Homework 1

Akshay Krishnamurthy
Due: Tuesday 9/19

September 12, 2017

Instructions: Turn in your homework in class on Tuesday 9/19/2017

1. **Linear Regression.** In class, we saw a risk bound for linear regression. Here we will study linear regression from another perspective. We consider the *fixed design* case, where the feature vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ are non-random and assume that $y_i = \langle \beta^\star, x_i \rangle + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid, but $\beta^\star$ is unknown. Note that the only randomness is in the error variables $\epsilon_1, \ldots, \epsilon_n$. Let $\Sigma = \frac{1}{n}\sum x_i x_i^T$ denote the second moment matrix, and assume that that Σ is non-singular with minimum eigenvalue $\lambda_{\min}(\Sigma) \geq \kappa$. Let $\hat{\beta}$ be the empirical risk minimizer with the square loss (e.g., the ordinary least squares estimator $\hat{\beta} = (n\Sigma)^{-1}\sum_{i=1}^n x_i y_i$). Show that

$$\mathbb{E}\|\hat{\beta} - \beta\|_2^2 \leq \frac{d\sigma^2}{\kappa n}.$$

   Note that while we didn't show it here, this basic result also applies for the random design setting, where instead we use $\kappa = \lambda_{\min}(\mathbb{E}x_i x_i^T)$, and it also holds with high probability. However this extensions require more powerful concentration inequalities than we'll see in this class.

2. **PAC Learning.** In this problem we will design and analyze an algorithm for PAC-learning the concept class of decision lists. The domain is $\{0,1\}^d$, the boolean hypercube. The hypotheses in consideration are known as *decision lists*, which are a sequence of if then else rules, which take the form "If ℓ_1 then b_1, else if ℓ_2 then b_2 else if ℓ_3 then, ... else b_k" where ℓ_is are boolean literals (either x_j or $\bar{x}_j$ for some $j \in [d]$), and $b_i \in \{-1, 1\}$.

   (a) Design an algorithm for computing an ERM decision list in the PAC model.

   (b) Provide a bound on $\log|\mathcal{H}|$, where $\mathcal{H}$ is the hypotheses induced by decision lists. What upper bound on the sample complexity of PAC-learning decision lists does this imply?

3. **Robust mean estimation.** Let $X_1, \ldots, X_n$ be iid random variables from a distribution P with unknown mean μ and variance $\sigma^2 < \infty$. We are interested in estimating the mean μ from the data. This is called robust mean estimation since we make very minimal assumptions on P, which could be very heavy tailed.

   (a) Prove the Efron-Stein inequality. For any function $f : \mathbb{R}^n \to \mathbb{R}$ and $X_1, \ldots, X_n$ iid,

$$\mathrm{Var}(f(X_1, \ldots, X_n)) \leq \mathbb{E}\sum_{i=1}^n \mathrm{Var}(f(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

   (b) Consider the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Show that, with probability at least $1 - \delta$,

$$|\bar{X} - \mu| \leq \sqrt{\frac{\sigma^2}{n\delta}}$$

   (c) A better estimator is the median-of-means estimator defined as follows. Choose a number k (assume for simplicity that $n = mk$ and that k is even) and partition the data into k groups. In each group compute the sample means $\mu_j = \frac{1}{m}\sum_{i=m(j-1)+1}^{mj} X_i$ for $j \in \{1, \ldots, k\}$. Then take the median of these k values, i.e. choose any number $\hat{\mu}$ that is larger than exactly $k/2$ of these sample means (mathematically,

$|\{\mu_j : \mu_j \le \hat{\mu}\}| = k/2)$. Show that, for appropriate settings of k and consequently m, with probability at least $1-\delta$

$$|\hat{\mu} - \mu| \le \sqrt{\frac{64\sigma^2 \log(1/\delta)}{n}},$$

Provided n is large enough. Other constants here are also fine.

4. **Uniform Convergence.** In this problem we will prove a stronger generalization error bound for the agnostic binary classification, that uses more information about the distribution. Let P be a distribution over (X, Y) pairs where $X \in \mathcal{X}$ and $Y \in \{+1, -1\}$ and let $\mathcal{H} \subset \mathcal{X} \to \{+1, -1\}$ be a finite hypothesis class and let ℓ denote the zero-one loss $\ell(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$. As usual let $R(h) = \mathbb{E}\ell(h(X), Y)$ denote the risk, and let $h^\star = \min_{h \in \mathcal{H}} R(h)$. Given n samples let $\hat{h}_n$ denote the empirical risk minimizer. The goal here is to prove a sample complexity bound of the form:

$$R(\hat{h}_n) - R(h^\star) \le c_1 \sqrt{\frac{R(h^\star)\log(|\mathcal{H}|/\delta)}{n}} + c_2 \frac{\log(|\mathcal{H}|/\delta)}{n}. \tag{1}$$

for constants c_1, c_2. This can be a much better bound than the usual excess risk bound we saw in class, if $R(h^\star)$ is small. In particular, if $R(h^\star) = 0$ as in the realizable setting, this bound recovers the $1/n$-rate.

(a) To prove the result, we will use Bernstein's inequality, which is a sharper concentration result.

Theorem 1 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be iid real-valued random variables with mean zero, and such that $|X_i| \le M$ for all i. Then for all $t > 0$*

$$\mathbb{P}[\sum_{i=1}^{n} X_i \ge t] \le \exp\left(-\frac{t^2/2}{\sum_{i=1}^{n}\mathbb{E}[X_i^2] + Mt/3}\right).$$

We will not prove this here. Use the inequality to show that with probability at least $1-\delta$

$$|\bar{X}| \le \sqrt{\frac{2\mathbb{E}X_1^2 \log(2/\delta)}{n}} + \frac{2M\log(2/\delta)}{3n}. \tag{2}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and X_is satisfy the conditions of Bernstein's inequality.

(b) Use Eq. (2) and the union bound to show Eq. (1).