# Lecture 7 CMS 165

Generalization Theory

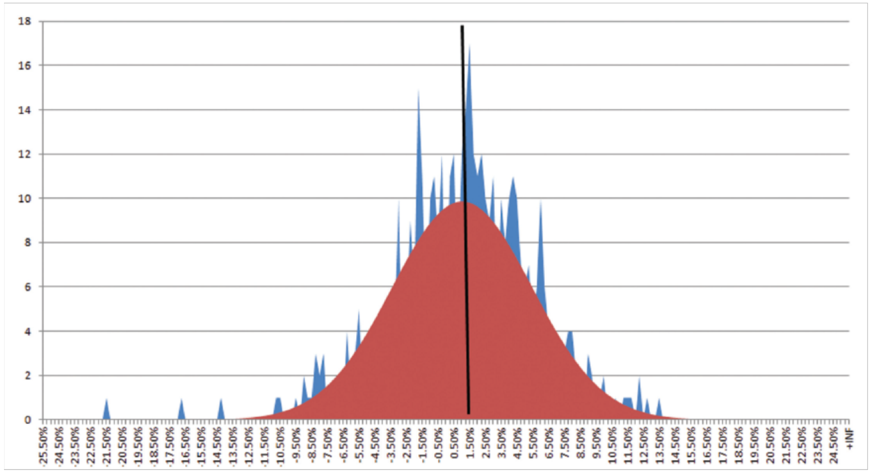
Recap:

Markov's inequality:

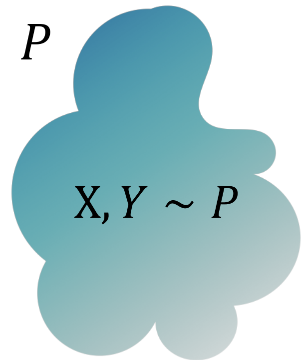
$$P(X \geq \epsilon) \leq \frac{E[X]}{\epsilon}$$

Hoeffding's inequality, $i.i.d.$ and $X \in [0,1]$ *(a simplified version):*

$$P\left(\frac{1}{n}\sum_i^n X_i - E\left[\frac{1}{n}\sum_i^n X_i\right] \geq \epsilon\right) \leq e^{-2n\epsilon^2}$$



Problem set-up:




Hypothesis class: $h \in H; h: X \to Y$

A loss function: $l(Y, h(X)) \in R, \quad e.g., \mathbb{I}(Y \neq h(X))$

Expected risk: $L(h) := E_P[l(Y, h(X))]$

Expected risk minimizer: $h^* \in \arg\min_{h \in H} L(h)$

Given a set of samples: $\{x_i, y_i\}_i^n$

Empirical risk: $\hat{L}(h) := \frac{1}{n} \sum_i^n l(y_i, h(x_i))$

Empirical risk minimizer: $\hat{h} \in \arg\min_{h \in H} \hat{L}(h)$

How good is $\hat{h}$ and how realistic is $\hat{L}(\hat{h})$?

$$L(\hat{h}) - \hat{L}(\hat{h})$$

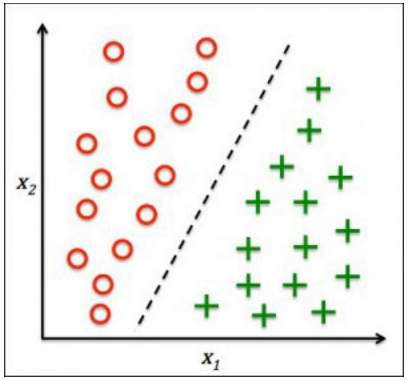
$$L(\hat{h}) - L(h^*)$$

Simple case, simple algorithm

1) Realizable setting: $L(h^*) = 0$

2) Finite $H$

3) Zero-one loss: $l(Y, h(X)) = \mathbb{I}(Y \neq h(X))$




Sample Complexity: $L(\hat{h}) \leq \epsilon$ with prob. $\delta$

Loss: How good we are after training on $n$ samples

$$\hat{L}(\hat{h}) = 0 \rightarrow \mathrm{P}[L(\hat{h}) \geq \epsilon] = ???$$

Consider a set $B$ such that $B \coloneqq \{h \in H : L(h) \geq \epsilon\}$

$$\mathrm{P}\left[L(\hat{h}) \geq \epsilon\right] = \mathrm{P}\left[L(\hat{h}) \in B\right]$$

$$\mathrm{P}\left[L(\hat{h}) \in B\right] \leq \mathrm{P}\left[\exists h \in B \colon \hat{L}(h) = 0\right]$$

Cool, what is the chance of an $h$ gives zero empirical loss?

$L(h)$ denotes the probability of mistake

$$P\left[\hat{L}(h) = 0\right] = \left(1 - L(\hat{h})\right)^n \leq (1 - \epsilon)^n \leq e^{-n\epsilon}$$

Now using the union bound;

$$\mathrm{P}\left[\exists h \in B \colon \hat{L}(h\,) = 0\right] \leq \sum_{h \in B} P\left[\hat{L}(h) = 0\right] \leq |B| e^{-n\epsilon} \leq |H| e^{-n\epsilon} \coloneqq \delta$$

By taking the log: $\mathrm{P}\left[L(\hat{h}) \geq \epsilon = \frac{\log(|H|/\delta)}{n}\right] \leq \delta$, it is also distribution free

Beyond the realizable case

$$L(\hat{h}) - L(h^*) = [L(\hat{h}) - \hat{L}(\hat{h})] + \underbrace{[\hat{L}(\hat{h}) - \hat{L}(h^*)]}_{\leq 0} + [\hat{L}(h^*) - L(h^*)]$$

For a given $h$, using the Hoeffding's inequality;

$$P\left(\frac{1}{n}\sum_{i}^{n} l(Y_i, h(X_i)) - L(h) \geq \epsilon\right) \leq e^{-2n\epsilon^2}$$

Also we know that

$$P\left(L(\hat{h}) - L(h^*) \geq \epsilon\right) \leq P\left([L(\hat{h}) - \hat{L}(\hat{h})] + [\hat{L}(h^*) - L(h^*)] \geq \epsilon\right)$$

Union bound: If for each $h$, $P\left(\left|L(h) - \hat{L}(h)\right| \geq \frac{\epsilon}{2}\right) \leq \frac{\delta}{2H}$

Then, $L(\hat{h}) - L(h^*) \leq \sqrt{\frac{2\left(\log\frac{2|H|}{\delta}\right)}{n}}$ with prob at least $1 - \delta$

What if we have some prior knowledge on hypothesis? e.g. $Pr(h)$

$$L(h) \geq \hat{L}(h) + \sqrt{\frac{log\frac{1}{Pr(h)\delta}}{2n}}$$ with prob at most $Pr(h)\delta$

PAC-Bayes

Beyond finite case:

$$P\left(L(\hat{h}) - L(h^*) \geq \epsilon\right) \leq P\left(\left[L(\hat{h}) - \hat{L}(\hat{h})\right] + \left[\hat{L}(h^*) - L(h^*)\right] \geq \epsilon\right)$$

$$\leq P\left(\sup_{h \in H} \left|L(h) - \hat{L}(h)\right| \geq \frac{\epsilon}{2}\right) := \delta$$

Rademacher Complexity

$$E\left[\sup_{h \in H} L(h) - \hat{L}(h)\right] \leq 2R_n(H, l)$$

$$R_n(H, l) = E\left[\sup_{h \in H} \frac{1}{n} \sum_{1}^{n} \sigma_i l(Y_i, h(X_i))\right], where\ \sigma_i\ is\ Rademcher\ random\ variable\ \{-1,1\}$$

$L(\hat{h}) - L(h^*) \leq 2R + \sqrt{\frac{2\left(\log\frac{2}{\delta}\right)}{n}}$ with prob at least $1 - \delta$

VC-Dimension: $R \leq \sqrt{\frac{2VC(H)(\log n+1)}{n}}$

Linear class $R \leq O\left(\sqrt{\frac{d(\log 2d)}{n}}\right)$

Bounded linear class $R \leq O\left(\sqrt{\frac{\beta(\log 2d)}{n}}\right)$