Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2
October 25, 2017
Due: November 08, 2017

A. Growth function

Growth function of stump functions. For any $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$, let ϕ_θ denote the threshold function that assigns sign $+1$ to $x \leq \theta$, -1 otherwise: $\phi_\theta(x) = 21_{x \leq \theta} - 1$. Let $\mathcal{H}$ be the family of functions mapping $\mathbb{R}^N$ to $\{-1, +1\}$ defined by

$$\mathcal{H} = \Big\{ \mathbf{x} \mapsto s\, \phi_\theta(x_i) \colon i \in [1, N], \theta \in \mathbb{R}, s \in \{-1, +1\} \Big\},$$

where x_i is the ith coordinate of $\mathbf{x} \in \mathbb{R}^N$.

1. Show that the following upper bound holds for the growth function of $\mathcal{H}$: $\Pi_m(\mathcal{H}) \leq 2mN$.

2. Let $\mathcal{H}_2$ be the family of functions mapping $\mathbb{R}^N$ to $\{-1, +1\}$ defined by

$$\mathcal{H}_2 = \Big\{ \mathbf{x} \mapsto s\, 1_{\phi_\theta(x_i)=+1}\, \phi_{\theta'}(x_j) + s'\, 1_{\phi_\theta(x_i)=-1}\, \phi_{\theta'}(x_j) \colon \\ i \neq j, i, j \in [1, N], \theta, \theta' \in \mathbb{R}, s, s' \in \{-1, +1\} \Big\}.$$

   Show that $\Pi_m(\mathcal{H}_2) = O(m^2 N^2)$. Give an explicit upper bound on $\Pi_m(\mathcal{H}_2)$.

B. VC-dimension

1. VC-dimension of circles in the plane.

   (a) Show that the VC-dimension of the circles in the planes is 3.

   (b) Let H_1 and H_2 be two families of functions mapping from $\mathcal{X}$ to $\{0, 1\}$ and $H = \{h_1 h_2 \colon h_1 \in H_1, h_2 \in H_2\}$ their product. Show that

$$\Pi_H(m) \leq \Pi_{H_1}(m) \Pi_{H_2}(m).$$

   (c) Give an upper bound on the VC-dimension of the family of intersections of k circles in the planes.

2. VC-dimension of Decision trees. A full binary tree is a tree in which each node is either a leaf or it is an internal node and admits exactly two child nodes.

   (a) Show that that a binary tree with n internal nodes has exactly $n + 1$ leaves (*Hint*: you can proceed by induction).

   (b) A binary decision tree is a full binary tree with each leaf labeled with $+1$ or -1 and each internal node labeled with a question.

   A binary decision tree classifies a point as follows: starting with the root of the tree, if the internal node question applied to the point admits a positive answer, then the current node becomes the right child, otherwise it becomes the left child. This is repeated until a leaf node is reached. The label assigned to the point is then the sign of that leaf node.

   Suppose the node questions are of the form $x_i > 0$, $i \in [1, N]$. Show that the VC-dimension of the set of binary decision trees with n nodes in dimension N is at most $(2n+1)\log_2(N+2)$ (*Hint*: bound the cardinality of the set). Use that to derive an upper bound on the Rademacher complexity of that set.

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/ml17/spambase.data.shuffled`

   Use the `libsvm` scaling tool to scale the features of all the data. Use the first 3000 examples for training, the last 1601 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. For any d, let K_d denote the polynomial kernel of degree d. Show that for any fixed integer $u > 0$, $G_u = \frac{2}{u(u+1)} \sum_{i \leq j \leq u} K_i K_j$ is a PDS kernel. Use SVMs combined with the polynomial kernel G_4 to tackle the same binary classification problem as in the previous questions: as in the previous questions, use ten-fold cross-validation to determine the best value of C; report the ten-fold cross-validation error and the test error of the hypothesis obtained when training with G_u.

D. Rademacher complexity

Let $p > 2$ and let q be its conjugate: $1/p + 1/q = 1$. Let $\mathcal{H}_p$ be the family of linear functions defined over $\{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|_p \leq r_p\}$ by

$$\mathcal{H} = \Big\{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} : \|\mathbf{w}\|_q \leq \Lambda_q\Big\},$$

for some $r_p > 0$ and $\Lambda_q > 0$. Give an upper bound on $\mathfrak{R}_m(\mathcal{H})$ in terms of Λ_q and r_p, assuming that for $p > 2$ the following inequality holds for all $z_1, \ldots, z_m \in \mathbb{R}$: $\mathbb{E}_{\boldsymbol{\sigma}} \left[|\sum_{i=1}^m \sigma_i z_i|^p \right] \leq [\frac{p}{2} \sum_{i=1}^m z_i^2]^{\frac{p}{2}}$.