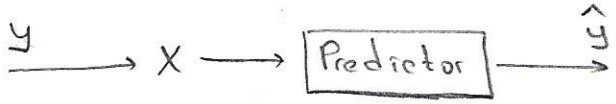
Lecture 6:



Data:

$$\{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d}}{\sim} P$$

(P is the true unknown distribution from which nature generates these samples)

Hypothesis class: $\mathcal{H} = \{h : h : \mathcal{X} \to \mathcal{Y}\}$
the class of functions we can use as predictors

Loss function: $\ell(y, \hat{y}) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$
quantifies the error associated with predicting y as $\hat{y}$.

Empirical loss of $h \in \mathcal{H}$:

$$\mathcal{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Population loss of $h \in \mathcal{H}$:

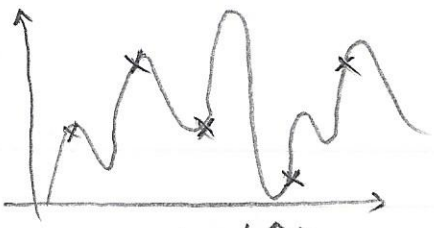
$$\mathcal{L}(h) = \mathbb{E}_P[\ell(h(x), y)]$$

Empirical Risk Minimization (ERM):

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \ \mathcal{L}_n(h)$$

→ The ERM solution or the empirical solution.

Key question: How can we ensure that $L(\hat{h})$ is close to $L_n(\hat{h})$?



$L_n(\hat{h}) = \hat{\epsilon}$

$z_i^{(h)} \overset{i.i.d}{\sim} P_Z$ $\quad$ $\mathbb{E}[z_i^{(h)}]$

$$L_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i) \xrightarrow[n\to\infty]{} \mathbb{E}_P\left[\ell(h(x), y)\right] = L(h)$$

$(x_i, y_i) \overset{i.i.d}{\sim} P$ $\quad$ because of LLN.

ERM: We approximate the true expectation of $z_i = \ell(h(x), y)$ with the empirical or sample mean.

Uniform Convergence: Show that if take n iid. samples from P and compute $L_n(h)$ for all $h \in \mathcal{H}$, then $L_n(h) \approx L(h)$ for all $h \in \mathcal{H}$.

Hoeffding's Inequality: Let $X_1, X_2, \dots X_n$ i.i.d.

$\mathbb{E}[X_i] = \mu$ & $a \le X_i \le b$, for any $t \ge 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge t\right) \le 2e^{-2nt^2/(b-a)^2}$$

Assume $0 \le \ell(\hat{y}, \hat{y}) \le 1$ (e.g. 0-1 loss

or $\ell(y, \hat{y}) = (y - \hat{y})^2$ $\quad$ $0 \le y, \hat{y} \le 1$.

Then for any $h \in \mathcal{H}$.

$$P\left(|L_n(h) - L(h)| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

We want $|L_n(h) - L(h)| \leq \varepsilon \quad \forall h \in \mathcal{H}$

$$P\left(\text{There exists } h \in \mathcal{H} \text{ s.t } |L_n(h) - L(h)| \geq \varepsilon\right)$$

$$= P\left(\bigcup_{h \in \mathcal{H}} \left\{|L_n(\hat{h}) - L(h)| \geq \varepsilon\right\}\right)$$

$$\leq \sum_{h=1}^{|\mathcal{H}|} P\left(|L_n(\hat{h}) - L(h)| \geq \varepsilon\right)$$

$$\leq 2|\mathcal{H}|\, 2e^{-2n\varepsilon^2} = 2e^{+\log|\mathcal{H}| - 2n\varepsilon^2} \quad (*)$$

$|\mathcal{H}|$ = # of functions in the class $\mathcal{H}$

If we want $(*)$ to be $\leq \delta$, then

$$e^{\log|\mathcal{H}| - 2n\varepsilon^2} \leq e^{\delta/2} \Rightarrow n > \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}.$$

If $\mathcal{H}$ is a function class with d parameters (e.g. a neural network with d coefficients) and each parameter is represented by a fixed precision, say k bits

$$|\mathcal{H}| \approx (2^k)^d.$$

$$\Rightarrow \quad n \gg d \cdot k$$

n: # of samples; $d \cdot k$: # of parameters.

Biggest open problem in ML: Deep network generalize well even when $n \ll d$.

Population Solution

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} L(h) = \operatorname*{argmin}_{h \in \mathcal{H}} \mathbb{E}_p[\ell(h(x), y)]$$

$$L(\hat{h}) = \underbrace{L(\hat{h}) - L(h^*)}_{\text{excess risk / estimation error}} + \underbrace{L(h^*)}_{\text{approximation error}}$$

$L(\hat{h})$: population loss of ERM solution

excess risk / estimation error: Typically larger if $\mathcal{H}$ is large

approximation error: Typically smaller if $\mathcal{H}$ is large

$$\underbrace{L(\hat{h}) - L(h^*)}_{\text{excess risk}} = L(\hat{h}) - L_n(\hat{h}) + \underbrace{L_n(\hat{h}) - L_n(h^*)}_{\leq 0} + L_n(h^*) - L(h^*)$$

$$\leq \underbrace{|L(\hat{h}) - L_n(\hat{h})|}_{\leq \varepsilon} + \underbrace{|L_n(h^*) - L(h^*)|}_{\leq \varepsilon}$$

$\leq 2\varepsilon$ if uniform convergence holds

$\leq 2\varepsilon$ with probability $1-\delta$

if $e^{\log|\mathcal{H}| - 2n\varepsilon^2} \leq e^{\log \delta/2}$

$$2n\varepsilon^2 \geq \log 2|\mathcal{H}|/\delta$$

$$n \geq \frac{\log 2|\mathcal{H}|/\delta}{2\varepsilon^2}$$