Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2
October 04, 2016
Due: October 18, 2016

A. Rademacher complexity

The definitions and notation are those introduced in the lectures slides.

1. What is the Rademacher complexity of a hypothesis set reduced to a single hypothesis? An alternative definition of the Rademacher is based on absolute values: $\mathfrak{R}'(H) = \frac{1}{m} \mathrm{E}_{\boldsymbol{\sigma},S}[\sup_{h \in H} |\sum_{i=1}^{m} \sigma_i h(x_i)|]$. Show the following upper bound for a hypothesis set reduced to a single hypothesis h:

$$\mathfrak{R}'(\{h\}) \leq \sqrt{\frac{\mathrm{E}_{x \sim D}[h^2(x)]}{m}}. \tag{1}$$

(*hint*: you can use the inequality $\mathrm{E}[|X|] \leq \sqrt{\mathrm{E}[X^2]}$ which is valid for any random variable X, by Jensen's inequality).

Solution: Let h be that single hypothesis. By definition,

$$\mathfrak{R}(\{h\}) = \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S} \left[\sum_{i=1}^{m} \sigma_i h(x_i) \right] = \frac{1}{m} \mathop{\mathrm{E}}_{S} \left[\sum_{i=1}^{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}[\sigma_i]\, h(x_i) \right] = 0,$$

since $\mathrm{E}_{\boldsymbol{\sigma}}[\sigma_i] = 0$ for all $i \in [1, m]$. Using Jensen's inequality, with the

alternative definition the Rademacher complexity can be bounded as follows:

$$\begin{aligned}
\mathfrak{R}'(\{h\}) &= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\left|\sum_{i=1}^{m} \sigma_i h(x_i)\right|\right] \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sqrt{\left|\sum_{i=1}^{m} \sigma_i h(x_i)\right|^2}\right] \\
&\leq \frac{1}{m} \sqrt{\mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\left|\sum_{i=1}^{m} \sigma_i h(x_i)\right|^2\right]} && \text{(by Jensen's inequality)} \\
&= \frac{1}{m} \sqrt{\mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sum_{i,j=1}^{m} \sigma_i \sigma_j h(x_i) h(x_j)\right]} \\
&= \frac{1}{m} \sqrt{\mathop{\mathrm{E}}_{S}\left[\sum_{i=1}^{m} h(x_i)^2\right]} && (\mathrm{E}[\sigma_i \sigma_j = 0] \text{ for } i \neq j) \\
&= \frac{1}{m} \sqrt{m \mathop{\mathrm{E}}_{S}\left[h(x_1)^2\right]} = \sqrt{\frac{\mathrm{E}_x[h^2(x)]}{m}}. && \text{(i.i.d. sample)}
\end{aligned}$$

□

2. Fix $m \geq 1$. Prove the following identities for any $\alpha \in \mathbb{R}$ and any two hypothesis sets H and H' of functions mapping from $\mathcal{X}$ to $\mathbb{R}$:

$$\mathfrak{R}_m(\alpha H) = |\alpha| \mathfrak{R}_m(H) \tag{2}$$

$$\mathfrak{R}_m(H + H') \leq \mathfrak{R}_m(H) + \mathfrak{R}_m(H'). \tag{3}$$

$$\mathfrak{R}_m(\{\max(h, h') \colon h \in H, h' \in H'\}) \leq \mathfrak{R}_m(H) + \mathfrak{R}_m(H'), \tag{4}$$

where $\max(h, h')$ denotes the function $x \mapsto \max_{x \in \mathcal{X}}(h(x), h'(x))$ (***hint***: you could use the identity $\max(a, b) = \frac{1}{2}[a + b + |a - b|]$ valid for all $a, b \in \mathbb{R}$ and the contraction lemma (Lecture 4)).

Solution: If $\alpha \geq 0$, then

$$\sup_{h \in \alpha H} \sum_{i=1}^{m} \sigma_i h(x_i) = \sup_{h \in H} \sum_{i=1}^{m} \alpha\, \sigma_i h(x_i) = \alpha \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i),$$

otherwise if $\alpha < 0$, then

$$\sup_{h \in \alpha H} \sum_{i=1}^{m} \sigma_i h(x_i) = \sup_{h \in \alpha H} \sum_{i=1}^{m} \alpha\, \sigma_i h(x_i) = (-\alpha) \sup_{h \in H} \sum_{i=1}^{m} (-\sigma_i) h(x_i).$$

Since σ_i and $-\sigma_i$ have the same distribution, this shows that $\mathfrak{R}_m(\alpha H) = |\alpha|\mathfrak{R}_m(H)$.

The second inequality actually holds with equality, and follows from:

$$\begin{aligned}
&\mathfrak{R}_m(H + H') \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i(h(x_i) + h'(x_i))\right] \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i h(x_i) + \sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i h'(x_i)\right] \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H} \sum_{i=1}^{m} \sigma_i h(x_i)\right] + \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h'\in H'} \sum_{i=1}^{m} \sigma_i h'(x_i)\right].
\end{aligned}$$

For the third inequality, using the identity $\max(a, b) = \frac{1}{2}[a + b + |a - b|]$ valid for all $a, b \in \mathbb{R}$, and the sub-additivity of $\sup$ we can write:

$$\begin{aligned}
&\mathfrak{R}_m(\{\max(h, h') \colon h \in H, h' \in H'\}) \\
&= \frac{1}{2m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i[h(x_i) + h'(x_i) + |h(x_i) - h'(x_i)|]\right] \\
&\leq \frac{1}{2}[\mathfrak{R}_m(H) + \mathfrak{R}_m(H')] + \frac{1}{2m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i|h(x_i) - h'(x_i)|\right].
\end{aligned}$$

Since the absolute value function is 1-Lipschitz, by the contraction lemma, the third term can be bounded as follows

$$\begin{aligned}
&\frac{1}{2m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i|h(x_i) - h'(x_i)|\right] \\
&\leq \frac{1}{2m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i[h(x_i) - h'(x_i)]\right] \\
&\leq \frac{1}{2m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i h(x_i) + \sup_{h\in H, h'\in H'} \sum_{i=1}^{m} -\sigma_i h'(x_i)]\right] \\
&= \frac{1}{2m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} \sigma_i h(x_i)\right] + \frac{1}{2m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H, h'\in H'} \sum_{i=1}^{m} -\sigma_i h'(x_i)]\right] \\
&= \frac{1}{2}[\mathfrak{R}_m(H) + \mathfrak{R}_m(H')],
\end{aligned}$$

using the fact that σ_i and $-\sigma_i$ follow the same distribution. □

B. VC-dimension

1. What is the VC-dimension of the family of subsets of the real line $[x, x+1] \cup [x+2, +\infty)$, with $x \in \mathbb{R}$?

   Solution: The VC-dimension of this family is less than 4 since the labeling $+-+-$ is not possible to obtain using any hypothesis: for any set of 4 points $\{x_1, x_2, x_3, x_4\}$ with $x_1 < x_2 < x_3 < x_4$, the labeling of the first 3 points imposes $x_1 \in [x, x+1]$, $x_2 \in (x+1, x+2)$, and $x_3 \in [x+2, +\infty)$ for some $x \in \mathbb{R}$, which implies a positive labeling for x_4.

   The VC-dimension is equal to 3 since there is a set of 3 points that can be fuly shattered. An example is $\{1/4, 9/8, 7/4\}$ with the following intervals for each labeling sequence:

$$
\begin{array}{ll}
+++ & [0, +1] \cup [2, +\infty) \\
++- & [1/4, 5/4] \cup [9/4, +\infty) \\
+-+ & [-1/2, +1/2] \cup [3/2, +\infty) \\
+-- & [0, 1] \cup [2, +\infty) \\
-++ & [-1, 0] \cup [1, +\infty) \\
-+- & [1/2, 3/2] \cup [5/2, +\infty] \\
--+ & [3/2, 5/2] \cup [7/2, +\infty) \\
--- & [2, 3] \cup [4, +\infty).
\end{array}
$$

   □

2. VC-dimension of sine functions. Consider the hypothesis family of sine functions: $\{x \to \operatorname{sign}(\sin(\omega x)) : \omega \in \mathbb{R}\}$.

   (a) Show that for any $x \in \mathbb{R}$ the points $x, 2x, 3x$ and $4x$ cannot be shattered by this family of sine functions.

   Solution:

   Fix $x \in \mathbb{R}$ and suppose there exists an ω that realizes the labeling $--+-$. Thus $\sin(\omega x) < 0$, $\sin(2\omega x) < 0$, $\sin(3\omega x) \geq 0$ and $\sin(4\omega x) < 0$. We will show that this implies $\sin^2(\omega x) < \frac{1}{2}$ and $\sin^2(\omega x) \geq \frac{3}{4}$, a contradiction.

   Using the identity $\sin(2\theta) = 2\sin(\theta)\cos(\theta)$ and the fact that $\sin(4\omega x) < 0$ we have

$$2\sin(2\omega x)\cos(2\omega x) = \sin(4\omega x) < 0.$$

Since $\sin(2\omega x) < 0$ we can divide both sides of this inequality by $2\sin(2\omega x)$ to conclude $\cos(2\omega x) > 0$. Applying the identity $\cos(2\theta) = 1 - 2\sin^2(\theta)$ yields $1 - 2\sin^2(\omega x) > 0$, or $\sin^2(\omega x) < \frac{1}{2}$.
Using the identity $\sin(3\theta) = 3\sin(\theta) - 4\sin^3(\theta)$ and the fact that $\sin(3\omega x) \geq 0$ we have

$$3\sin(\omega x) - 4\sin^3(\omega x) = \sin(3\omega x) \geq 0$$

Since $\sin(\omega x) < 0$ we can divide both sides of this inequality by $\sin(\omega x)$ to conclude $3 - 4\sin^2(\omega x) \leq 0$, or $\sin^2(\omega x) \geq \frac{3}{4}$.

□

Interestingly, it is possible to find a set of four evenly spaced points that can be shattered by this family of sine functions, provided that the distance from the origin to the nearest point is allowed to differ from the distance between the points. For example, the points $\{1, 1.5, 2, 2.5\}$ are shattered by the frequencies

$$\omega \in \{-10, -9, -8, -7, -6, -5, -4, -1, 1, 2, 4, 5, 11, 14, 15, 23\}.$$

(b) Show that the VC-dimension of the family of sine functions is infinite. (*hint*: show that $\{2^{-m} : m \in \mathbb{N}\}$ can be fully shattered for any $m > 0$.)

Solution:
For any $m > 0$, consider the set of points $(x_1, \ldots, x_m)$ with arbitrary labels $(y_1, \ldots, y_m) \in \{-1, +1\}^m$. Now, choose the parameter $\omega = \pi(1 + \sum_{i=1}^{m} 2^i y'_i)$ where $y'_i = \frac{1-y_i}{2}$. We show that this single parameter will always correctly classify the entire sample for any $m > 0$ and choice of labels. For any $j \in [1, m]$ we have

$$\begin{aligned}\omega x_j = \omega 2^{-j} &= \pi(2^{-j} + \sum_{i=1}^{m} 2^{i-j} y'_i) \\ &= \pi(2^{-j} + (\sum_{i=1}^{j-1} 2^{i-j} y'_i) + y'_j + (\sum_{i=1}^{m-j} 2^i y'_i)) \,.\end{aligned}$$

The last term can be dropped from the sum since it only contributes multiples of 2π. Since $y'_i \in \{0, 1\}$ the remaining term, that is, $\pi(2^{-j} + (\sum_{i=1}^{j-1} 2^{i-j} y'_i) + y'_j) = \pi(\sum_{i=1}^{j-1} 2^{-i} y'_i + 2^{-j} + y'_j)$, can be upper and

lower bound as follows:

$$\pi(\sum_{i=1}^{j-1} 2^{-i} y'_i + 2^{-j} + y'_j) \leq \pi(\sum_{i=1}^{j} 2^{-i} + y'_j) < \pi(1 + y'_j)\,,$$
$$\pi(\sum_{i=1}^{j-1} 2^{-i} y'_i + 2^{-j} + y'_j) > \pi y'_j\,.$$

Thus, if $y_j = 1$ we have $y'_j = 0$ and $0 < \omega x_j < \pi$, which implies $\text{sign}(\omega x_j) = 1$. Similarly, for $y_j = -1$ we have $\text{sign}(\omega x_j) = -1$. □

C. Support Vector Machines

1. Download and install the `libsvm` software library from:

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

2. Consider the `spambase` data set

   `http://archive.ics.uci.edu/ml/datasets/Spambase`.

   Download a shuffled version of that dataset from

   `http://www.cs.nyu.edu/~mohri/ml16/spambase.data.shuffled`

   Use the `libsvm` scaling tool to scale the features of all the data. Use the first 3450 examples for training, the last 1151 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 that consists of predicting if the e-mail message if a spam using the 57 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

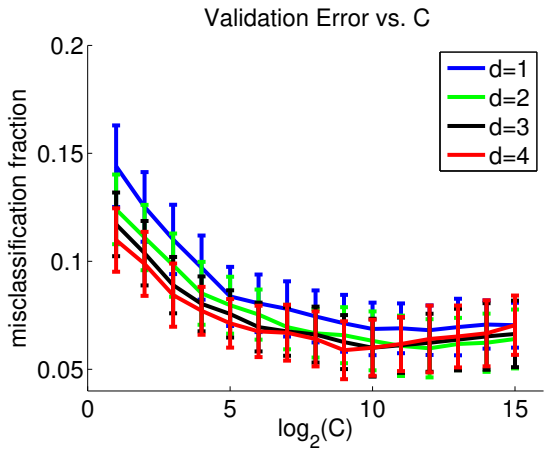

Figure 1: Average error according to 10-fold cross-validation, with error-bars indicating one standard deviation.

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. Note that the algorithm starts to exhibit some over-fitting as C becomes very large. The performance for several choices of d and C are essentially indistinguishable; one suitable choice of optimal parameters is $C^* = 2^9$ and $d^* = 4$.

4. Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the ten-fold cross-validation error and the test errors for the hypotheses obtained as a function of d. Plot the average number of support vectors obtained as a function of d. How many of the support vectors lie on the margin hyperplanes?

   Solution: The first plot in Figure 2 shows that the test error decreases (slightly) with an increase in degree and also that the cross-validation error is (slightly) optimistic when compared to the test error on the held-out dataset.

   The second plot shows that the total number of marginal support vectors increases with d (as does the dimension of the feature space) while the total number of overall support vectors decreases. This also implies that the number of support vectors due to mistakes is decreasing, which agrees with the

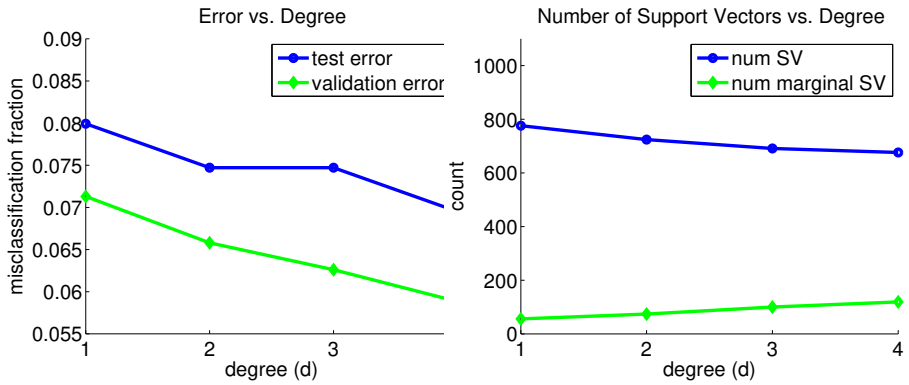

Figure 2: The test and validation error as a function of degree (left panel) as well as the number of total and marginal support vectors (right panel).

first plot.

5. 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:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \\ \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi \forall i \in [1,m] \\ & \xi \geq 0. \end{aligned} \tag{5}$$

   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:

$$\nabla_{\mathbf{w}} L = 0 \implies \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = 0 \implies \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\nabla_\xi L = 0 \implies \sum_{i=1}^{m} \alpha_i + \beta = C.$$

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

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

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

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: $1/p + 1/q = 1$. □