Statistical Learning Theory

Section 4. General loss functions

Yongdai Kim

Review of Empirical Risk Minimization for classification

- In the previous lectures we have focused on binary losses for the classification problem and developed VC theory for it.
- In particular, we consider a classification function $h : \mathcal{X} \to \{0, 1\}$ and binary loss function todefine the risk

$$R(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{E}[\mathbb{I}(h(X) \neq Y)].$$

Review of Empirical Risk Minimization for classification

- In this section, we will consider a general loss function and a general regression model where Y is not necessarily a binary variable.
- Note that for the binary classification problem we used the followings:
  - Hoeffding's inequality: it requires boundedness of the loss functions.
  - Bounded difference inequality: again it requires boundedness of the loss functions.
  - VC theory: it requires binary nature of the loss function.

Review of Empirical Risk Minimization for classification

- There are many limitations of the VC theory.
- It would be hard to find the optimal classification. That is, the empirical risk minimization optimization, i.e.,

$$\min_h \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(h\left(X_i\right) \neq Y_i\right)$$

  is a difficult optimization.
- This is not suited for regression.
- Indeed, classification problem is a subset of regression problem as in regression the goal is to find $\mathbb{E}[Y \mid X]$ for a general Y (not necessarily binary).

Empirical Risk Minimization for general losses

- In this section, we assume that $Y \in [-1, 1]$ (this is not a limiting assumption as all the results can be derived for any bounded Y) and we have a regression problem where $(X, Y) \in \mathcal{X} \times [-1, 1]$.
- Most of the results that we preset here are the analogous to the results we had in binary classification.
- we will explain how to extend the techniques for the binary loss to general losses.

Empirical Risk Minimization for general losses

Loss functions

- In binary classification the loss function was $\mathbb{1}(h(X) \neq Y)$.
- Here, we replace this loss function by $\ell(Y, f(X))$, where $f \in \mathcal{F}$, $f : \mathcal{X} \to [-1, 1]$ is the regression functions.
- Examples of loss functions include
  - $\ell(a, b) = \mathbb{1}(a \neq b)$ (this is the classification loss function).
  - $\ell(a, b) = |a - b|$
  - $\ell(a, b) = (a - b)^2$
  - $\ell(a, b) = |a - b|^p, p \geq 1$

Empirical Risk Minimization for general losses

- We further assume that $0 \leq \ell(a, b) \leq 1$.
- Risk: the risk is the expectation of the loss function, i.e.

$$R(f) = \mathbb{E}_{X,Y}[\ell(Y, f(X))]$$

  where the joint distribution is typically unknown and it must be learned from data.
- Data: we observe a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent draws from a joint distribution $P_{X,Y}$, where $(X, Y) \in \mathcal{X} \times [-1, 1]$.
- We denote the data points by $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

Empirical Risk Minimization for general losses

- Empirical Risk: the empirical risk is defined as

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i, f\left(X_i\right)\right).$$

- The empirical risk minimizer denoted by $\hat{f}^{\text{erm}}$ (or $\hat{f}$) is defined as the minimizer of empirical risk, i.e.,

$$\underset{f\in\mathcal{F}}{\text{argmin}}\hat{R}_n(f)$$

Empirical Risk Minimization for general losses

- In order to control the risk of $\hat{f}$ we shall compare its performance with the following oracle:

$$\bar{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} R(f).$$

- Note that this is an oracle as in order to find it one need to have access to P_{XY} and then optimize $R(f)$ (we only observe the data D_n).
- Since $\hat{f}$ is the minimizer of the empirical risk minimizer, we have that $\hat{R}_n(\hat{f}) \leq \hat{R}_n(\bar{f})$, which leads to

$$\begin{aligned} R(\hat{f}) &\leq R(\hat{f}) - \hat{R}_n(\hat{f}) + \hat{R}_n(\hat{f}) - \hat{R}_n(\bar{f}) + \hat{R}_n(\bar{f}) - R(\bar{f}) + R(\bar{f}) \\ &\leq R(\bar{f}) + R(\hat{f}) - \hat{R}_n(\hat{f}) + \hat{R}_n(\bar{f}) - R(\bar{f}) \\ &\leq R(\bar{f}) + 2 \sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| \end{aligned}$$

Empirical Risk Minimization for general looses

- Therefore, the quantity of interest that we need to bound is

$$\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right|.$$

- Moreover, from the bounded difference inequality, we know that since the loss function $\ell(\cdot, -)$ is bounded by $1, \sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right|$ has the bounded difference property with $c_i = \frac{1}{n}$ for $i = 1, \ldots, n$.
- Hence, the bounded difference inequality establishes

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right|\right] \geq t\right] \leq \exp\left(\frac{-2t^2}{\sum_i c_i^2}\right) = \exp\left(-2nt^2\right).$$

Empirical Risk Minimization for general losses

- In turn, we have

$$\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| \right] + \sqrt{\frac{\log(1/\delta)}{2n}}, \text{ w.p. } 1 - \delta.$$

- As a result we only need to bound the expectation

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| \right].$$

Symmetrization and Rademacher Complexity

- Similar to the binary loss case, we first use symmetrization technique and then introduce Rademacher random variables.
- Let $D_n = \{(X_1, Y_1), \ldots (X_n, Y_n)\}$ be the sample set and define an independent sample (ghost sample) with the same distribution denoted by $D_n' = \{(X_1', Y_1'), \ldots (X_n', Y_n')\}$ (for each $i, (X_i', Y_i')$ is independent from D_n with the same distribution as of (X_i, Y_i)).
- Also, let $\sigma_i \in \{-1, +1\}$ be i.i.d. Rad $\left(\frac{1}{2}\right)$ random variables independent of D_n and D_n'.

Symmetrization and Rademacher Complexity

Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i,f\left(X_i\right)\right)-\mathbb{E}\left[\ell\left(Y_i,f\left(X_i\right)\right)\right]\right|\right]\\
&=\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i,f\left(X_i\right)\right)-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i',f\left(X_i'\right)\right)\mid D_n\right]\right|\right]\\
&=\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i,f\left(X_i\right)\right)-\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i',f\left(X_i'\right)\right)\mid D_n\right]\right|\right]\\
&\overset{(a)}{\leq}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i,f\left(X_i\right)\right)-\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i',f\left(X_i'\right)\right)\right|\mid D_n\right]\right]\\
&\leq\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i,f\left(X_i\right)\right)-\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i',f\left(X_i'\right)\right)\right|\right]\\
&\overset{(b)}{=}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\ell\left(Y_i,f\left(X_i\right)\right)-\ell\left(Y_i',f\left(X_i'\right)\right)\right)\right|\right]\\
&\overset{(c)}{\leq}2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell\left(Y_i,f\left(X_i\right)\right)\right|\right]\\
&\leq 2\sup_{D_n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mid\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell\left(y_i,f\left(x_i\right)\right)\right]\Big]
\end{aligned}
$$

where

- (a) follows from Jensen's inequality with convex function $f(x) = |x|$,
- (b) follows from the fact that (X_i, Y_i) and (X_i', Y_i') has the same distributions,
- (c) follows from triangle inequality.

Symmetrization and Rademacher Complexity

- Rademacher complexity of a class $\mathcal{F}$ of functions for a given loss function $\ell(\cdot, \cdot)$ and samples D_n is defined as

$$\mathcal{R}_n(\ell \circ \mathcal{F}) = \sup_{D_n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell\left(y_i, f\left(x_i\right)\right) \right| \right].$$

- Therefore, we have

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \mid \frac{1}{n} \sum_{i=1}^{n} \ell\left(Y_i, f\left(X_i\right)\right) - \mathbb{E}\left[\ell\left(Y_i, f\left(X_i\right)\right)\right]\right]\right] \leq 2\mathcal{R}_n(\ell \circ \mathcal{F})$$

and we only require to bound the Rademacher complexity.

Finite class of functions

- Suppose that the class of functions $\mathcal{F}$ is finite.
- We have the following bound:

Theorem
Assume that $|\mathcal{F}|$ *is finite and that* ℓ *takes values in* $[0, 1]$. *Then, we have*

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}$$

Finite class of functions

Proof

From the previous lecture, for $B \subseteq \mathbb{R}^n$, we have that

$$\mathcal{R}_n(B) = \mathbb{E}\left[\max_{b \in B}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i b_i\right|\right] \leq \max_{b \in B}|b|_2 \frac{\sqrt{2\log(2|B|)}}{n}$$

Here, we have

$$B = \left\{ \begin{pmatrix} \ell\left(y_1, f\left(x_1\right)\right) \\ \vdots \\ \ell\left(y_n, f\left(x_n\right)\right) \end{pmatrix}, f \in \mathcal{F} \right\}$$

Since ℓ takes values in $[0,1]$, this implies $B \subseteq \left\{b : |b|_2 \leq \sqrt{n}\right\}$.
Plugging this bound in the above inequality completes the proof. □

Covering numbers

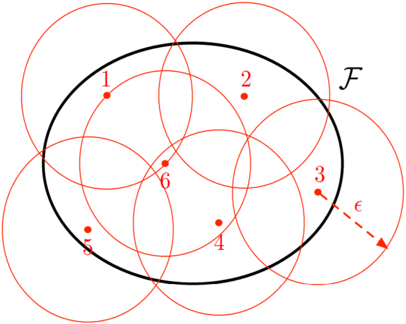
- Recall that for the classification problem, we had $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$.
- We have seen that the cardinality of the set $\{(f(x_1), \dots, f(x_n)), f \in \mathcal{F}\}$ plays an important role in bounding the risk of f^{erm} .
- However, this set might be uncountable and thus we need to introduce a measure of the size of the set.
- To this end we will define covering numbers, which basically plays the role of VC dimension in the classification.

Covering numbers

Definition
Given a set of functions $\mathcal{F}$ and a pseudo metric d on $\mathcal{F}((\mathcal{F}, d)$ is a metric space) and $\varepsilon > 0$. An ε-net of $(\mathcal{F}, d)$ is a set V such that for any $f \in \mathcal{F}$, there exists $g \in V$ such that $d(f, g) \leq \varepsilon$. Moreover, the covering numbers of $(\mathcal{F}, d)$ are defined by

$$N(\mathcal{F}, d, \varepsilon) = \inf\{|V| : V \text{ is an } \varepsilon \text{ -net } \}$$

Covering numbers



- For instance, for the $\mathcal{F}$ shown in the above figure, the set of points $\{1, 2, 3, 4, 5, 6\}$ is a covering.
- However, the covering number is 5 as point 6 can be removed from V and the resulting points are still a covering.

Covering numbers

Definition
Given $x = (x_1, \ldots, x_n)$, the conditional Rademacher average of a class of functions $\mathcal{F}$ is defined as

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f\left(x_i\right) \right|\right]$$

Covering numbers

- Note that when we apply the above result to learning theory at the end of this section, we will take x_i to be (x_i, y_i) and $\mathcal{F}$ to be $\ell \circ \mathcal{F}$.
- We define the empirical l_1 distance as

$$d_1^x(f, g) = \frac{1}{n} \sum_{i=1}^{n} \left| f\left(x_i\right) - g\left(x_i\right) \right|.$$

Covering numbers

Theorem
If $0 \leq f \leq 1$ *for all* $f \in \mathcal{F}$, *then for any* $x = (x_1, \ldots, x_n)$, *we have*

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq \inf_{\varepsilon \geq 0} \left\{ \varepsilon + \sqrt{\frac{2 \log \left(2N\left(\mathcal{F}, d_1^x, \varepsilon\right)\right)}{n}} \right\}$$

Proof

Fix $x = (x_1, \dots, x_n)$ and $\varepsilon > 0$. Let V be a minimal ε-net of $(\mathcal{F}, d_1^x)$. Thus, by definition we have that $|V| = N(\mathcal{F}, d_1^x, \varepsilon)$. For any $f \in \mathcal{F}$, define $f^\circ \in V$ such that $d_1^z(f, f^\circ) \leq \varepsilon$.

Proof

We have that

$$\begin{aligned}
\hat{\mathcal{R}}_n^x(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(x_i\right)\right|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mid \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(f\left(x_i\right)-f^{\circ}\left(x_i\right)\right)\right]\Bigg] \\
&+ \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f^{0}\left(x_i\right)\right|\right] \\
&\leq \varepsilon+\mathbb{E}\left[\max_{f\in V} \mid \frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(x_i\right)\right]\Bigg] \\
&\leq \varepsilon+\sqrt{\frac{2\log(2|V|)}{n}} \\
&= \varepsilon+\sqrt{\frac{2\log\left(2N\left(\mathcal{F}, d_1^x, \varepsilon\right)\right)}{n}}
\end{aligned}$$

Proof

Since the previous bound holds for any ε, we can take the infimum over all $\varepsilon \geq 0$ to obtain

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq \inf_{\varepsilon \geq 0} \left\{ \varepsilon + \sqrt{\frac{2 \log \left(2N \left(\mathcal{F}, d_1^x, \varepsilon \right) \right)}{n}} \right\}$$

The previous bound clearly establishes a trade-off because as ε decreases $N \left(\mathcal{F}, d_1^x, \varepsilon \right)$ increases. $\square$

Computing covering numbers

For any $p \geq 1$, define

$$d_p^x(f, g) = \left(\frac{1}{n} \sum_{i=1}^{n} \left| f\left(x_i\right) - g\left(x_i\right) \right|^p \right)^{\frac{1}{p}},$$

and for $p = \infty$, define

$$d_\infty^x(f, g) = \max_i \left| f\left(x_i\right) - g\left(x_i\right) \right|.$$

Computing covering numbers

- Using the previous theorem, in order to bound $\mathcal{R}_n^x$ we need to bound the covering number with d_1^x norm.
- We will show that it is sufficient to bound the covering number for the infinity norm.
- In order to show this, we will compare the covering number of the norms $d_p^x(f,g) = \left(\frac{1}{n}\sum_{t=1}^{n}|f(x_i) - g(x_i)|^p\right)^{\frac{1}{p}}$ for $p \geq 1$ and conclude that a bound on $N(\mathcal{F}, d_\infty^x, \varepsilon)$ implies a bound on $N\left(\mathcal{F}, d_p^x, \varepsilon\right)$ for any $p \geq 1$.

Proposition
For any $1 \leq p \leq q$ *and* $\varepsilon > 0$, *we have that*

$$N\left(\mathcal{F}, d_p^x, \varepsilon\right) \leq N\left(\mathcal{F}, d_q^x, \varepsilon\right)$$

Proof.
This is because $d_p^x(f) \leq d_q^x(f)$ for any $p \leq q$ (from HW). □

- Using this propositions we only need to bound $N(\mathcal{F}, d_\infty^x, \varepsilon)$.

Computing covering numbers

Example

- Let the function class be
  $\mathcal{F} = \left\{ f(x) = \langle f, x \rangle, f \in B_\infty^d, x \in B_1^d \right\}$, where
  $B_p^d = \left\{ x \in \mathbb{R}^d : |x|_p \leq 1 \right\}$ and $|x|_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$.
- Note that $|f(x)| \leq 1$ (HW).
- It can be shown that

$$N(\mathcal{F}, d_1^x, \epsilon) \leq c/\epsilon^d$$

  for a certain constant $c > 0$ (HW).

Computing covering numbers

Example (continue)

- Hence, we have

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq \inf_{\varepsilon \geq 0} \left\{ \varepsilon + \sqrt{\frac{2 \log \left(c/\varepsilon^d\right)}{n}} \right\}.$$

- Optimizing over all choices of ε gives

$$\varepsilon^* = c\sqrt{\frac{d \log(n)}{n}} \Rightarrow \quad \hat{\mathcal{R}}_n^x(\mathcal{F}) \leq c\sqrt{\frac{d \log(n)}{n}}.$$

Chaining: A techniuqe to derive a tighter upper bound

Theorem
Assume that $|f| \leq 1$ for all $f \in \mathcal{F}$. Then

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq \inf_{\varepsilon > 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{1} \sqrt{\log\left(N\left(\mathcal{F}, d_2^x, t\right)\right)} dt \right\}$$

(Note that the integrand decays with t.)

Chaining: A techniuqe to derive a tighter upper bound

- Let the function class be
  $\mathcal{F} = \left\{ f(x) = \langle f, x \rangle, f \in B_2^d, x \in B_2^d \right\}.$
- It can be shown (HW) that

$$N\left(\mathcal{F}, d_2^x, \varepsilon\right) \leq c/\varepsilon^d.$$

- Hence, we have

$$\mathcal{R}_n^x(\mathcal{F}) \leq \inf_{\varepsilon > 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^1 \sqrt{\log\left((c'/t)^d \right)}\, dt \right\}.$$

- Since $\int_0^1 \sqrt{\log(c/t)} dt = \bar{c}$ is finite, we then have

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq 12\bar{c}\sqrt{d/n}.$$

- Using chaining, we've been able to remove the log factor!

Back to Learning

- Recall that we want to bound

$$\mathcal{R}_n(\ell \circ \mathcal{F}) = \sup_{(x_1,y_1),\dots,(x_n,y_n)} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell\left(y_i, f\left(x_i\right)\right) \right| \right].$$

- We consider $\hat{R}_n^x(\Phi \circ \mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{t=1}^{n} \sigma_i \Phi \circ f\left(x_i\right) \right| \right]$ for some L -Lipschitz function Φ, that is $|\Phi(a) - \Phi(b)| \leq L|a-b|$ for all $a, b \in [-1, 1]$. We have the following lemma.

Back to Learning

Theorem (Contraction Inequality)
Let Φ be L -Lipschitz and such that $\Phi(0) = 0$, then

$$\hat{R}_n^x(\Phi \circ \mathcal{F}) \leq 2L \cdot \mathcal{R}_n^x(\mathcal{F})$$

- As a final remark, note that requiring the loss function to be Lipschitz prohibits the use of $\mathbb{R}$ -valued loss functions, for example $\ell(Y, \cdot) = (Y - \cdot)^2$.
- Examples of Lipschitz losses are the logistic loss, hinge loss and absolute loss.