Mehryar Mohri
Foundations of Machine Learning 2018
Courant Institute of Mathematical Sciences
Homework assignment 2
October 07, 2018
Due: October 21, 2018

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 $a, b \in \mathbb{R}$. Give an upper bound on the growth function $\Pi_{\mathcal{H}}(m)$ and use it to derive an upper bound on the $\mathfrak{R}_m(\mathcal{H})$.

Solution:

$$\Pi_H(m) = \binom{n+1}{2} + 1 = \frac{1}{2}n^2 + \frac{1}{2}n + 1$$

2. Prove that for any hypotheses class $\mathcal{H}$ and any function $h\colon \mathcal{X} \mapsto \mathbb{R}$, $\mathfrak{R}_m(\mathcal{H}) = \mathfrak{R}_m(\mathcal{H} + h)$.

Solution: Expand the definition of empirical Radmacher complexity.

3. Prove that if for two hypotheses classes $\mathcal{H}$ and $\mathcal{F}$ the inclusion $\mathcal{H} \subseteq \mathcal{F}$ holds, then the following inequality holds for any finite sample S: $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{F})$.

Solution: Definition of Radmacher complexity and supremum over $\mathcal{H}$ is at least as supremum over $\mathcal{F}$.

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

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

(*hint*: write $h_1 h_2$ in a way such that you can apply Talagrand's inequality.)

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

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

As ϕ is an 1-Lipschitz function, by Talagrands Contraction Lemma,

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

B. VC-dimension

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

   Solution: 3. First we prove 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 this outcome: A+, B-, C+, D-.

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. (Bonus) Let C be a concept class whose VC-dimension is 3. Show that the VC-dimension of intersections of k concepts from C is upper bounded by $6k \log_2(3k)$. (*hint*: use Sauer's lemma.)

   Solution: First we prove the following key property

   Lemma 1. *For two concept sets C_1, C_2 denote*

   $$C = \{c_1 c_2 \mid c_1 \in C_1, c_2 \in C_2\}$$

   we have $\Pi_C(m) \leq \Pi_{C_1}(m)\Pi_{C_2}(m)$.

Proof. For any set $\{x_1, ...x_m\} \subset \mathcal{X}$, it is easy to see that

$$\begin{aligned} &|\{(c_1(x_1)c_2(x_1), ..., c_1(x_m)c_2(x_m)) \mid c_1 \in C_1, c_2 \in C_2\}| \\ \leq &|\{(c_1(x_1), ..., c_1(x_m)) \mid c_1 \in C_1\}||\{(c_2(x_1), ..., c_2(x_m)) \mid c_2 \in C_2\}| \\ &\leq \Pi_{C_1}(m)\Pi_{C_2}(m) \end{aligned}$$

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

Back to this problem we denote C^k as the set of intersections of k concepts from C. Then by above lemma $\Pi_{C^k}(m) \leq (\Pi_C(m))^k$ for any $m \in \mathbb{N}$. We only need to prove that $(\Pi_C(m))^k < 2^m$ for $m = 6k\log_2(3k)$. By Sauer's lemma and the fact that $VC - dim(C) = 3$ we get $\Pi_C(m) \leq (\frac{em}{3})^3$. Thus $(\Pi_C(m))^k \leq (\frac{em}{3})^{3k}$. 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. Download and install the libsvm software library from:

   `https://www.csie.ntu.edu.tw/~cjlin/libsvm`

   and briefly consult the documentation to become more familiar with the tools.

2. Consider the `svmguide1` dataset

   `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/svmguide1`

   Download a shuffled version of that dataset from

   `http://www.cs.nyu.edu/~mohri/ml18/svmguide1.shuffled`

   Use the `libsvm` scaling tool to scale the features of all the data. Use the first 2316 examples for training, the last 773 for testing. The scaling parameters should be computed only on the training data and then applied to the test data.

3. Consider the binary classification task in `svmguide1`, using the 4 features. Use SVMs combined with polynomial kernels to tackle this binary classification problem.

   To do that, randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, $d = 1, 2, 3, 4$, plot the average cross-validation error plus or minus one standard deviation as a function of C (let other parameters of polynomial kernels in `libsvm` be equal to their default values), varying C in powers of 2, starting from a small value $C = 2^{-k}$ to $C = 2^k$, for some value of k. k should be chosen so that you see a significant variation in training error, starting from a very high training error to a low training error. Expect longer training times with `libsvm` as the value of C increases.

   Solution: Figure 1 shows the average cross-validation performance as a function of the regularization parameter C from 2^{-6} to 2^6. The performance for several choices of d and C are essentially indistinguishable; one suitable choice of optimal parameters is $C^* = 2^5$ and $d^* = 4$.

4. Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the following results as a function of d:

   (a) The average ten-fold cross-validation error, and the test error for the hypotheses obtained by running SVMs on the whole training set.

   (b) The average number of support vectors, and the average number of support vectors lie on the margin hyperplanes.

   Solution: The first plot in Figure 2 shows that the test error first decreases and then increases, as a function of increasing degree d. Also, the cross-validation error is (slightly) optimistic when compared to the test error on the held-out dataset, for larger values of d. The second plot shows that the total number of marginal support vectors slightly increases with d, while the total number of overall support vectors decreases first and then increases.

5. 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

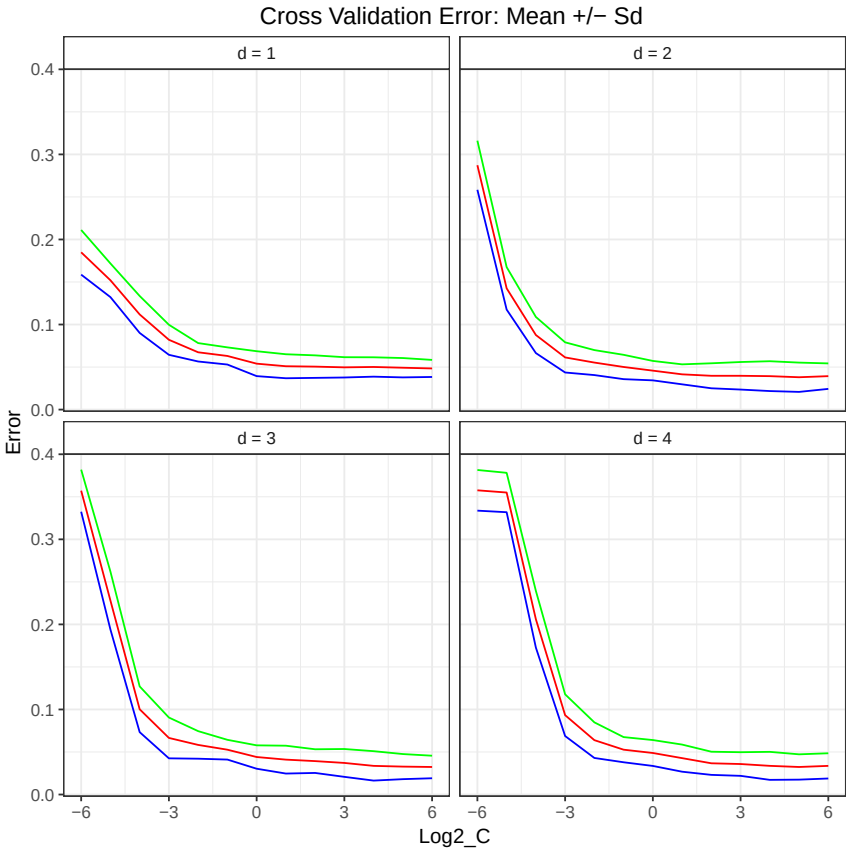

Figure 1: Average error according to 10-fold cross-validation, with error-bars indicating one standard deviation.

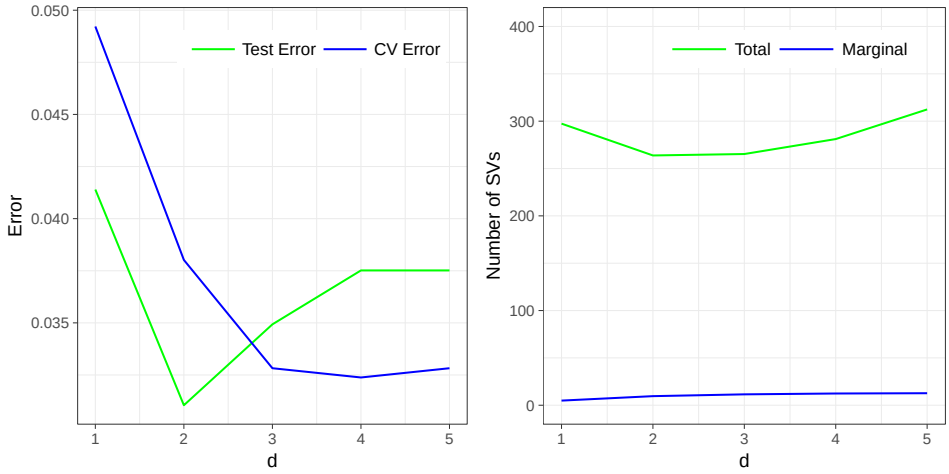

Figure 2: The test and validation error as a function of degree d (left panel) as well as the number of total and marginal support vectors (right panel).

α that defines the weight vector w:

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

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

Solution: Let

$$x_i' = \Big(y_1(x_1 \cdot x_i), \ldots, y_m(x_m \cdot x_i) \Big).$$

Then the optimization problem becomes

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

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

Another way is to show that the dual forms are equivalent. Define Lagrange variables $p_i \geq 0, q_i \geq 0, r_i \geq 0$. The Lagrangian is

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

Note that

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

To meet KKT conditions,

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

And

$$\begin{aligned} & \forall i \in [m], \quad \sum_{i=1}^{m} p_i \left\{ y_i \left[\left(\sum_{j=1}^{m} \alpha_j y_j x_j \right) \cdot x_i + b \right] - 1 + \xi_i \right\} = 0, \\ & \forall i \in [m], \quad q_i \xi_i = 0, \, r_i \alpha_i = 0. \end{aligned}$$

Plugging in the expression of α in L gives

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

Note that

$$r_i \alpha_i = 0 \Rightarrow y_i x_i r_i \cdot \left(\sum_{j=1}^{m} p_j y_j x_j\right) + r_i^2 = 0.$$

Thus,

$$\begin{aligned}
&\parallel \alpha \parallel^2 \\
&= \sum_{i=1}^{m} \left[y_i x_i \cdot \left(\sum_{j=1}^{m} p_j y_j x_j\right) \right]^2 - \sum_{i=1}^{m} r_i^2 \\
&= \sum_{i=1}^{m} \left(\sum_{j,k=1}^{m} p_j y_j p_k y_k (x_i y_i \cdot x_j)(x_i y_i \cdot x_k) \right) - \sum_{i=1}^{m} r_i^2 \\
&= \sum_{j,k=1}^{m} p_j y_j p_k y_k \left(\sum_{i=1}^{m} (x_i y_i \cdot x_j)(x_i y_i \cdot x_k) \right) - \sum_{i=1}^{m} r_i^2 \\
&= \sum_{j,k=1}^{m} p_j p_k y_j y_k K(x_j, x_k) - \sum_{i=1}^{m} r_i^2,
\end{aligned}$$

where $K(x_j, x_k) = \sum_{i=1}^{m} (x_i y_i \cdot x_j)(x_i y_i \cdot x_k)$. Putting everything to-

gether, the dual optimization problem is

$$\max_{p,r} \quad \sum_{i=1}^{m} p_i - \frac{1}{2} \sum_{i,j=1}^{m} p_i p_j y_i y_j K(x_i, x_j) + \frac{1}{2} \sum_{i=1}^{m} r_i^2$$

$$\text{subject to} \quad 0 \leq p_i \leq C \wedge r_i \geq 0 \wedge \sum_{i=1}^{m} p_i y_i = 0, i \in [m].$$

One can show that this dual form coincides with the dual form of the standard SVM primal problem with non-negative constraints on w, using kernel $K(\cdot, \cdot)$ instead of inner product. Note that if we let $x'_i = \Big(y_1(x_1 \cdot x_i), \ldots, y_m(x_m \cdot x_i)\Big)$ as defined earlier, then $K(x_j, x_k) = (x'_j \cdot x'_k)$.