Theoretical Foundations of Machine Learning - Homework #3

Jacob Abernethy Due: 10/30/2015

Homework Policy: Working in groups is fine, but *every student* must submit their own writeup. Please write the members of your group on your solutions. There is no strict limit to the size of the group but we may find it a bit suspicious if there are more than 4 to a team. Questions labelled with **(Challenge)** are not strictly required, but you'll get some participation credit if you have something interesting to add, even if it's only a partial answer.

1) **Rademacher Complexity Identities.** Problem 3.4 from the book, parts (a) and (b).

2) **Model Selection.** Assume our space $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$, $\mathcal{Y} := \{-1, 1\}$, we have an unknown distribution D on $\mathcal{X} \times \mathcal{Y}$ our loss is the hinge loss $\ell(y', y) = \max(0, 1 - yy')$. With a slight abuse of notation we will write $\ell(\mathbf{w}, (\mathbf{x}, y))$ to mean $\ell((\mathbf{w}^\top \mathbf{x}), y)$. Our class of functions H are linear thresholds $H := \{h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x}) : \mathbf{w} \in \mathbb{R}^d\}$. The risk of $\mathbf{w}$ (that is, of $h_{\mathbf{w}}$) is $R(\mathbf{w}) := \mathbb{E}_{(\mathbf{x},y)\sim D}[\ell(\mathbf{w}, (\mathbf{x}, y))]$. Let $R^*_{\text{lin}} := \inf_{\mathbf{w}\in\mathbb{R}^d} R(\mathbf{w})$ be the optimal risk for linear threshold functions. Given m samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ then the empirical risk is $\hat{R}_m(\mathbf{w}) := \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{w}, (\mathbf{x}_i, y_i))$.

In high-dimension, it is often the case that the optimal $\mathbf{w}^*$, that is the minimizer of $R(\cdot)$, will have to be very large, or possibly grow without bound. In other words, the margin may be tiny or even infinitesimally small. On the other hand it is typical that a slight amount of regularization can help. Let us assume that

$$\inf_{\mathbf{w}\in\mathbb{R}^d} R(\mathbf{w}) + \lambda\|\mathbf{w}\|^2 \;\to\; R^*_{\text{lin}} \qquad \text{as } \lambda \to 0. \tag{1}$$

This suggests that it is a good idea to regularize our objective when we have limited data. Let

$$\mathbf{w}_m = \arg\min_{\mathbf{w}} \hat{R}_m(\mathbf{w}) + \lambda_m\|\mathbf{w}\|^2 \quad \text{and} \quad \mathbf{w}^*_m = \arg\min_{\mathbf{w}} R(\mathbf{w}) + \lambda_m\|\mathbf{w}\|^2$$

for some choice of regularization parameter λ_m (decaying with m and to be determined), and note that the arg mins are unique since the objectives are strictly convex. Let's analyze the performance of $\mathbf{w}_m$, and figure out how to select the sequence λ_m.

(a) Prove that both $\|\mathbf{w}_m\|^2 \leq \frac{1}{\lambda_m}$ and $\|\mathbf{w}^*_m\|^2 \leq \frac{1}{\lambda_m}$. (*Hint:* note that $\mathbf{0}$ is a possible choice for $\mathbf{w}$)

(b) Let m be fixed. Prove that with probability at least $1 - \delta$,

$$|R(\mathbf{w}_m) - \hat{R}_m(\mathbf{w}_m)| \leq \epsilon(m, \lambda_m, \delta) \quad \text{and} \quad |R(\mathbf{w}^*_m) - \hat{R}_m(\mathbf{w}^*_m)| \leq \epsilon(m, \lambda_m, \delta)$$

for a suitable choice of function $\epsilon(\cdot, \cdot, \cdot)$ (and notice that $\epsilon(\cdot, \cdot, \cdot)$ need not depend on d).

(c) Now prove the same statement but *uniformly* over m. That is, find a function $\epsilon(\cdot, \cdot, \cdot)$ such that with probability at least $1 - \delta$ we have for any m that

$$|R(\mathbf{w}_m) - \hat{R}_m(\mathbf{w}_m)| \leq \epsilon(m, \lambda_m, \delta) \quad \text{and} \quad |R(\mathbf{w}^*_m) - \hat{R}_m(\mathbf{w}^*_m)| \leq \epsilon(m, \lambda_m, \delta).$$

(*Hint:* Union bound followed by choosing a reasonable sequence δ_m, so that $\delta = \sum_{m=1}^\infty \delta_m$. It is likely you will get additional $\log m$ terms.)

(d) For the choice of $\epsilon(\cdot,\cdot,\cdot)$ above, prove that with probability $1-\delta$ it holds that for all m

$$R(\mathbf{w}_m) \leq R(\mathbf{w}_m) + \lambda_m \|\mathbf{w}_m\|^2 \leq R(\mathbf{w}_m^*) + \lambda_m \|\mathbf{w}_m^*\|^2 + 2\epsilon(m, \lambda_m, \delta).$$

(*Hint:* the actual form of $\epsilon(\cdot,\cdot,\cdot)$ doesn't matter, as long is it works in part (c))

(e) Find a suitable choice for the sequence λ_m that depends only on m, and prove that for this sequence it holds that $R(\mathbf{w}_m) \to R^*_{\text{lin}}$ as $m \to \infty$ with high probability. (*Hint*: Your solution should combine equation 1 and part (d).)

What we are showing here is that our regularization algorithm is *consistent*, i.e. our chosen hypothesis $\mathbf{w}_m$ will approach optimality as the size of our sample grows towards infinity.