CSC2547: Topics in Statistical Learning Theory	Winter 2019
Homework 3	
Ali Harakeh, Hitarth Choubisa	*Collaborators:*

1 Question 1:

1.1 Symmetrization for concentration:

We want to obtain a concentration for the empirical process $\hat{\xi}(f_+) = \frac{1}{n}\sum_{i=1}^{n} f_+(\langle x_i, \hat{\mu}\rangle)$ and we define $\xi(f_+) = E[f_+(\langle x, \mu\rangle)]$ where $\mu = E[x]$ and $|x|_2 < \kappa$ almost surely. Also, let $f, f_+ \in \mathcal{G}$ where $\mathcal{G}$ is the set of all functions which are L-lipschitz and bounded.
For $\hat{f} \in \mathcal{G}$, we have

$$|\hat{\xi}(\hat{f}) - \xi(f_+)| \leq |\hat{\xi}(\hat{f}) - \xi(\hat{f})| + |\xi(\hat{f}) - \xi(f_+)|$$

$$\leq \sup_{f\in\mathcal{G}} |\hat{\xi}(f) - \xi(f)| + 2B$$

Thus, we can write $P[\hat{\xi}(f) - \xi(f) \leq \epsilon] \geq P[\sup_{f\in\mathcal{G}} \hat{\xi}(f) - \xi(f) \leq \epsilon - 2B]$. So, we will try to bound $\Phi(S) = \sup_{f\in\mathcal{G}}(\hat{\xi}(f) - \xi(f))$ where S denotes a sample dataset over which we want to calculate the empirical estimate and S' is sample set same as S except for one different element x'_m in place of x_m.

Clearly,

$$|\Phi(S) - \Phi(S')| = |\sup_{f\in\mathcal{G}}(\hat{\xi}(f, S) - \xi(f)) - \sup_{h\in\mathcal{G}}(\hat{\xi}(h, S') - \xi(h))|$$

$$= |\hat{\xi}(f_*, S) - \xi(f_*) - \sup_{h\in\mathcal{G}}(\hat{\xi}(h, S') - \xi(h))|$$

where f_* maximizes the expression $\sup_{f\in\mathcal{G}}(\xi(\hat{f, S}) - \xi(f))$,

$$\leq |\hat{\xi}(f_*, S) - \xi(f_*) - \hat{\xi}(f_*, S') + \xi(f_*)|$$

since $\sup_{h\in\mathcal{G}}(\xi(\hat{h, S'}) - \xi(h)) \geq \hat{\xi}(f_*, S') - \xi(f_*)$

$$= |\hat{\xi}(f_*, S) - \hat{\xi}(f_*, S')| = |\hat{\xi}(S) - \hat{\xi}(S')|$$

$$= |\hat{\xi}(x_1, .., x_m, .., x_n) - \hat{\xi}(x_1, .., x'_m, ..x_n)| = \frac{1}{n}|\sum_{i=1}^{n} f(\langle x_i, \hat{\mu}\rangle) - \sum_{i=1}^{n} f(\langle x_i, \hat{\mu'}\rangle)|$$

The expression can be broken down into sum over $f(\langle x_i, \hat{\mu}\rangle) - f(\langle x_i, \hat{\mu'}\rangle)$ which in turn can be seen to be bounded as,

$$f(\langle x_i, \hat{\mu}\rangle) - f(\langle x_i, \hat{\mu'}\rangle) \leq |f(\langle x_i, \hat{\mu}\rangle) - f(\langle x_i, \hat{\mu'}\rangle)|$$

$$\leq L|\langle x_i, \hat{\mu'}\rangle - \langle x_i, \hat{\mu}\rangle|$$

$$= \frac{L}{n}|\langle x_i, x'_m\rangle - \langle x_i, x_m\rangle| \leq \frac{2L\kappa^2}{n} : (i \neq m)$$

$$= \frac{L}{n}||x'_m|^2 - |x_m|^2| \leq \frac{2L\kappa^2}{n}(i = m)$$

Thus, using $c_i = \frac{2L\kappa^2}{n}$ and applying McDearmid's inequality, we get,

$$P[\Phi(S) - E[\Phi(S)] \geq \epsilon] \leq exp\bigg\{\frac{-n\epsilon^2}{2L^2\kappa^4}\bigg\}$$

Next, we need to use Rademacher complexity to obtain bounds on $E[\Phi(S)]$. Consider a ghost sample $Q = (x'_1, x'_2, ..., x'_n)$. We get,

$$E[\Phi(S)] = E_S[\Phi(S)] = E_S[\sup_{f\in\mathcal{G}}(\hat{\xi}(f,S) - \xi(f))]$$

$$= E_S[\sup_{f\in\mathcal{G}}(\hat{\xi}(f,S) - E_Q[\hat{\xi}(f,Q)])]$$

$$= E_S[\sup_{f\in\mathcal{G}} E_Q[(\hat{\xi}(f,S) - \hat{\xi}(f,Q)])]$$

Using Jensen's inequality,

$$\leq E_{S,Q}[\sup_{f\in\mathcal{G}}(\hat{\xi}(f,S) - \hat{\xi}(f,Q))]$$

Using Symmetrization argument we get,

$$= E_{S,Q,\sigma}[\sup_{f\in\mathcal{G}} \frac{1}{n}\sum_i \sigma_i(f(x_i) - f(x'_i))]$$

where σ_i are Radamacher random variables.

$$\leq E_{S,Q,\sigma}[\sup_{f\in\mathcal{G}} \frac{1}{n}\sum_i \sigma_i f(x_i) + \sup_{f\in\mathcal{G}} \frac{1}{n}\sum_i (-\sigma_i) f(x'_i)]$$

$$\leq E_{S,Q,\sigma}[\sup_{f\in\mathcal{G}} \frac{1}{n}\sum_i f(x_i)] + E_{S,Q,\sigma}[\sup_{f\in\mathcal{G}} \frac{1}{n}\sum_i f(x'_i)]$$

$$= \mathcal{R}_m(\mathcal{G}) + \mathcal{R}_m(\mathcal{G}) = 2\mathcal{R}_m(\mathcal{G})$$

Thus, with probability at least $1 - exp\Big\{\frac{-n\epsilon^2}{2L^2\kappa^4}\Big\}$, we will have

$$\hat{\xi}(f) - \xi(f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i, \hat{\mu}) - E[f(x,\mu)] \leq 2\mathcal{R}_m(\mathcal{G}) + \epsilon$$

Also, the functional class $\mathcal{G}$ is composition of f and linear functions $\mathcal{F}$ constrained in l_2 ball of radius κ since μ should lie in the convex hull of x and $|x|_2 < \kappa$. Rademacher Complexity for the linear functions is

$$\mathcal{R}_m(\mathcal{F}) = E[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)] = \frac{1}{n}E[\sup_{w\in B_2(\kappa)} \sum_{i=1}^{n} \sigma_i \langle w, z_i\rangle] = \frac{1}{n}E[\sup_{w\in B_2(\kappa)} \langle w, \sum_{i=1}^{n} \sigma_i z_i\rangle]$$

$$\leq \frac{1}{n}E[\sup_{w\in B_2(\kappa)} |w| \cdot |\sum_{i=1}^{n} \sigma_i z_i|] \leq \frac{2\kappa}{n}E[|\sum_{i=1}^{n} \sigma_i z_i|]$$

$$= \frac{2\kappa}{n}E\left[\sqrt{\sum\sum \sigma_i\sigma_j\langle z_i, z_j\rangle}\right]$$

Using Jensen's inequality,

$$\leq \frac{2\kappa}{n}\sqrt{E\Big[\sum\sum \sigma_i\sigma_j\langle z_i, z_j\rangle\Big]} \leq \frac{2\kappa}{n}\sqrt{n\kappa^2} = \frac{2\kappa^2}{\sqrt{n}}$$

Using Talagrand's contraction, we get,

$$\mathcal{R}_m(\mathcal{G}) = \mathcal{R}_m(fo\mathcal{F}) = L \cdot \mathcal{R}_m(\mathcal{F}) = \frac{2\kappa^2 L}{\sqrt{n}}$$

Thus, we can summarize the final result as: With probability at least $1-\delta$, we will have

$$\hat{\xi}(f) \leq \xi(f) + 2\mathcal{R}_m(\mathcal{G}) + \kappa^2 L\sqrt{\frac{2log(1/\delta)}{n}}$$

Thus, with probability 1-δ we have,

$$\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \hat{\mu}\rangle) \leq E[f(x,\mu)] + \frac{4\kappa^2 L}{\sqrt{n}} + \kappa^2 L\sqrt{\frac{2log(1/\delta)}{n}}$$

1.2 Rademacher complexity of linear functions constrained in $\ell - 1$ ball:

Given:

- $\mathcal{F} : \{f(z) =< \beta, z >, ||\beta||_1 \leq r\}$
- $||z||_\infty \leq \kappa$

The empirical Rademacher complexity of $\mathcal{F}$ is written as:

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_\sigma[\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)] \\
&= \mathbb{E}_\sigma[\sup_{||\beta||_1 \leq r} \frac{1}{n} < \beta, \sum_{i=1}^{n}\sigma_i z_i >] \\
&= \frac{r}{n}\mathbb{E}_\sigma[\sup_{||\beta||_1 \leq 1} < \beta, \sum_{i=1}^{n}\sigma_i z_i >],
\end{aligned}$$

where we implicitly condition on z_i $\forall i = 1 \ldots n$. The unit $\ell - 1$ norm ball in R^d can be expressed as the convex hull of $2d$ points $[e_1, -e_1, \ldots, e_d, -e_d]$. Here e_j is the j-th standard basis vector.

Following from the above, and the fact that the Rademacher complexity of the convex hull of a set is the same as the set, $\mathcal{R}_n(\mathcal{F})$ can be expressed as:

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= r\mathcal{R}_n(cnvx([e_1, -e_1, \ldots, e_d, -e_d])) \\
&= r\mathcal{R}_n([e_1, -e_1, \ldots, e_d, -e_d]) \\
&= \frac{r}{n}\mathbb{E}[\sup_j \sum_{i=1}^{n} \sigma_i x_{ij}] \\
&\leq \frac{r}{n}\mathbb{E}[(\sup_j \sum_{i=1}^{n} \sigma_i^2 x_{ij}^2)^{\frac{1}{2}}] \quad (Jensen's) \\
&\leq \frac{r}{n}\mathbb{E}[(\sup_j \sum_{i=1}^{n} x_{ij}^2)^{\frac{1}{2}}] \\
&\leq \frac{r\kappa\frac{\sqrt{2\log d}}{\sqrt{n}}}{n} \quad (Massart's\ Finite\ Lemma) \\
&\leq \frac{r\kappa\sqrt{2\log d}}{\sqrt{n}}
\end{aligned}$$

The Rademacher complexity of the linear function class with parameters constrained in the $\ell - 1$ norm ball has an additional dependence on the dimension of data d over the the linear function class with parameters constrained in the $\ell - 2$ norm ball.

1.3 Generalization of binary classification:

Given:

- $(x_i, y_i) \in R^d$
- $||x_i||_\infty \leq \kappa$
- $y_i \in \{-1, +1\}$
- $\ell((x, y), \beta) = \min(2, \max(0, 1 - y < \beta, x >))$
- $||\beta||_1 \leq r$

First, let $z = y < \beta, x >$. The loss function can be rewritten as:

$$\ell(z) = \begin{cases} 0 & z \geq 1 \\ 1 - z & -1 \leq z \leq 1 \\ 2 & z \leq -1 \end{cases}$$

This function is 1-Lipschitz. In addition, it is bounded as: $0 \leq \ell(z) \leq 2$. Using RC theorem, with probability at least $1 - \delta$:

$$R(\hat{f}) - R(f^*) \leq 4\mathcal{R}_n(\mathcal{A}) + 2B\sqrt{\frac{\log \frac{2}{\delta}}{n}}$$

Let $\phi(z) = \min(2, \max(0, 1 - z))$. Then $\ell((x, y), \beta) = \phi(y < \beta, x >)$. Futhermore, let $G : \{s = (x, y) \rightarrow y.f(x).f \in \mathcal{F}\}$. Then $\mathcal{A} = \phi.G$.

By Talagrand's Contraction Principle: $\mathcal{R}_n(\mathcal{A}) \leq L.\mathcal{R}_n(\mathcal{F})$ where the Lipschitz constant $L = 1$. In question 1.2, it was shown that $\mathcal{R}_n(\mathcal{F}) \leq \frac{r\kappa\sqrt{2\log d}}{\sqrt{n}}$. Replace in the above equation, with the bound $B = 2$, the generalization bound on the empirical risk minimizer is:

$$R(\hat{f}) \leq R(f^*) + \frac{4r\kappa\sqrt{2\log d}}{\sqrt{n}} + 4\sqrt{\frac{\log \frac{2}{\delta}}{n}}$$

with probability at least $1 - \delta$.

1.4 Course Evaluation

Yes.