FoSML: Lecture 11

- It may be the case that ERM involves NP-hard computations if, for example, we need to do a brute force search over the hypothesis class.
- However, the NP-hardness sometimes results from the non-convexity of the 0-1 loss function. We can address this by developing a convex surrogate loss which upper bounds the 0-1 loss function and has the advantage of being ~~complex~~ convex.
- Let us consider a hypothesis class of $\mathbb{R}$-valued functions $\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R}\}$. We can define a binary classifier given h as $\operatorname{sgn}(h)$, i.e. $f_h(x) = \begin{cases} +1, & h(x) \geq 0 \\ -1, & h(x) < 0 \end{cases}$.

- The loss of h at $x \in \mathcal{H}$ is the binary classification error of f_h:

$$\mathbb{1}_{f_h(x) \neq y} = \mathbb{1}_{yh(x) \leq 0} + \mathbb{1}_{h(x)=0 \wedge y=-1}$$

$yh(x) \leq 0$: $y = 1$ and $h(x) \leq 0$, or $y = -1$ and $h(x) \geq 0$

$h(x) = 0 \wedge y = -1$: boundary case

- The key observation is we can upper bound this as

$$\mathbb{1}_{yh(x) < 0} + \mathbb{1}_{h(x)=0 \wedge y=-1} \leq \mathbb{1}_{yh(x) \leq 0}$$ ✡ Key inequality!

- We will denote by $R(h)$ the expected error of h:

$$R(h) = \mathbb{E}_{(x,y) \sim D}\left(\mathbb{1}_{f_h(x) \neq y}\right)$$

- For any $x \in \mathcal{X}$, let $\eta(x)$ denote $\eta(x) = \underbrace{\mathbb{P}(y=+1 \mid x)}_{\text{stochastic setting} \Rightarrow \neq 1,0 \text{ in general}}$. Let D_X denote the marginal distribution over $\mathcal{X}$.

- We can decompose the generalization error as $R(h)$

$$= \mathop{\mathbb{E}}_{(x,y)\sim D} \left(\mathbb{1}_{f_h(x) \neq y} \right)$$

$$= \mathop{\mathbb{E}}_{x \sim D_X} \left[\eta(x) \mathbb{1}_{h(x)<0} + (1-\eta(x)) \mathbb{1}_{h(x)\geq 0} + (1-\eta(x)) \cdot \mathbb{1}_{h(x)=0} \right]$$

$$= \mathop{\mathbb{E}}_{x \sim D_X} \left[\eta(x) \mathbb{1}_{h(x)<0} + (1-\eta(x)) \mathbb{1}_{h(x)\geq 0} \right]$$

- Minimizing over h, we see the Bayes classifier h^* satisfies

$$h^*(x) = \begin{cases} -1, & \eta(x) \geq \frac{1}{2} \\ -1, & \eta(x) < \frac{1}{2} \end{cases} \Rightarrow \quad h^*(x) = \eta(x) - \frac{1}{2}.$$

- Let $R^* = R(h^*)$ denote the Bayes error.

Lemma: For any $h : \mathcal{X} \to \mathbb{R}$ satisfies $R(h) - R^* = 2 \mathop{\mathbb{E}}_{x \sim D_X} \left(|h^*(x)| \mathbb{1}_{h(x)h^*(x) \leq 0} \right)$.

Proof: For any h, we can write

$$R(h) = \mathop{\mathbb{E}}_{x \sim D_X} \left(\eta(x) \mathbb{1}_{h(x)<0} + (1-\eta(x)) \mathbb{1}_{h(x)\geq 0} \right)$$

$$= \mathop{\mathbb{E}}_{x \sim D_X} \left(\eta(x) \mathbb{1}_{h(x)<0} + (1-\eta(x)) \left(1 - \mathbb{1}_{h(x)<0}\right) \right)$$

$$= \mathop{\mathbb{E}}_{x\sim D_X}\left(\left[2\eta(x)-1\right]\mathbb{1}_{h(x)<0} + \left(1-\eta(x)\right)\right)$$

$$= \mathop{\mathbb{E}}_{x\sim D_X}\left(2h^*(x)\mathbb{1}_{h(x)<0} + \left(1-\eta(x)\right)\right).$$

We then compute $R(h) - R^*$

$$= \mathop{\mathbb{E}}_{x\sim D_X}\left(2h^*(x)\mathbb{1}_{h(x)<0} + \left(1-\eta(x)\right)\right) - \mathop{\mathbb{E}}_{x\sim D_X}\left(h^*(x)\right)$$

use definition of $h^*(x)$: $\eta(x) - \frac{1}{2}$

$$= \mathop{\mathbb{E}}_{x\sim D_X}\left(2|h^*(x)|\left(\mathbb{1}_{h(x)\le 0} - \mathbb{1}_{h^*(x)\le 0}\right)\right)$$

$$= 2\mathop{\mathbb{E}}_{x\sim D_X}\left(|h^*(x)|\,\mathrm{sgn}(h^*(x))\,\mathbb{1}_{(h(x)h^*(x)\le 0)\wedge((h(x),h^*(x))\neq(0,0))}\right)$$

$$= 2\mathop{\mathbb{E}}_{x\sim D_X}\left(|h^*(x)|\,\mathbb{1}_{h(x)h^*(x)\le 0}\right)$$

- Now, let $\Phi: \mathbb{R}\to\mathbb{R}$ be convex and non-decreasing. Suppose that for any $u\in\mathbb{R}$, $\mathbb{1}_{u\le 0} \le \Phi(-u)$. Define the Φ-loss of $h: \mathcal{X}\to\mathbb{R}$ at a point $(x,y)\in \mathcal{X}\times\{-1,1\}$ as $\Phi(-yh(x))$. Then its expected Φ loss is

$$\mathcal{L}_\Phi(h) = \mathop{\mathbb{E}}_{(x,y)\sim D}\left(\Phi(-yh(x))\right)$$

$$= \mathop{\mathbb{E}}_{x\sim D_X}\left(\eta(x)\,\Phi(-h(x)) + (1-\eta(x))\,\Phi(h(x))\right).$$

- By construction, $R(h) \le \mathcal{L}_\Phi(h)$.

- For any $x \in \mathcal{X}$, let $u \mapsto L_\Phi(x,u)$ be the function defined for $u \in \mathbb{R}$ by $L_\Phi(x,u) = \eta(x)\,\Phi(-u) + (1-\eta(x))\,\Phi(u)$. Then

$$\mathcal{L}_\Phi(h) = \mathop{\mathbb{E}}_{x \sim D_X} \left(L_\Phi(x, h(x)) \right).$$

- Φ is convex $\Rightarrow$ $u \mapsto L_\Phi(x,u)$ is convex, being a sum of two convex functions.
- Let $h_\Phi^* : \mathcal{X} \to [-\infty, \infty]$ be the Bayes solution for the loss function L_Φ,

i.e. $h_\Phi^*(x) = \operatorname*{argmin}_{u \in [-\infty, \infty]} \eta(x)\,\Phi(-u) + (1-\eta(x))\,\Phi(u)$.

- When $\eta(x) = 0$, $h_\Phi^*(x)$ is a minimizer of $\Phi(u)$, and since Φ is non-decreasing we can set $h_\Phi^*(x) = -\infty$. Similarly, when $\eta(x) = 1$, we can set $h_\Phi^*(x) = +\infty$.
- When $\eta(x) = \frac{1}{2}$, $L_\Phi(x,u) = \frac{1}{2}\left[\Phi(u) + \Phi(-u)\right]$ so that by convexity, $L_\Phi(x,u) \geq \Phi(0)$, so we can choose $h_\Phi^*(x) = 0$ in this case.
- In between these extremal values, we may have non-uniqueness of solutions. Pick one arbitrarily.
- Let $\mathcal{L}_\Phi^*$ be the Φ-loss of h_Φ^*, i.e. $\mathop{\mathbb{E}}_{(x,y) \sim D} \left(\Phi(-y h_\Phi^*(x)) \right)$.
- Our goal is the following theorem, which estimates generalization error in terms of $\mathcal{L}_\Phi$:

Theorem: Let ϕ be convex and non-decreasing. Suppose $\exists s \geq 1$ and $c > 0$ s.t. for all $x \in \mathcal{X}$, $|h^*(x)|^s = |\eta(x) - \frac{1}{2}|^s \leq c^s \left[L_\phi(x,0) - L_\phi(x, h^*_\phi(x)) \right]$.

Then $\forall$ hypotheses h, $R(h) - R^* \leq 2c \left[\mathcal{L}_\phi(h) - \mathcal{L}^*_\phi \right]^{\frac{1}{s}}$.

- Proof next time, but we remark that these assumptions hold broadly:

ex: $\phi(u) = \max(1+u, 0) \Rightarrow s=1,\ c = \frac{1}{2}$

$\phi(u) = \exp(u) \Rightarrow s=2,\ c = 1/\sqrt{2}$

$\phi(u) = \log_2(1+e^u) \Rightarrow s=2,\ c = 1/\sqrt{2}$.

$\vdots$