Homework 3

Akshay Krishnamurthy
Due: Tuesday 10/17

October 11, 2017

Instructions: Turn in your homework in class on Tuesday 10/17/2017

1. **Model Selection.** This problem builds on the $R(h^\star)$ bound we proved in Homework 1. The goal here is to do this simultaneously while doing structural risk minimization. Specifically, given a family of hypothesis classes $\mathcal{H}_1 \subset \mathcal{H}_2 \ldots, \subset \mathcal{H}_L$, of sizes $N_1 \le N_2 \le \ldots \le N_L < \infty$, a loss function bounded on $[0,1]$ and a sample of size n, design an algorithm that guarantees

   $$R(\hat{h}) \le \min_{i \in [L]} \min_{h^\star \in \mathcal{H}_i} \left\{ R(h^\star) + c_1 \sqrt{\frac{R(h^\star)\log(LN_i/\delta)}{n}} + c_2 \frac{\log(LN_i/\delta)}{n} \right\}$$

   for $n \ge 2$. Your algorithm may use ERM (so need not be efficient) and your constants may vary.

   You may find it useful to use the following *empirical bernstein inequality.*

   Theorem 1. *Let $X_1, \ldots, X_n$ be i.i.d. random variables from a distribution P supported on $[0,1]$ and define the sample variance $V_n = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (X_i - X_j)^2$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$*

   $$\mathbb{E}X - \frac{1}{n}\sum_{i=1}^n X_i \le \sqrt{\frac{2V_n \log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)}.$$

2. **Boosting.** Boosting can be understood from many perspectives, and here we will explore a game-theoretic perspective. As some background, a two-player zero-sum game is specified by matrix M and the two players are denoted the row player and the column player. The game is played by the row player choosing a row i and the column player choosing a column j, and the *loss* for the row player is $M(i,j)$, which is also the *reward* for the column player. Instead of choosing individual rows/columns, we will allow both players to choose distributions over rows/columns, so if row player choose P and column player choose Q the loss/reward is

   $$\sum_i \sum_j P(i)M(i,j)Q(j) \triangleq M(P,Q)$$

   If we are acting as the row player, we'd like to choose a distribution P that achieves low loss, no matter what the column player does, in other words, we'd like to choose P that minimizes $\max_Q M(P,Q)$. On the other hand, if we were the column player, we'd like to choose Q that maximizes $\min_P M(P,Q)$. Von Neumann's celebrated minimax theorem states that in fact both of these values are the same, or in some sense it doesn't matter which player goes first in the game:

   $$\max_Q \min_P M(P,Q) = \min_P \max_Q M(P,Q) \triangleq V$$

   V is referred to as the value of the game. In this problem, we'll use boosting to compute the optimal strategy in a particular game.

   Let $\mathcal{H}$ be finite, and fix a target concept $c : \mathcal{X} \to \{+1,-1\}$ (not necessarily in $\mathcal{H}$) and sample $S = \{(X_i, c(X_i)\}_{i=1}^n$ of size n. We will form a matrix $M \in \{0,1\}^{n \times |\mathcal{H}|}$ where $M(i,h) = \mathbf{1}\{h(X_i) = c(X_i)\}$. Here the row player specifies distributions over examples, just as in boosting, and the column player chooses distributions over hypotheses.

(a) Assume that the empirical γ-weak learning assumption holds so that for every distribution P over examples, there exists a hypothesis $h \in \mathcal{H}$ such that $\mathbb{P}_{x\sim P}[h(x) \neq c(x)] \leq 1/2 - \gamma$. What does this mean about the value of the game?

(b) Let $Q^\star$ be the distribution achieving the value of this game, i.e. $\min_P M(P, Q^\star) = V$. Since $Q^\star$ is a distribution over hypotheses, what can you say about the empirical error of $Q^\star$?

(c) Boosting can be viewed as an iterative algorithm to compute $Q^\star$ using a weak learner. At every round we choose P_t, a distribution over samples, and then compute

$$Q_t = \max_Q M(P_t, Q)$$

which is actually a single hypothesis h_t due to linearity. Then we update P_t to be

$$P_{t+1}(x) = \frac{P_t(x)}{Z_t} \times (\exp(-\eta \mathbf{1}\{h_t(x) = c(x)\}))$$

After T rounds, we output $\bar{Q} = \frac{1}{T}\sum_{t=1}^{T} Q_t$ which is a distribution over hypothesis and the final predictor is $H(x) = \text{sign}(\bar{Q}(x))$.

Prove that once $T = \Omega(\log(n)/\gamma^2)$ rounds and for appropriate choice of η, this variant of boosting guarantees

$$\forall x \in X. \frac{1}{T}\sum_{i=1}^{T} M(x, h_t) > 1/2,$$

which implies that $H(x)$ has zero training error.

Hint: You may find it helpful to look at the Taylor expansion of $\exp(-x)$.

(d) Informally, what does this mean about $\bar{Q}$, what is it converging to as $T \to \infty$?

3. **Margins and Geometry.** In this problem we briefly investigate the different notions of margin used in the boosting analysis and the SVM analysis. The translation between boosting and SVM is as follows. For boosting, fix the dataset and let $\mathcal{H}$ be finite. Let $w \in \mathbb{R}^{|\mathcal{H}|}$ be a distribution over hypothesis and let $\Phi(x) = (h_1(x), \ldots, h_{|\mathcal{H}|}(x))$ where the hypotheses are numbered arbitrarily. In boosting, we defined the margin for (x, y) as $y\langle w, \Phi(x)\rangle$, which is exactly how we defined the margin for SVM but where $\Phi(x) = x \in \mathbb{R}^d$ and $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$ (for hard-margin SVM). The primary difference here is that in Boosting we have that w is a distribution (so $\|w\|_1 = 1$) and $\|\Phi(x)\|_\infty = 1$ while in SVM we have $\|w\|_2 = 1$ and $\|\Phi(x)\|_2 \leq 1$, so that the geometries are different. Here we will investigate the differences. Let $|\mathcal{H}| = d$ so both problems are in the same dimension.

(a) Prove that the same margin-type generalization bound applies for SVM: For any $\theta > 0$ and for all weight vectors $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$ and where $\mathcal{X} = \{x \mid \|x\|_2 \leq 1\}$, with probability at least $1 - \delta$,

$$\mathbb{P}_{\mathcal{D}}[y\langle w, x\rangle \leq 0] \leq \mathbb{P}_S[y\langle w, x\rangle \leq \theta] + \frac{1}{\theta\sqrt{n}} + 4\sqrt{\frac{\log(4/\delta)}{n}}.$$

Hint: for a simpler proof, use the Rademacher complexity of the linear class with the ramp function $\varphi(u) = \mathbf{1}\{u \leq 0\} + (1 - u/\theta)\mathbf{1}\{0 \leq u \leq \theta\}$.

(b) Suppose that in the boosting case $\Phi(x) \in \{-1, +1\}^d$ and in the SVM case $\Phi(x) \in \{-1/\sqrt{d}, +1/\sqrt{d}\}^d$ so that the normalization works out. Let k be an odd number and suppose that w is the majority vote over k of the hypotheses/dimensions. What are the worst-case Boosting and SVM margins in this case? When k is small, why is this favorable for boosting?

(c) Suppose instead that $\Phi(x)$ are s-sparse binary vectors (with s odd and entries ± 1 in the boosting case and entries $\pm 1/\sqrt{s}$ in the SVM case). For boosting, you can think of this as some most classifiers abstaining from making a prediction on each example. Let w be a majority over all hypotheses dimensions. What are the worst case boosting and SVM margins here? When s is small why is this favorable for SVM?

(d) Looking at the generalization bounds, in words explain how the problem geometry can dramatically impact performance and why the choice between boosting and SVM can problem-dependent.

4. **SVM.** In class we saw the hard margin SVM formulation

$$\underset{w}{\text{minimize}} \frac{1}{2}\|w\|_2^2 \text{ s.t. } \forall i \in [n], y_i\langle w, x_i\rangle \geq 1.$$

We also saw the *soft margin SVM* optimization

$$\underset{w,\xi}{\text{minimize}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i \text{ s.t. } \forall i \in [n], y_i\langle w, x_i\rangle \geq 1 - \xi_i, \text{ and } \xi_i \geq 0.$$

Use Lagrange duality to show that this problem can be expressed purely in terms of inner products between the points, so that we can apply the Kernel trick.

$$\underset{\alpha\in\mathbb{R}^n}{\text{maximize}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle \text{ s.t. } \forall i \in [n], 0 \leq \alpha_i \leq C.$$