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: $\mathfrak{R}'(H) = \frac{1}{m}\operatorname{E}_{\boldsymbol{\sigma},S}[\sup_{h\in H}|\sum_{i=1}^{m}\sigma_i h(x_i)|]$. Show the following upper bound for a hypothesis set reduced to a single hypothesis h:

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

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

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)).

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 $x \in \mathbb{R}$?

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

(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.

(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 $m > 0$.)

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 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 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?

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 $\|\boldsymbol{\alpha}\|_1$.