STA 250: Theoretical Foundations for Machine Learning

Lecture 2: Empirical Risk Minimization

Dogyoon Song

Spring 2025, UC Davis

Recap: Supervised learning

Given

- Unknown distribution μ on $\mathcal{X} \times \mathcal{Y}$
- A sample $\mathcal{D}_n = \{(x_i, y_i) : i \in [n]\}$ from μ
- A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

We want to design $\texttt{Alg} : \mathcal{D}_n \mapsto f$ such that the excess risk $R(f) - R^*$ is small

Challenge: we do not have access to μ but $\mathcal{D}_n$!

Idea: Use a sample-analogue estimator of the risk function

Agenda

- Empirical risk minimization
- Asymptotic analysis
- Non-asymptotic analysis

Empirical risk minimization

Empirical risk: $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$

Typically, we consider a parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, and try to find

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \hat{R}(f_\theta)$$

We want guarantees for the excess risk $R(f_{\hat{\theta}}) - R^*$

Notation: From now on, I may use $R(\theta)$ to denote $R(f_\theta)$ and $\hat{R}(\theta)$ for $\hat{R}(f_\theta)$ for brevity

Asymptotic analysis

Asymptotic guarantees

Assume that (1) θ^* is the unique minimizer of $R(\theta)$ and (2) $\nabla^2 R(\theta^*) \succ 0$. Then

- $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, H^{-1} G H^{-1})$
- $n(R(\hat{\theta}) - R(\theta^*)) \xrightarrow{d} \frac{1}{2}\|S\|^2$ where $S \sim \mathcal{N}(0, H^{-1/2} G H^{-1/2})$

where $G = \mathrm{Cov}(\nabla \ell(f_{\theta^*}(x), y)) = \mathbb{E}[\nabla \ell(f_{\theta^*}(x), y) \cdot \nabla \ell(f_{\theta^*}(x), y)^\top]$ and $H = \nabla^2 R(\theta^*)$

This yields an asymptotic upper bound on the excess risk: $R(\hat{\theta}) - R^* \leq \frac{c}{n} + o\left(\frac{1}{n}\right)$

- Benefits: (1) asymptotic normality with explicit distribution; (2) fast rate of $O(1/n)$
- Drawbacks: asymptotic...

Example: Well-specified linear regression

- $G = \sigma^2 H \implies$ excess risk $\sim \frac{\sigma^2 d}{2n}$

Proof sketch

Step 1: Taylor expansion

$$
\begin{aligned}
0 = \hat{R}(\hat{\theta}) &= \nabla \hat{R}(\theta^*) + \nabla^2 \hat{R}(\theta^*)(\hat{\theta} - \theta^*) + O(\|\hat{\theta} - \theta^*\|^2) \\
&\implies \hat{\theta} - \theta^* \approx -\left[\nabla^2 \hat{R}(\theta^*)\right]^{-1} \cdot \nabla \hat{R}(\theta^*) \\
&\implies \sqrt{n}(\hat{\theta} - \theta^*) \approx -\left[\nabla^2 \hat{R}(\theta^*)\right]^{-1} \cdot \sqrt{n} \nabla \hat{R}(\theta^*)
\end{aligned}
$$

Step 2: Central limit theorem

- Recall that $\nabla \hat{R}(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(f_{\theta^*}(x_i), y_i)$
- Due to linearity, $\mathbb{E}\left[\nabla \hat{R}(\theta^*)\right] = \nabla R(\theta^*) = 0$
- By CLT,

$$
\sqrt{n}(\nabla \hat{R}(\theta^*) - \nabla R(\theta^*)) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}(\nabla \ell(f_{\theta^*}(x), y)))
$$

- Since $\nabla^2 \hat{R}(\theta^*) \xrightarrow{p} \nabla^2 R(\theta^*)$ by LLN,

$$
\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \nabla^2 R(\theta^*)^{-1} \mathcal{N}(0, G) \overset{d}{=} \mathcal{N}(0, H^{-1} G H^{-1})
$$

Non-asymptotic analysis

We decompose excess risk as

$$R(\hat{\theta}) - R^* = \underbrace{\left(R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \right)}_{\text{estimation error}} + \underbrace{\left(\inf_{\theta \in \Theta} R(\theta) - R^* \right)}_{\text{approximation error}}$$

Our main focus is to control the estimation error; we may discuss approximation error later

Uniform convergence: To this end, we will utilize a bound of the form

$$\Pr\left(\sup_{\theta \in \Theta} \left| \hat{R}(\theta) - R(\theta) \right| \leq \epsilon \right) \geq 1 - \delta$$

Note that

- For a single fixed θ, this inequality might look trivial by the LLN
- However, a uniform bound is not straightforward
- Note that $\epsilon = \epsilon_\delta(\Theta)$ depends on the hypothesis class Θ

From uniform convergence to generalization bound

Suppose we have uniform convergence

Then

$$
\begin{aligned}
R(\hat{\theta}) - R(\theta^*) &= R(\hat{\theta}) - \hat{R}(\hat{\theta}) + \underbrace{\hat{R}(\hat{\theta}) - \hat{R}(\theta^*)}_{\leq 0 \ \because \hat{\theta} \text{ minimizes } \hat{R}} + \hat{R}(\theta^*) - R(\theta^*) \\
&\leq \left|R(\hat{\theta}) - \hat{R}(\hat{\theta})\right| + \left|\hat{R}(\theta^*) - R(\theta^*)\right| \\
&\leq 2 \sup_{\theta \in \Theta} \left|\hat{R}(\theta) - R(\theta)\right|
\end{aligned}
$$

Now the question reduces to establishing uniform convergence (for Θ of interest)

Finite hypothesis class

Suppose that (1) $|\Theta| < \infty$ and (2) $\ell(f_\theta(x), y) \in [0, B]$

Then

$$\Pr\left(\sup_{\theta\in\Theta} |\hat{R}(\theta) - \hat{R}(\theta)| > t\right) \leq \sum_{\theta\in\Theta} \Pr\left(|\hat{R}(\theta) - \hat{R}(\theta)| > t\right) \qquad \because \text{union bound}$$

$$\leq 2|\Theta| \cdot \exp\left(-\frac{2nt^2}{B^2}\right) \qquad \because \text{Hoeffding ineq}$$

This implies that with probability at least $1 - \delta$,

$$\sup_{\theta\in\Theta} |\hat{R}(\theta) - \hat{R}(\theta)| \leq \underbrace{B\sqrt{\frac{\log(2|\Theta|)}{2n}}}_{\text{overhead for uniform control}} + B\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}$$

Remarks: (1) overhead for uniform control; (2) explicit bound but slow rate $O(n^{-1/2})$

Infinite hypothesis class

Suppose that (1') Θ is compact and (2, 3) $\ell(f_\theta(x), y) \in [0, B]$, and L-Lipschitz (w.r.t. θ)

Definition (ϵ-net)

Let (T, d) be a metric space. Let $K \subseteq T$ and $\epsilon > 0$. A subset $N \subseteq S$ is an ϵ-**net** of S if for all $x \in S$, there exists $x' \in N$ such that $d(x, x') \leq \epsilon$.

Let N be an ϵ-net of Θ. For any $\theta \in \Theta$, there exists $\theta' \in N$ such that

$$\begin{aligned} |\hat{R}(\theta) - R(\theta)| &\leq |\hat{R}(\theta) - \hat{R}(\theta')| + |\hat{R}(\theta') - R(\theta')| + |R(\theta') - R(\theta)| \\ &\leq 2L \underbrace{\|\theta - \theta'\|}_{\leq \epsilon} + |\hat{R}(\theta') - R(\theta')| \end{aligned}$$

It follows that with probability at least $1 - \delta$,

$$\sup_{\theta \in \Theta} |\hat{R}(\theta) - \hat{R}(\theta)| \leq 2L\epsilon + B\sqrt{\frac{\log(2N(\Theta, \epsilon))}{2n}} + B\sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}$$

where $N(\Theta, \epsilon)$ is the covering number of Θ