Lecturer: Prof. McDonald Scribe: Yisu Peng

16.1 Empirical risk

First, we introduce some notations,

- Data point $(x, y) \sim P \in \mathcal{P}$,
  where P is the joint distribution and $\mathcal{P}$ is the space of distributions.
- Loss function $\ell : \mathcal{Y} \times \mathcal{U} \to [0, 1]$,
  $\mathcal{Y}$ is the space of true labels and $\mathcal{U}$ is the space of predicted labels
- Hypothesis space $\mathcal{F} : \mathcal{X} \to \mathcal{U}$
- Data $Z^n = (Z_1, ..., Z_N) \overset{iid}{\sim} P$

Definition 16.1. *some function $f \in \mathcal{F}$, the empirical risk is*

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(y_i, f(x_i))] \tag{16.1}$$

Now we have,

$$\begin{aligned} \mathbb{E}\Big[\widehat{R}_n(f)\Big] &= \frac{1}{n} \sum_{i=1} n\mathbb{E}[\ell(y_i, f(x_i))] \\ &= \mathbb{E}[\ell(y_i, f(x_i))] \\ &= R(f) \end{aligned}$$

So based on the Hoeffding inequality,

$$P(|\widehat{R}_n(f) - R(f)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

$\implies$ w.p. $1 - 2e^{-2n\epsilon^2}$

$$|\widehat{R}_n(f) - R(f)| < \epsilon$$

We could find $\widehat{R}_n(f)$ for any $f \in \mathcal{F}$, then why not $\widehat{f_n} = \text{argmin}_{f \in \mathcal{F}} \widehat{R}_n(f)$

Here is how the logic goes: $f^* := \text{argmin}_{f \in \mathcal{F}} R(f)$, and

$$R(\widehat{f_n}) \approx \widehat{R}_n(\widehat{f_n}) \approx \widehat{R}_n(f^*) \approx R(f^*)$$

However, thid doesn't work. We need conditions on $\mathcal{P}, \mathcal{F}, \ell$. First we intoduce the "induced loss class",

$$\mathcal{L}(\mathcal{F}) = \{\ell(\cdot, f(\cdot)) : f \in \mathcal{F}\}$$

and

$$q(n,\epsilon) := \sup_{P\in\mathcal{P}} P^n(Z^n \in \mathbb{Z}^n : \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| \geq \epsilon)$$

This is corresponding to the worst case probability distribution over all $\mathcal{P}$ of "bad" samples (those with worst/large risk diviation over $\mathcal{F}$).

We say $\mathcal{L}(\mathcal{F})$ has "uniform convergence of empirical means" (UCEM) property w.r.t. $\mathcal{P}$, if

$$\lim_{n\to\infty} q(n,\epsilon) = 0 \qquad \forall \epsilon > 0$$

Theorem 16.2. *if $\mathcal{L}(\mathcal{F})$ has UCEM property the ERM is PAC.*

Proof.

$$\begin{aligned}
R(\widehat{f}_n) - R(f^*) &= R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) + \widehat{R}_n(\widehat{f}_n) - \widehat{R}_n(f^*) + \widehat{R}_n(f^*) - R(f^*) \\
\widehat{R}_n(\widehat{f}_n) - \widehat{R}_n(f^*) &\leq 0 \quad \text{by ERM} \\
R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) &\leq \sup_f (R(f) - R_n(f)) \leq \sup_f |R(f) - \widehat{R}_n(f)| \\
\widehat{R}_n(f^*) - R(f^*) &\leq \sup_f (R(f) - R_n(f)) \leq \sup_f |R(f) - \widehat{R}_n(f)|
\end{aligned}$$

So,

$$R(\widehat{f}) - R(f^*) \leq 2 \sup_f |R(f) - \widehat{R}_n(f)|$$

then with the UCEM property we have, $\exists n_0(\epsilon,\delta)$ s.t.

$$q(n,\epsilon/2) \leq \delta \qquad \forall n > n_0$$

■

16.2 ERM algorithm

Now, let's consider an algorithm,

$$\begin{aligned}
r_{\mathcal{A}}(n,\epsilon) &= \sup_{P\in\mathcal{P}} P^n(Z^n : R(\widehat{f}) \geq R(f^*) + \epsilon) \\
&\leq \sup_{P\in\mathcal{P}} P^n(Z^n : \sup_{f\in\mathcal{F}} |\widehat{R}_n(\widehat{f}) - R(f)| \geq \epsilon/2) \\
&= q(n,\epsilon/2) \leq \delta
\end{aligned}$$

UCEM is sufficient for the ERM algorithm to "work". Then our question is, when do we have UCEM? It turns out $\mathbb{E}\Big[\sup_{f\in\mathcal{F}} |R(f) - \widehat{R}_n(f)|\Big] = o(1)$ is sufficient for UCEM when $\ell : \mathcal{Y}\times\mathcal{U} \to [0,M]$.

Further more, we define

$$\Delta_n = \sup_f |R(f) - \widehat{R}_n(f)|$$

Then we have the following with McDiarmid's inequality,

$$P^n(\Delta_n - E\Delta_n \geq t) \leq e^{-2nt^2} \qquad \forall \ell \to [0,1]$$

What we want is $P^n(\Delta_n \geq \epsilon) \to 0$

$$\begin{aligned} P^n(\Delta_n \geq \epsilon) &= P^n(\Delta_n - E\Delta_n \geq \epsilon - E\Delta_n) \\ &\exists n \text{s.t.} E\Delta_n < \epsilon/2 \\ &\leq P^n(\Delta_n - E\Delta_n \geq \epsilon/2) \leq e^{-n\epsilon^2/2} \\ &\leq f \qquad \text{with n large enough} \end{aligned}$$

When is $E\Delta_n$ is small?

16.3 Rademacher complexity

Theorem 16.3. *ERM satisfies*

- $R(\widehat{f}_n) \leq R(f^*) + 2\Delta_n(\mathcal{F})$,
- $R(\widehat{f}_n) \leq \widehat{R}_n(\widehat{f}_n) + \Delta_n(\mathcal{F})$.
- $\Delta_n = \sup_{f\in\mathcal{F}} |R(f) - \widehat{R}_n(f)|$

We need to control Δ_n,

Definition 16.4.

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} |\frac{2}{n} \sum_{i=1} n\sigma_i f(Z_i)| \middle| Z^n \right] \tag{16.2}$$

σ_i's are Rademacher variables, $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$

- Think of this as "projection" of $\mathcal{F}$ on to Z^n
- "Correlation" of $\mathcal{F}$ with noise
- $\sup_f |f| \leq M$ or this doesn't work
- "empirical Rademacher complexity"

Definition 16.5. *Expected Rademacher complexity*

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{P^n}\left[\widehat{\mathfrak{R}}_n(\mathcal{F})\right]$$

Theorem 16.6.

$$\mathbb{E}[\Delta_n(\mathcal{F})] \leq \mathfrak{R}_n(\mathcal{F})$$

Proof. (by symmetrization) First we need Z^n and "ghost" samples $\widetilde{Z}^n$

$$\begin{aligned} \Delta_n(\mathcal{F}) &= \sup_{f\in\mathcal{F}} |R(f) - \widehat{R}_n(f)| \\ &= \sup_{f\in\mathcal{F}} \widetilde{\mathbb{E}}\left[|\widetilde{\widehat{R}}_n(f) - \widehat{R}_n(f)| \right] \\ &\leq \widetilde{E}\left[\sup_{f\in\mathcal{F}} |\widetilde{\widehat{R}}_n(f) - \widehat{R}_n(f)| \right] \end{aligned}$$

So,

$$\mathbb{E}_{Z^n}\Delta_n(\mathcal{F}) = \mathbb{E}_{Z^n}\mathbb{E}_{\widetilde{Z}^n}\left[\sup_{f\in\mathcal{F}}|\widetilde{\widehat{R}}_n(f) - \widehat{R}_n(f)|\right]$$

Because Z^n and $\widetilde{Z}^n$ are independent, with the symmetrization

$$\begin{aligned}\widetilde{\widehat{R}}_n(f) - \widehat{R}_n(f) &= \frac{1}{n}\sum_{i=1}^n f(\widetilde{Z}_i) - f(Z_i) \\ &\overset{dist}{=} \frac{1}{n}\sum_{i=1}^n f(Z_i) - f(\widetilde{Z}_i) \\ &\overset{dist}{=} \frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i) - f(\widetilde{Z}_i))\end{aligned}$$

So,

$$\begin{aligned}\mathbb{E}_{Z^n}\Delta_n(\mathcal{F}) &= \mathbb{E}_{Z^n}\mathbb{E}_{\widetilde{Z}^n}\left[|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i) - f(\widetilde{Z}_i))|\right] \\ &\leq \mathbb{E}_{Z^n\sigma}\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i))|\right] + \mathbb{E}_{\widetilde{Z}^n\sigma}\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(\widetilde{Z}_i))|\right] \\ &= 2\mathbb{E}_{Z^n\sigma}\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i))|\right] \\ &= \mathbb{E}_{Z^n}\left[\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}|\frac{2}{n}\sum_{i=1}^n \sigma_i(f(Z_i))|\,\middle|\, Z^n\right]\right] \\ &= \mathfrak{R}_n(\mathcal{F})\end{aligned}$$

■

References