MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2018**

Problem Set 4

Due Date: **Mon, Nov. 26 2018, 11:59 pm** *(online)*

Instructions: There are **2 problems** in total in this problem set. The breakdown of individual scores per sub-problem are provided. Use the provided LaTeX template to typeset your report. Provide sufficient explanations in all solutions but avoid proving lecture or out-of-scope material (unless explicitly asked to). An 8-page submission maximum is allowed (do not change the font or margin of the template).

What to submit: Submit your report **online through Stellar** by the due date/time. Submission must be a single pdf in LaTeX format. Include code files separately, if applicable.

Policies: Collaborative reports are not allowed. Even if you discuss problems with classmates, you are expected to write and submit **individual reports**.

Problem 1 [30 points] (Adversarially-Robust Learning) Recall that hinge loss is defined as $\ell_h(w,(x,y)) = \max\{0, 1 - y\langle w, x\rangle\}$. You are worried about adversarial examples, and you'd like to define a loss function that will encourage the decision boundary to be sufficiently far from your training data. A common approach is to define an adversarially-robust loss

$$\ell_{adv}(w,(x,y)) = \max_{\delta:\|\delta\|\le\Delta} \max\{0, 1 - y\langle w, x+\delta\rangle\}$$

where the Euclidean norm of the adversarial perturbation δ is constrained by $\Delta \ge 0$.

Suppose $\mathscr{S} = \{(x_1,y_1),\dots,(x_n,y_n)\}$ are your data, with $y_i \in \{\pm 1\}$ and $\|x_i\| \le 1$. Let $S^{d-1} = \{w \in \mathbb{R}^d : \|w\| = 1\}$ be the Euclidean sphere. Consider the following classes of functions:

$$\mathcal{G} = \{(x,y) \mapsto \ell_h(w,(x,y)) : w \in S^{d-1}\}$$

$$\mathcal{G}_{adv} = \{(x,y) \mapsto \ell_{adv}(w,(x,y)) : w \in S^{d-1}\}$$

As a reminder,

$$\widehat{\mathscr{R}}_n(\mathcal{G}|_{\mathscr{S}}) = \mathbb{E}_\epsilon \max_{w\in S^{d-1}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(w,(x_i,y_i)).$$

1.1 [10pts] Suppose $\|w\| = 1$. Prove that ℓ_{adv} is equal to $\max\{0, (1+\Delta) - y\langle w, x\rangle\}$, which is essentially the same as the original hinge loss.

1.2 [5pts] Prove that $|\max\{0,a\} - \max\{0,b\}| \le |a-b|$ for all $a, b \in \mathbb{R}$.

1.3 [10pts] Let

$$G = \{\ell_h(w,(x_1,y_1)),\ldots,\ell_h(w,(x_n,y_n)) : w \in S^{d-1}\} \subset \mathbb{R}^n.$$

Use 1.2 to argue that G can be written as

$$G = \{(\phi_1(g_1),\ldots,\phi_n(g_n)) : (g_1,\ldots,g_n) \in G'\}$$

where

$$G' = \{(\langle w, x_1\rangle,\ldots,\langle w, x_n\rangle) : w \in S^{d-1}\}$$

and $\phi_1,\ldots,\phi_n$ are 1-Lipschitz functions $\mathbb{R} \to \mathbb{R}$. Use the contraction property of Rademacher averages to establish

$$\widehat{\mathscr{R}}_n(\mathcal{G}|_{\mathscr{S}}) \le \frac{1}{\sqrt{n}}.$$

(As mentioned in class, contraction property holds even if different ϕ_i are applied to different coordinates).

1.4 [5pts] Argue that the same upper bound holds for $\widehat{\mathscr{R}}_n(\mathcal{G}_{adv}|_{\mathscr{S}})$. Hence, at least in terms of these upper bounds, sample complexity of learning with the adversarially-robust loss is the same as that of learning with the standard hinge loss in the current setup (this may not be true for non-linear functions).

Problem 2 [40 points] (Stability of k-Nearest Neighbors, Leave-One-Out) Let $\widehat{f}_n[\mathscr{S}]$ be the function $\mathcal{X} \to \mathcal{Y}$ obtained by training on data $\mathscr{S}$, and let $\widehat{f}_{n-1}[\mathscr{S}^{-i}]$ be the result of training on $\mathscr{S}^{-i} = \{Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n\}$ (ith point removed). As in class, we only consider symmetric algorithms. Define the shorthands $\widehat{g}_n(Z) = \ell(\widehat{f}_n(X),Y)$ and $Z = (X,Y)$.

We say that an algorithm is β-stable in L_1 sense if

$$\mathbb{E}\left|\widehat{g}_n[\mathscr{S}](Z) - \widehat{g}_{n-1}[\mathscr{S}^{-i}](Z)\right| \le \beta \tag{4.1}$$

where the expectation is over $\mathscr{S}$ and a new point Z. Suppose we always have $0 \le \widehat{g}_n \le 1$, for any n.

2.1 [10pts] Consider the Leave-One-Out estimate

$$\mathbf{L}^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \ell(\widehat{f}_{n-1}[\mathscr{S}^{-i}](X_i),Y_i) = \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_{n-1}[\mathscr{S}^{-i}](Z_i)$$

That is, we leave out one example from the dataset, then test on it, and repeat this for all examples. Show that $\mathbf{L}^{\mathrm{loo}}$ is an unbiased estimate of $\mathbb{E}\mathbf{L}(\widehat{f}_{n-1})$.

Unfortunately, unbiasedness is a weak notion since the variance can be large, rendering the estimate unreliable. The rest of the problem deals addresses this issue.

2.2 [10pts] Show that $\mathbf{L}^{\text{loo}}$ is an almost unbiased estimate of $\mathbb{E}\mathbf{L}(\widehat{f}_n)$ for an L_1-stable algorithm, in the sense that $\left|\mathbb{E}\left[\mathbf{L}(\widehat{f}_n) - \mathbf{L}^{\text{loo}}\right]\right| \leq \beta$. Hint: add and subtract $\mathbf{L}(\widehat{f}_{n-1})$ and use Jensen's inequality.

2.3 [10pts] Consider k-Nearest-Neighbor rule for classification. That is, $\widehat{f}_n[\mathscr{S}](x)$ outputs $-1/1$ according to the majority vote of the k nearest (to x) neighbors in the dataset, with ties broken in some manner. We are assuming $Y_i \in \{\pm 1\}$. Take the zero-one loss function $\mathbf{I}\left\{\widehat{f}_n(X) \neq Y\right\}$. Prove that k-Nearest-Neighbor rule is L_1 stable with $\beta = \frac{k}{n}$.

Hint: first relate the left-hand-side of (4.1) to

$$\mathbb{P}\left(\widehat{f}_n[\mathscr{S}](X) \neq \widehat{f}_{n-1}[\mathscr{S}^{-i}](X)\right)$$

where probability is over both $\mathscr{S}$ and X. Next argue by symmetry.

2.4 [10pts] It is possible to show that

$$\mathbb{E}\left(\mathbf{L}(\widehat{f}_n) - \mathbf{L}^{\text{loo}}\right)^2 \leq 3\mathbb{E}\left|\widehat{g}_n[\mathscr{S}](Z) - \widehat{g}_{n-1}[\mathscr{S}^{-i}](Z)\right| + \frac{1}{n}. \tag{4.2}$$

The proof is not too hard, but we will not do it here.

Use 2.2 and 2.3, and Chebyshev's inequality to deduce a statement of the form: with probability at least $1 - \delta$,

$$\left|\mathbf{L}^{\text{loo}} - \mathbf{L}(\widehat{f}_n)\right| \leq \Psi(\delta, n, k)$$

for k-NN. Find an appropriate upper bound Ψ.

We conclude that one can use the leave-one-out value for kNN to "reliably" estimate the true out-of-sample performance. We remark that kNN effectively produce a class of infinite VC dimension, and so stability analysis really saves the day.