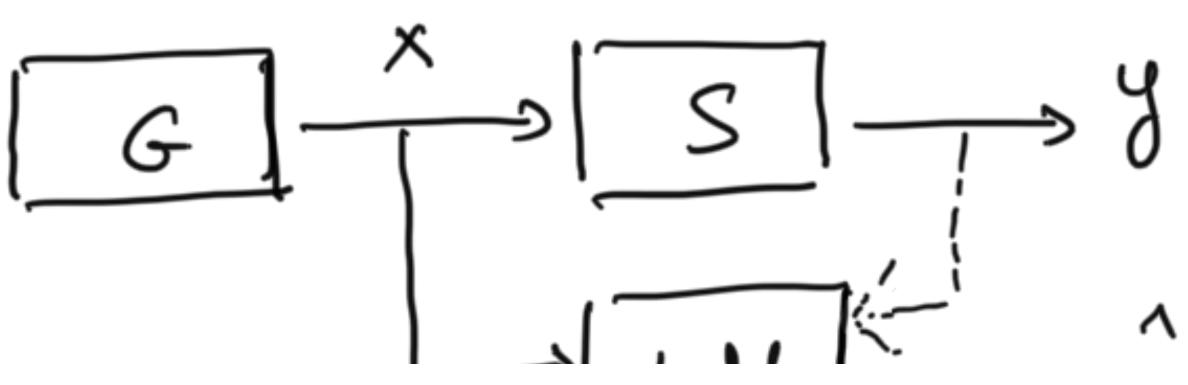
Brief overview of statistical learning theory

Literature:
- V. N. Vapnik, "The nature of statistical learning", 2nd edition, Springer, 2000
- M. Hardt, B. Recht, "Patterns, Predictions, and actions", arXiv, 2021
- U. v. Luxburg, B. Schölkopf, "Statistical Learning Theory: Models, Concepts, and Results", arXiv, 2008

1. Problem description (simplified setting)



$\longrightarrow \boxed{LM} \longrightarrow y$

- Generator (G): generates iid vectors $X_i \in \mathbb{R}^d$ from <u>unknown</u> but <u>fixed</u> cdf F_X
- Supervisor (S): returns an output $y_i \in \{0,1\}$ for every X_i according to <u>unknown</u> but <u>fixed</u> cdf $F_{y|X}$.
- Learning Machine (LM): implements a function $f(X,\alpha)$ that predicts y from X, where $\alpha \in \underline{\Lambda}$ are parameters.

- Goal: Given $(X_1, y_1), (X_2, y_2), \ldots$ (training examples) find $\alpha \in \underline{\Lambda}$ such that for a new X, the learning machine predicts the correct y.

→ Formally : Find $\alpha \in \Lambda$ that minimizes

$$R(\alpha) = E_{xy}[\ell(f(X,\alpha), y)]$$

risk

loss function

$$= \int \ell(f(x,\alpha), y)\, dF_{xy}$$

$$\ell(y,\hat{y}) \geq 0 \quad \forall y, \hat{y} \in \{0,1\}^2$$

→ Fundamental problem : we don't know F_{xy}.

2. Empirical risk minimization

Idea: minimize $\hat{R}_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(X_i,\alpha))$

instead.

- main questions of statistical learning theory: Under what circumstances is $\hat{R}_n(\hat{\alpha}_n)$ a good approx. of $R(\hat{\alpha}_n)$ and $\inf_{\alpha \in \Lambda} R(\alpha)$, where $\hat{\alpha}_n = \arg\min_{\alpha \in \Lambda} \hat{R}_n(\alpha)$?

  i). consistency: $R(\hat{\alpha}_n) \xrightarrow[n \to \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha)$

  $\hat{R}_n(\hat{\alpha}_n) \xrightarrow[n \to \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha)$?

  ii). at what rate? dependence on n, $|\Lambda|$?

  $|\Lambda|$ ~ cardinality of Λ

3. Main assumptions

i). (X_i, Y_i), $i = 1, \dots, n$ are independent samples

⊕ very convenient → often the only way forward

⊖ problematic in many practical applications

→ trajectories of a dynamical system

→ search path of animal looking for food → as an example for a random walk

ii). cdf is fixed

⊕ mathematically very convenient

⊖ could be problematic in practice

iii). no assumption on F_{xy}

↳ if knew F_{xy} → we could evaluate $R(a)$

↳ however, we have often some prior information available.

→ B. Recht et al. "Do Imagenet Classifiers Generalize to ImageNet?", arXiv, 2018

4. Results from statistical learning (Λ finite)

- $$\ell(y,\hat{y}) = \begin{cases} 0 & \text{if } y=\hat{y} \\ 1 & \text{else} \end{cases} \Rightarrow$$ R reduces to the probability of making a mistake.

- $\Lambda = \{\alpha_1, \dots, \alpha_m\}$ finite

- $$\begin{aligned} R(\alpha) &= \hat{R}_n(\alpha) + R(\alpha) - \hat{R}_n(\alpha) \\ &\le \underbrace{\hat{R}_n(\alpha)}_{\text{this we can compute}} + \underbrace{\sup_{\alpha' \in \Lambda} R(\alpha') - \hat{R}_n(\alpha')}_{\text{this is our goal}} \end{aligned}$$ for any $\alpha \in \Lambda$

- $$\Pr\Big(\sup_{\alpha' \in \Lambda} R(\alpha') - \hat{R}_n(\alpha') \geq \varepsilon \Big)$$

"OR"

$$= \Pr\Big(R(\alpha_1) - \hat{R}_n(\alpha_1) \geq \varepsilon \vee R(\alpha_2) - \hat{R}_n(\alpha_2) \geq \varepsilon \vee \dots \vee R(\alpha_m) - \hat{R}_n(\alpha_m) \geq \varepsilon \Big)$$

union bound

$$\leq \sum_{i=1}^{m} \Pr\big(R(\alpha_i) - \hat{R}_n(\alpha_i) \geq \varepsilon \big)$$

- if we fix α_i, $\ell(f(X, \alpha_i), y)$ is a Bernoulli-RV with mean $R(\alpha_i)$.

- $\hat{R}_n(\alpha_i)$ corresponds to the empirical mean.

$$\hat{R}_n(\alpha_i) \xrightarrow[n \to \infty]{a.s} R(\alpha_i).$$

- Hoeffdings inequality:

$$\Pr\big(R(\alpha_i) - \hat{R}_n(\alpha_i) \geq \varepsilon \big) \leq e^{-2n\varepsilon^2}$$

$$\Rightarrow \Pr\Big(\sup R(\alpha') - \hat{R}_n(\alpha') \geq \varepsilon \Big) \leq \underbrace{m e^{-2n\varepsilon^2}}$$

$\alpha' \in \Lambda$ δ

- Set $\varepsilon = \sqrt{\dfrac{\log(1/\delta) + \log(|\Lambda|)}{2n}}$

with prob. $1-\delta$

$$\boxed{R(\alpha) \le \hat{R}_n(\alpha) + \sqrt{\frac{\log(1/\delta) + \log(|\Lambda|)}{2n}} \quad \forall \alpha \in \Lambda}$$

$$\sqrt{\frac{\log(\text{"number of hypothesis"})}{\text{"number of samples}}}$$

- convergence is slow in the number of training examples.

- In case $\hat{R}_n(\alpha) = 0$ we can show that

$$R(\alpha) \le \frac{\log(1/\delta) + \log(|\Lambda|)}{}$$

n

5. Results from statistical learning when (Λ infinite)

- use an argument called symmetrization

For any $\varepsilon \geq \sqrt{2/n}$ we have

$$\Pr\left(\sup_{\alpha \in \Lambda} |R(\alpha) - R_n^1(\alpha)| \geq \varepsilon \right)$$

$$\leq 2 \Pr\left(\sup_{\alpha \in \Lambda} |R_n^2(\alpha) - R_n^1(\alpha)| \geq \varepsilon/2 \right)$$

where $R_n^2(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i', \alpha), y_i')$,

$(x_i', y_i') \overset{iid}{\sim} F_{xy}$.

- (x_i', y_i') is ghost sample.

- $\sup |R_n^2(\alpha) - R_n^1(\alpha)|$

$\alpha \in \Lambda$

↳ remains the same if we classify sample and ghost sample in the same way

↳ reduces to a finite set of possibilities

→ in general there are 2^{2n} different ways to classify sample and ghost sample.

→ $N(\Lambda, n) \hat{=}$ "the number of functions from Λ, which can be distinguished based on their values on n samples"

→ $N(\Lambda, n)$ either grows ($\sim 2^n$) or polynomially ($\sim n^d$), d is constant.

→ $R(\alpha) \leq \hat{R}^n(\alpha) + \sqrt{\log(1/\delta) + d\log(n)}$

$\Rightarrow \quad \mathcal{R}(\alpha) \leq \mathcal{R}_n(\alpha) + \quad \forall \quad n$

with prob. $1-\delta$.

---

6. Counting parameters?

- $f(x, \alpha) = \begin{cases} 1 & \text{if } \cos(x^T\omega + b)^T \alpha \geq 0.5 \\ 0 & \text{else} \end{cases}$

$x \in \mathbb{R}^2$, $\omega \in \mathbb{R}^{2\times D}$ (randomly generated), D can be large, $b \in \mathbb{R}^D$ randomly generated.

$$\alpha = \underset{\alpha' \in \mathbb{R}^D}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{2} |y_i - \cos(X_i^T\omega + b)^T \alpha'|^2 + \frac{\lambda}{2} n |\alpha'|^2$$

fixed λ to 0.01