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 $\mathfrak{R}_m(\mathcal{H})$.

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

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: $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{F})$.

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.)

B. VC-dimension

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

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

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.)

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.

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.

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 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.