Mehryar Mohri
Foundations of Machine Learning 2023
Courant Institute of Mathematical Sciences
Homework assignment 2
October 6, 2023
Due: October 31, 2023

A. Radmacher complexity

1. Consider the class of functions $\mathcal{H}$ mapping from $\mathbb{R}$ to $\{+1, -1\}$ such that

$$h(x) = \begin{cases} +1 \text{ for } x \in [a, b], \\ -1 \text{ otherwise }, \end{cases}$$

for some $a, b \in \mathbb{R}$. Use Sauer's lemma to give an upper bound on the growth function $\Pi_{\mathcal{H}}(m)$ and prove that the upper bound is tight in this example. Use it to derive an upper bound on $\mathfrak{R}_m(\mathcal{H})$.

Solution: The VC-dimension of the hypothesis class of intervals on the real line is 2. Therefore, by Sauer's lemma, the following inequality holds:

$$\Pi_{\mathcal{H}}(m) \le \binom{m}{0} + \binom{m}{1} + \binom{m}{2}.$$

The above is actually an equality since we can compute the growth function as follows:

$$\Pi_{\mathcal{H}}(m) = \binom{m+1}{2} + 1 = \frac{1}{2}m^2 + \frac{1}{2}m + 1.$$

The Rademacher complexity can be bounded in terms of the growth function as follows:

$$\mathfrak{R}_m(\mathcal{H}) \le \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} = \sqrt{\frac{2 \log\left(\frac{1}{2}m^2 + \frac{1}{2}m + 1\right)}{m}}.$$

2. Prove that for any $\alpha, \beta \in \mathbb{R}$ and any two hypothesis sets $\mathcal{H}_1$ and $\mathcal{H}_2$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, the equality $\mathfrak{R}_m(\alpha\mathcal{H}_1 + \beta\mathcal{H}_2) = |\alpha|\mathfrak{R}_m(\mathcal{H}_1) + |\beta|\mathfrak{R}_m(\mathcal{H}_2)$ holds, where the linear combination of the two hypothesis sets are defined by $\alpha\mathcal{H}_1 + \beta\mathcal{H}_2 = \{\alpha h_1 + \beta h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$.

Solution: Expand the definition of empirical Radmacher complexity.

3. Prove that if for two hypothesis sets $\mathcal{H}_1$ and $\mathcal{H}_2$ the inclusion $\mathcal{H}_1 \subseteq \mathcal{H}_2$ holds, then the following inequality holds for any finite sample S: $\widehat{\mathfrak{R}}_S(\mathcal{H}_1) \le \widehat{\mathfrak{R}}_S(\mathcal{H}_2)$.

Solution: Definition of Radmacher complexity and supremum over $\mathcal{H}_1$ is upper bounded by supremum over $\mathcal{H}_2$.

4. Let $\mathcal{H}_1$ be a family of functions mapping from $\mathcal{X}$ to $\{0, 1\}$ and let $\mathcal{H}_2$ be a family of functions mapping from $\mathcal{X}$ to $\{-1, +1\}$. Let $\mathcal{H} = \{h_1 h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Show that the empirical Rademacher complexity of $\mathcal{H}$ for any sample S of size m can be bounded as follows:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \le \widehat{\mathfrak{R}}_S(\mathcal{H}_1) + \widehat{\mathfrak{R}}_S(\mathcal{H}_2).$$

[hint: write $h_1 h_2$ in a way such that you can apply Talagrand's lemma.]

Solution: Consider $\phi(x) = |x| - 1$. Then, one can verify that $h_1 h_2$ can be written as

$$h_1 h_2 = \phi(h_1 + h_2).$$

As ϕ is a 1-Lipschitz function, by Talagrand's Contraction Lemma, we have

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{H}_1 + \mathcal{H}_2) = \widehat{\mathfrak{R}}_S(\mathcal{H}_1) + \widehat{\mathfrak{R}}_S(\mathcal{H}_2).$$

B. VC-dimension

1. What is the VC-dimension of axis-aligned squares in $\mathbb{R}^2$? Is this value the same as the VC-dimension of squares (not necessarily axis-aligned) in $\mathbb{R}^2$? Why?

   Solution: 3. First we prove that there exists a 3-point set such that it can be fully shattered by axis-aligned squares. For example, suppose 3 points are vertices of an isosceles right triangle. It is easy to see that they can be fully shattered. We also need to prove no 4-points set could be fully shattered by axis-aligned squares. It is easy to see when 3 points are collinear, they can not be fully shattered (for example + - +). Suppose no 3 points are collinear and mark the 4 points clockwise as A,B,C,D. Assume $|AC| > |BD|$ and we can not generate both A+, B-, C+, D- and A-, B+, C-, D+.

   The VC-dimension of squares (not necessarily axis-aligned) in $\mathbb{R}^2$ is larger, since there exists a 4-point set that can be fully shattered by squares.

2. What is the VC-dimension of intersections of 2 axis-aligned squares in $\mathbb{R}^2$?

   Solution: 4. Same as axis-aligned rectangles.

3. (a) For two concept classes $\mathcal{C}_1, \mathcal{C}_2$, define the concept class $\mathcal{C}$ by

   $$\mathcal{C} = \{c_1 c_2 \mid c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2\}.$$

   Prove that the following inequality holds:

   $$\Pi_{\mathcal{C}}(m) \leq \Pi_{\mathcal{C}_1}(m)\Pi_{\mathcal{C}_2}(m).$$

   Solution: For any set $\{x_1, \ldots, x_m\} \subset \mathcal{X}$, it is straightforward to see that the following inequalities hold:

   $$\begin{aligned} &|\{(c_1(x_1)c_2(x_1), ..., c_1(x_m)c_2(x_m)) \mid c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2\}| \\ &\leq |\{(c_1(x_1), ..., c_1(x_m)) \mid c_1 \in \mathcal{C}_1\}||\{(c_2(x_1), ..., c_2(x_m)) \mid c_2 \in \mathcal{C}_2\}| \\ &\leq \Pi_{\mathcal{C}_1}(m)\Pi_{\mathcal{C}_2}(m). \end{aligned}$$

   Taking max on the left hand side we close the proof.

   (b) Let $\mathcal{C}$ be a concept class whose VC-dimension is 3. Show that the VC-dimension of intersections of k concepts from $\mathcal{C}$ is upper bounded by $6k\log_2(3k)$. [hint: use Sauer's lemma and the result of (a).]

   Solution: We denote $\mathcal{C}^k$ as the set of intersections of k concepts from $\mathcal{C}$. Then by the previous question, we have $\Pi_{\mathcal{C}^k}(m) \leq (\Pi_{\mathcal{C}}(m))^k$ for any $m \in \mathbb{N}$. We only need to prove that $(\Pi_{\mathcal{C}}(m))^k < 2^m$ for $m = 6k\log_2(3k)$. By Sauer's lemma and the fact that $\text{VCdim}(\mathcal{C}) = 3$ we get $\Pi_{\mathcal{C}}(m) \leq (\frac{em}{3})^3$. Thus $(\Pi_{\mathcal{C}}(m))^k \leq (\frac{em}{3})^{3k}$. We substitute m by $6k\log_2(3k)$ then the inequality turns out to be $2e\log_2(3k) < 9k$, which is trivially true.

C. Support Vector Machines

1. (a) SVMs are "sparse" in the sense that the number of support vectors is usually small compared to total number of observations. Suppose we explicitly maximize sparsity by penalizing the L_2 norm of the vector $\boldsymbol{\alpha}$ that defines the weight vector $\mathbf{w}$:

$$\begin{aligned}
\min_{\boldsymbol{\alpha},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{\alpha}\|^2 + C\left(\sum_{i=1}^m \xi_i\right) \qquad (1)\\
\text{subject to} \quad & y_i\left(\left(\sum_{j=1}^m \alpha_j y_j \mathbf{x}_j\right)\cdot \mathbf{x}_i + b\right) \geq 1-\xi_i,\\
& \xi_i \geq 0, \alpha_i \geq 0, i \in [m].
\end{aligned}$$

Show that the problem coincides with an instance of the primal optimization problem of SVMs, modulo the non-negativity constraint on $\boldsymbol{\alpha}$. You should indicate exactly how to view it as such.

Solution: Let

$$\mathbf{x}'_i = \Big(y_1(\mathbf{x}_1\cdot\mathbf{x}_i),\ldots,y_m(\mathbf{x}_m\cdot\mathbf{x}_i)\Big).$$

Then the optimization problem becomes

$$\begin{aligned}
\min_{\boldsymbol{\alpha},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{\alpha}\|^2 + C\left(\sum_{i=1}^m \xi_i\right)\\
\text{subject to} \quad & y_i\left(\boldsymbol{\alpha}\cdot\mathbf{x}'_i + b\right) \geq 1-\xi_i,\\
& \xi_i \geq 0, \alpha_i \geq 0, i \in [m].
\end{aligned}$$

This is the standard formulation of the primal SVM optimization problem on samples $(\mathbf{x}'_1, y_1),\ldots,(\mathbf{x}'_m, y_m)$, modulo the non-negativity constraints on α_i.

(b) Derive the dual optimization problem of (1).

Solution: Define Lagrange variables $p_i \geq 0, q_i \geq 0, r_i \geq 0$. The Lagrangian is

$$\begin{aligned}
L(\boldsymbol{\alpha},b,\boldsymbol{\xi},p,q,r) = & \frac{1}{2}\|\boldsymbol{\alpha}\|^2 + C\left(\sum_{i=1}^m \xi_i\right)\\
& - \sum_{i=1}^m p_i\left\{y_i\left[\left(\sum_{j=1}^m \alpha_j y_j\mathbf{x}_j\right)\cdot\mathbf{x}_i + b\right] - 1 + \xi_i\right\}\\
& - \sum_{i=1}^m q_i\xi_i - \sum_{i=1}^m r_i\alpha_i.
\end{aligned}$$

Note that

$$\begin{aligned}
& \sum_{i=1}^m p_i\left\{y_i\left[\left(\sum_{j=1}^m \alpha_j y_j\mathbf{x}_j\right)\cdot\mathbf{x}_i + b\right] - 1 + \xi_i\right\}\\
&= \left(\sum_{i=1}^m p_i y_i\mathbf{x}_i\right)\cdot\left(\sum_{i=1}^m \alpha_i y_i\mathbf{x}_i\right) + \sum_{i=1}^m p_i y_i b - \sum_{i=1}^m p_i + \sum_{i=1}^m p_i\xi_i.
\end{aligned}$$

Set the gradient of the Lagrangian with respect to the primal variables to zero:

$$\begin{aligned}
\nabla_{\alpha_i} L &= \alpha_i - y_i\mathbf{x}_i\cdot\left(\sum_{j=1}^m p_j y_j\mathbf{x}_j\right) - r_i = 0 && \Rightarrow \alpha_i = y_i\mathbf{x}_i\cdot\left(\sum_{j=1}^m p_j y_j\mathbf{x}_j\right) + r_i\\
\nabla_b L &= -\sum_{i=1}^m p_i y_i = 0 && \Rightarrow \sum_{i=1}^m p_i y_i = 0\\
\nabla_{\xi_i} L &= C - p_i - q_i = 0 && \Rightarrow p_i + q_i = C
\end{aligned}$$

Plugging in the expression of $\boldsymbol{\alpha}$ in L gives

$$
\begin{aligned}
&L(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, p, q, r) \\
&= \frac{1}{2}\|\boldsymbol{\alpha}\|^2 + C\left(\sum_{i=1}^m \xi_i\right) - \sum_{i=1}^m \alpha_i(\alpha_i - r_i) \\
&- \sum_{i=1}^m p_i y_i b + \sum_{i=1}^m p_i - \sum_{i=1}^m (p_i + q_i)\xi_i - \sum_{i=1}^m r_i \alpha_i \\
&= \frac{1}{2}\|\boldsymbol{\alpha}\|^2 - \sum_{i=1}^m \alpha_i^2 + \sum_{i=1}^m p_i \\
&= -\frac{1}{2}\|\boldsymbol{\alpha}\|^2 + \sum_{i=1}^m p_i \\
&= -\frac{1}{2}\|\sum_{i=1}^m p_i y_i \mathbf{x}_i' + r\|^2 + \sum_{i=1}^m p_i.
\end{aligned}
$$

Putting everything together, the dual optimization problem is

$$
\begin{aligned}
\max_{p,r} \quad & \sum_{i=1}^m p_i - \frac{1}{2}\|\sum_{i=1}^m p_i y_i \mathbf{x}_i' + r\|^2 \\
\text{subject to} \quad & 0 \le p_i \le C \wedge r_i \ge 0 \wedge \sum_{i=1}^m p_i y_i = 0, i \in [m].
\end{aligned}
$$

2. Suppose we replace in the primal optimization problem of SVMs the penalty term $\sum_{i=1}^m \xi_i = \|\boldsymbol{\xi}\|_1$ with $\|\boldsymbol{\xi}\|_\infty = \max_{i=1}^m \xi_i$. Give the associated dual optimization problem. Show that it differs from the standard dual optimization problem of SVMs only by the constraints, which can be expressed in terms of $\|\boldsymbol{\alpha}\|_1$.

Solution: The optimization problem for this version of SVMs can be written as follows:

$$
\begin{aligned}
\min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \\
\text{subject to} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \ge 1 - \xi \forall i \in [m] \\
& \xi \ge 0.
\end{aligned} \tag{2}
$$

The corresponding Lagrange function can be written as

$$
L = \frac{1}{2}\|\mathbf{w}\|^2 + C\xi - \sum_{i=1}^m \alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi] - \beta\xi.
$$

Differentiating with respect to the primal variables gives:

$$
\begin{aligned}
&\nabla_{\mathbf{w}} L = 0 \implies \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \\
&\nabla_b L = 0 \implies \sum_{i=1}^m \alpha_i y_i = 0 \\
&\nabla_\xi L = 0 \implies \sum_{i=1}^m \alpha_i + \beta = C.
\end{aligned}
$$

Plugging in the first equality in L and using the second and third yields:

$$
L = \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j).
$$

In view of the third equality, the condition $\beta \geq 0$ can be equivalently written as $\sum_{i=1}^{m} \alpha_i \leq C$. Thus, the equivalent dual optimization problem can be written as

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to} \quad (\boldsymbol{\alpha} \geq 0) \wedge (\|\boldsymbol{\alpha}\|_1 \leq C) \wedge (\sum_{i=1}^{m} \alpha_i y_i = 0).$$

More generally, a $\|\cdot\|_p$-constraint on $\boldsymbol{\xi}$ in the primal optimization problem leads to a $\|\cdot\|_q$-constraint (dual norm constraint) on $\boldsymbol{\alpha}$ in the dual, where p and q are conjugate: $1/p + 1/q = 1$.