Homework 2
CS229T/STATS231 (Winter 2014–2015)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something's true (by algebra, convexity, etc.). There's no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived. Up to two additional points will awarded for especially well-organized and elegant solutions.

Due date: Wednesday, Feb. 11 (at the beginning of class)

1. Improved generalization in low error regimes (*10 points*)

Recall that in the realizable setting (where the expected risk minimizer h^* satisfies $L(h^*) = 0$), we obtained excess risk bounds of $O(1/n)$, but in the unrealizable setting, we had $O(\sqrt{1/n})$. What if the learning problem is *almost* realizable, in that $L(h^*)$ is small? Can we obtain bounds that gracefully interpolate between the two? This problem explores this possibility.

To start out, let's revisit Hoeffding's inequality, which was used to prove generalization bounds in the unrealizable setting. Recall that Hoeffding's inequality states that if $X_1, \ldots, X_n$ are independent random variables such with $\mu = \mathbb{E}[X_i]$, and $a \leq X_i \leq b$ with probability 1 for each i, then

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right] \leq \exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right). \tag{1}$$

Since Hoeffding's inequality only depends on the upper and lower bounds a and b of X_i, it can be very loose when the X_i has low variance. For example, compare (i) $X_i = -1$ or $X_i = +1$, each with probability $\frac{1}{2}$; and (ii) $X_i = 0$ with probability 0.98 and $X_i = -1$ or $X_i = +1$, each with probability 0.01. Example (ii) should intuitively enjoy a sharper bound because it has smaller variance. In this problem, we will derive better generalization bounds that depend on variance.

Instead of using Hoeffding's inequality, we will use *Bernstein's inequality*. The setup is the same as in Hoeffding's inequality, except that we also define $\sigma^2 = \mathrm{Var}[X_i]$. The bound is as follows:

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2(\sigma^2 + (b-a)\epsilon/3)}\right). \tag{2}$$

(**a**) We first consider hypotheses with expected risk that is bounded above by a constant E. (The existence of a small upper bound E is what makes the problem almost realizable.) Equipped with Bernstein's inequality, we prove a concentration bound, relating empirical and expected risk for such hypotheses.

Assume our loss function is bounded as follows: $\ell(y, p) \in [0, 1]$. Suppose that we have a fixed predictor $h : \mathcal{X} \to \mathbb{R}$ that achieves expected risk at most E; that is, $L(h) \leq E$, where

$$L(h) \stackrel{\text{def}}{=} \mathbb{E}_{(x,y)\sim p^*}[\ell(y, h(x))].$$

Recall that we defined the empirical risk as the random variable:

$$\hat{L}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, h(x^{(i)})).$$

Show that

$$\mathbb{P}[\hat{L}(h) - L(h) \geq \epsilon] \leq \exp\left(\frac{-n\epsilon^2}{2(E + \epsilon/3)}\right).$$

Remark: When $E = 0$, the exponent behaves like $O(-n\epsilon)$, which is much better than the usual $O(-n\epsilon^2)$ when ϵ is small.

(**b**) Next, we will prove a sort of converse to the above. Consider hypotheses with expected risk that is *at least* some amount $E' + \epsilon$. (If such an amount is large, these hypotheses are far from "realizing" the minimum expected problem.) We will show that, as ϵ increases from zero, it is increasingly unlikely for the empirical risk of such a hypothesis to fall below the risk threshold E'.

Formally, suppose that instead we now have another fixed predictor h' with expected risk at least $E' + \epsilon$:

$$L(h') \geq E' + \epsilon.$$

Show that it is unlikely that the empirical risk $\hat{L}(h')$ is less than E':

$$\mathbb{P}[\hat{L}(h') \leq E'] \leq \exp\left(\frac{-n\epsilon^2}{2(E' + 4\epsilon/3)}\right).$$

(**c**) We now bound the excess risk in terms of the smallest valid expected risk bound, $E = L(h^*)$.

Suppose that our hypothesis class $\mathcal{H}$ is finite with $|\mathcal{H}|$ elements. Use the preceding parts to conclude that the empirical risk minimizer $\hat{h}$ achieves:

$$\mathbb{P}[L(\hat{h}) - L(h^*) \geq 2\epsilon] \leq |\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2(L(h^*) + 7\epsilon/3)}\right).$$

(**d**) In the previous part, we proved a generalized excess risk bound with the dependencies that we had originally desired. The bound applies beyond the realizable setting, instead depending on the "extent of realizability" $L(h^*)$.

Compare this bound with (i) the bound we have for the realizable case and (ii) the usual bound one obtains with Hoeffding's inequality. Comment on the relationship between them. (Hint: you may want to have a discussion on the risk threshold E.)

2. Complexity of hypothesis classes (*15 points*)

Generalization bounds for the empirical risk minimizer (ERM) depend on the complexity of the hypothesis class that the ERM is defined over. Recall that there are several ways to measure the complexity of $\mathcal{H}$. In this problem, we will compute the VC dimensions or Rademacher complexity of certain hypothesis classes, in order to develop an intuition for the difficulty of learning these hypothesis classes.

(**a**) Let the input space be $\mathcal{X} = \mathbb{R}^p$ for $p \geq 2$ and consider hypotheses consisting of all convex sets:

$$\mathcal{H} = \{h_S(x) = \mathbb{I}[x \in S] : S \text{ convex}\}.$$

Compute the VC dimension of $\mathcal{H}$.

(**b**) A decision tree T is a binary tree that classifies points in $\mathbb{R}^d$. Each internal node (non-leaf node) v in T has an attribute $j_v \in \{1, 2, \ldots, d\}$ and a threshold $t_v \in \mathbb{R}$. Each leaf node is labeled with one of the two classes, +1 or -1. Given a point $x \in \mathbb{R}^d$, we start from the root, and every time we encounter an internal node v, we check the condition $\mathbb{I}[x_{j_v} \geq t_v]$. We go to the left child if the condition is not met,

and the right child otherwise. We repeat such process until we reach a leaf node, and classifies the point according to the label of the node.

Show that the VC dimension of the hypothesis class corresponding to all depth-k decision trees defined above is $\Omega(2^k \log d)$.

(**c**) Recall that the Rademacher complexity of a class of functions $\mathcal{F}$ is defined as

$$R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right],$$

where $Z_1, \ldots, Z_n$ are drawn i.i.d. from some distribution p^* and $\sigma_1, \ldots, \sigma_n$ are Rademacher variables drawn i.i.d. from $\{-1, 1\}$ with equal probability of $+1$ and -1.

Let $f : \mathcal{X} \to \mathbb{R}$ be a function, and let $\mathcal{F} := \{-f, f\}$ be a function class containing only two functions. Upper bound $R_n(\mathcal{F})$ using a function of n and $\mathbb{E}[f(X)^2]$.

(**d**) In applications such as natural language processing, we often have sparse feature vectors. Suppose that $x \in \{0, 1\}^d$ has only k non-zero entries. For example, in document classification, one feature might be "$x_{17} = 1$ iff the document contains the word *cat*."

Define the class of linear functions whose coefficients have bounded L_∞ norm:

$$\mathcal{F} = \{x \mapsto w \cdot x : \|w\|_\infty \le B\}.$$

Compute an upper bound on the Rademacher complexity $R_n(\mathcal{F})$. Express your answer as a function of B, k, d, n. Note that this allows us to effectively control the complexity of learning using L_∞ regularization.

(**e**) Consider a prediction problem from $x \in \mathbb{R}$ to $y \in \{0, \ldots, k\}$. For every parameter vector $\theta \in \mathbb{R}^k$, define the prediction function $h_\theta(x) = \sum_{i=1}^{k} \mathbb{I}[x \ge \theta_i]$ (monotonically increasing piecewise constant functions). Define the loss function to be $\ell(y, p) = |y - p|$, yielding the following loss class:

$$\mathcal{A} = \{(x, y) \mapsto \ell(y, h_\theta(x)) : \theta \in \mathbb{R}^k\}.$$

Compute an upper bound on the Rademacher complexity of $\mathcal{A}$.

(**f**) Let $\mathcal{F}$ be the class of all continuous functions $f : [0, 1] \to [0, 1]$ with at most k local maxima. Find an upper bound of the Rademacher complexity of $\mathcal{F}$.

3. Online-to-batch conversion in high probability (*5 points*)

In online learning, the learner receives a sequence of convex functions $f_1, \ldots, f_T$, and returns a sequence of weight vectors $w_1, \ldots, w_T$, where w_t only depends on $f_1, \ldots, f_{t-1}$. Assume that all weight vectors have bounded norm: $w \in S$, where $S \stackrel{\text{def}}{=} \{w : \|w\|_2 \le B\}$. Furthermore, assume that all the functions f_t have bounded subgradients: for all $w \in S$, we have each $z_t \in \partial f_t(w)$ satisfying $\|z_t\|_2 \le L$. In class, we proved a regret bound, namely for all $u \in S$:

$$\text{Regret}(u) = \sum_{t=1}^{T} [f_t(w_t) - f_t(u)] \le BL\sqrt{T}. \tag{3}$$

To perform online-to-batch conversion, we assume all the f_t's are i.i.d. with $L(w) = \mathbb{E}[f_t(w)]$ and let $\hat{w} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} w_t$. In class, we showed a generalization bound in expectation, but ideally, we'd like a result that holds with high probability. Let $w^* \in \arg\min_{w \in S} L(w)$. Show that with probability at least $1 - \delta$,

$$L(\hat{w}) \leq L(w^*) + \sqrt{\frac{B^2 L^2}{T}} + \sqrt{\frac{8B^2 L^2 \log(1/\delta)}{T}}.$$

(**a**)

4. Feedback (*0 points*)

(**a**) On a scale of 1 to 10, how difficult was this assignment?

(**b**) On a scale of 1 to 10, how useful was this assignment?

(**c**) Any other comments?