. Thus, VC dimension of the class of union of k intervals on the real line is $2k$.

3. [VC Dimension: 20 points] 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: $u(f) = (f(x_1), \ldots, f(x_m))$.]

Solution: Show that no set of size $m = r+1$ can be shattered by H. Let $x_1, \ldots, x_m$ be m arbitrary points. Define the linear mapping $l : F \mapsto \mathbb{R}^m$ defined by:

$$l(f) = (f(x_1), \ldots, f(x_m)).$$

Since the dimesion of $\dim(F) = m-1$, the rank of l is at most $m-1$ and there exists $\alpha \in \mathbb{R}^m$ orthogonal to $l(F)$:

$$\forall f \in F, \sum_{i=1}^{m} \alpha_i f(x_i) = 0.$$

We can assume that at least one α_i is negative. Then,

$$\forall f \in F, \sum_{i:\alpha_i \geq 0}^{m} \alpha_i f(x_i) = -\sum_{i:\alpha_i < 0}^{m} \alpha_i f(x_i).$$

Now, assume that there exists a set $\{x : f(x) \geq 0\}$ selecting exactly the x_is on the left-hand side. Then all the terms on the left-hand side are non-negative, while those on the right-hand side are negative, which cannot be. Thus, $\{x_1, \ldots, x_m\}$ cannot be shattered.

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) [10 points] Express the objective function above in terms of matrices X and the vectors W, Y, together with the tradeoff constant γ.

Solution:

$$F = \frac{1}{2}\mathsf{W}^T\mathsf{W} + \gamma \left((\mathsf{XW} - \mathsf{Y})^T(\mathsf{XW} - \mathsf{Y})\right).$$

(b) [20 points] 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].

Solution:

$$\mathsf{W}^* = 2\gamma \left[\mathsf{I} + 2\gamma \mathsf{X}^T \mathsf{X}\right]^{-1} \mathsf{X}^T \mathsf{Y}.$$

(c) [10 points] 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 $x \in \mathbb{R}^n$?

Solution: The complexity of multiplying two matrices A, B of dimensions $a \times b, b \times c$ respectively is $O(abc)$. The complexity of inverting a square matrix A of dimension $a \times a$ is $O(a^3)$ (faster algorithms exist, for e.g. Strassen's algorithm) but the more well-known $O(a^3)$ is accepted as a reference for this problem.

Based on these observations, the complexity of computing $\mathsf{I} + 2\gamma \mathsf{X}^T \mathsf{X}$ is $O(n^2 m)$. The complexity of inverting this matrix is $O(n^3)$. The complexity of multiplying the resulting $n \times n$ matrix with $\mathsf{X}^T \mathsf{Y}$ is $O(n^2 m)$. Thus, the overall complexity is $O(n^2(m+n))$.

(d) [10 points] The matrix $\mathsf{X}\mathsf{X}^T$ is called the Gram matrix K. Using the observation that

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

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

Solution: The optimal hyperplane W^* in the *dual* is given by:

$$\mathsf{W}^* = 2\gamma \mathsf{X}^T \left[\mathsf{I} + 2\gamma \mathsf{K}\right]^{-1} \mathsf{Y}.$$

The complexity of obtaining this solution is $O(m^2(m+n))$. Using this solution is more efficient when the number of sample points m is far smaller than the number of features n.