Empirical Risk Minimization for Hyperplanes

classifier : $c : \mathbb{R}^d \times \underbrace{\{\mathbb{R}^d \times \{0,1\}\}^n}_{\mathcal{Z}} \longrightarrow \{0,1\}$

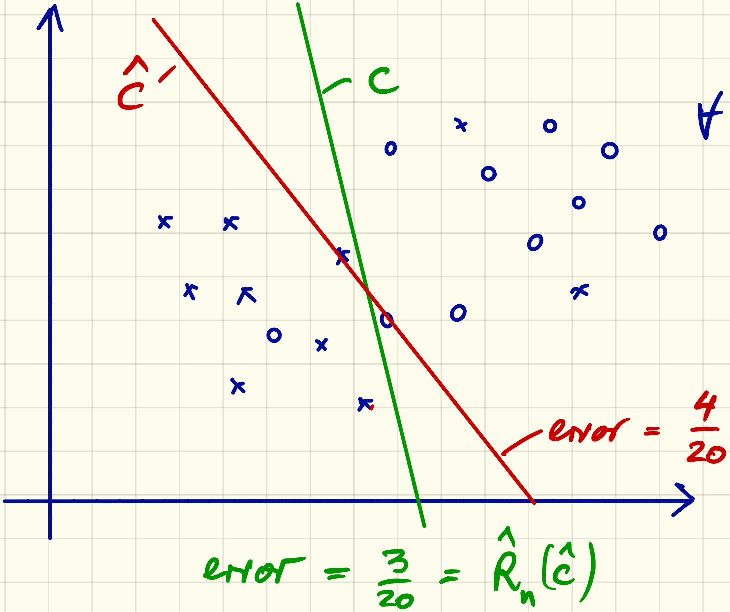
training data : $\mathcal{Z}_n = \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathcal{Z}$

set of independent variables $\mathcal{X} = \{x_1, \dots x_n\} \subset \mathbb{R}^d$

Hypothesis class : consider the set of hyperplanes $\mathcal{H}$

$$\mathcal{H} = \{ (a_0, a^T) \in \mathbb{R}^{d+1} : \text{s.t. } \exists\, \mathcal{X}_j = \{x_{i_1}, \dots x_{i_d}\} \subset \mathcal{X}$$
$$1 \leq j \leq \binom{n}{d}, \ \forall\, \tilde{x} \in \mathcal{X}_j \quad a^T \tilde{x} + a_0 = 0 \}$$

classifiers $\mathcal{C} = \left\{ c(x) = \begin{cases} 1 & a^T x + a_0 \gtreqless 0 \\ 0 & \text{otherwise} \end{cases}, \ (a_0, a^T) \in \mathcal{H} \right\}$



$$\forall c: \hat{R}_n(c) \geq \hat{R}_n(\hat{c}) - \frac{d}{n}$$

The ERM of all linear separators has bounded error by the ERM of classifiers through data points minus d errors

VC-inequality $\quad c^* = \arg\min_c R(c)$ (best linear classifier)

$$R(\hat{c}) - R(c^*) = R(\hat{c}) - \hat{R}_n(\hat{c}) + \underbrace{\hat{R}_n(\hat{c})}_{\hat{R}_n(c^*) + \frac{d}{n}} - R(c^*)$$

$$\leq \max_{1 \leq i \leq 2\binom{n}{d}} \left\{ R(c_i) - \hat{R}_n(c_i) \right\} + \hat{R}_n(c^*) - R(c^*) + \frac{d}{n}$$

Large deviation probability

$$P\left(R(\hat{c}) - R(c^*) > \varepsilon \right) \le$$

$$P\left(\max_{1 \le i \le 2\binom{n}{d}} \{ R(c_i) - \hat{R}_n(c_i) \} + \hat{R}_n(c^*) - R(c^*) + \frac{d}{n} > \varepsilon \right) \le$$

$$P\left(\max_{1 \le i \le 2\binom{n}{d}} \{ R(c_i) - \hat{R}_n(c_i) \} > \frac{\varepsilon}{2} \vee \hat{R}_n(c^*) - R(c^*) + \frac{d}{n} > \frac{\varepsilon}{2} \right) \le$$

$$P\left(\max_{1 \le i \le 2\binom{n}{d}} \{ R(c_i) - \hat{R}_n(c_i) \} > \frac{\varepsilon}{2} \right) + P\left(\hat{R}_n(c^*) - R(c^*) > \frac{\varepsilon}{2} - \frac{d}{n} \right)$$

r. v. $\xi = n \hat{R}_n(c^*)$ is binomially distributed with parameter n and $R(c^*)$

2019 / 12 / 11 (4)

$$P\left(n \hat{R}_n(c^*) = k\right) = \binom{n}{k} R(c^*)^k \left(1 - R(c^*)\right)^{n-k}$$

use Chernoff tail bound $\Rightarrow$

$$P\left(\hat{R}_n(c^*) - R(c^*) > \frac{\varepsilon}{2} - \frac{d}{n}\right) \leq \exp\left(-2n\left(\frac{\varepsilon}{2} - \frac{d}{n}\right)^2\right)$$

$$\leq e^{2d\varepsilon} e^{-n\frac{\varepsilon^2}{2}}$$

Remark: Note that pointwise convergence does not depend on data.

"Optimization term"

$$P\left(\max_{1 \leq i \leq 2\binom{n}{d}} \left\{ R(c_i) - \hat{R}_n(c_i) \right\} \geq \frac{\varepsilon}{2}\right) \leq$$

$$\sum_{1 \leq i \leq 2\binom{n}{d}} \mathbb{E}_{X_{i_1} \dots X_{i_d}} P\left(R(c_i) - \hat{R}_n(c_i) \geq \frac{\varepsilon}{2} \,\middle|\, X_{i_1} \dots X_{i_d}\right)$$

Proof idea:

d samples (data points) are used to define the classifier:

Replace them with new samples in the analysis, i.e.

$$(X_i', Y_i') = \begin{cases} (X_i'', Y_i'') & X_i \in \{X_{i_1}, \dots, X_{i_d}\} \\ (X_i, Y_i) & \text{otherwise} \end{cases}$$

$$P\left(R(c_i) - \hat{R}_n(c_i) \geq \frac{\varepsilon}{2} \,\middle|\, X_{i_1} \dots X_{i_d} \right) \leq$$

$$P\left(R(c_i) - \frac{1}{n} \sum_{j \notin \{i_1 \dots i_d\}} \mathbb{I}_{\{c_i(X_j) \neq Y_j\}} > \frac{\varepsilon}{2} \,\middle|\, X_{i_1} \dots X_{i_d} \right) \leq$$

$$P\left(R(c_i) - \underbrace{\frac{1}{n} \sum_{j=1}^{n} \mathbb{I}_{\{c_i(X_j') \neq Y_j'\}}}_{\sim \text{ binomial}(n, R(c_i))} + \frac{d}{n} > \frac{\varepsilon}{2} \,\middle|\, X_{i_1} \dots X_{i_d} \right)$$

$$\leq \exp\left(-2n\left(\frac{\varepsilon}{2}-\frac{d}{n}\right)^2\right) \leq \exp\left(-n\frac{\varepsilon^2}{2}+2d\varepsilon\right)$$

Since all the $2\binom{n}{d}$ terms are symmetric, it holds

$$\boxed{P\left(R(\hat{c})-R(c^*)>\varepsilon\right) \leq \left(2\binom{n}{d}+1\right)e^{2d\varepsilon}e^{-n\frac{\varepsilon^2}{2}}}$$

$$\leq \exp\left(\underbrace{\log\left(2\binom{n}{d}+1\right)}_{\substack{\text{entropic term} \\ \approx\, d\log n}}+2d\varepsilon-\underbrace{\frac{n\varepsilon^2}{2}}_{\substack{\text{fitting} \\ \text{(energy) term}}}\right)$$

with assumption $\varepsilon > 2\frac{d}{n}$

this "fingering" argument explores the richness of functions on samples.