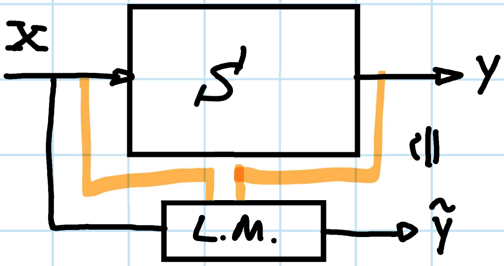
ML AND DATA ANALYSIS	OCCAM'S RAZOR	Set. 22. 2023



$D = \{(\underline{x}_1, y_1), \dots (\underline{x}_N, y_N)\}$ ↳ i.i.d.

$\underline{x} \in \mathbb{R}^d$

$y \in$
- $\{0,1\}$ B.C.
- $\{1,2,\dots,C\}$ M.C.C.
- $\mathbb{R}$ REGR.

L.M. : $f(\underline{x}) = \tilde{y}$ MAPS $\underline{x}$ INTO $\tilde{y}$, CHOOSEN IN $f \in \underline{F}$ ← set of functions

LOSS FUNCTION: $\ell(f(\underline{x}), y)$ HOW GOOD IS MY PREDICTION IN RESPECT TO MY TARGET

EMPIRICAL ERROR: $\hat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(\underline{x}_i), y_i)$ ERROR THAT I COMPUTE ON MY DATA

(TRUE) GENERALIZATION ERROR: $R(f) = \mathbb{E}_{\underline{x},y}\, \ell(f(\underline{x}), y)$ ERROR THAT I COMPUTE ON THE POPULATION, THE ONE TO MINIMIZE

$\boxed{\text{CASE } |F| = 1}$

$F = f \Rightarrow$ I HAVE ONLY 1 FUNCTION $\Rightarrow$ I HAVE NO LEARNING PROCESS

$$\hat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} \ell_i$$

$$R(f) = \mathbb{E}_p\{\ell\}$$

VAR. OF RANDOM VAR ℓ

$$P\{|R - \hat{R}| \geq \varepsilon\} \leq \frac{6_\ell^2}{m\varepsilon^2} = \delta \rightarrow \varepsilon = \sqrt{\frac{6_\ell^2}{m\delta}}$$

$$P\left\{|R - \hat{R}| \geq \sqrt{\frac{6_\ell^2}{m\delta}}\right\} \leq \delta$$

? : UNKNOWN

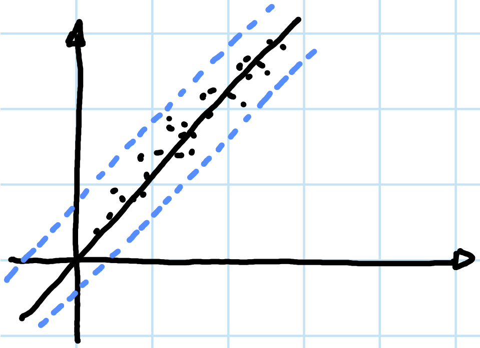
$$|R - \hat{R}|_{1-\delta} \leq \sqrt{\frac{6_\ell^2}{m\delta}}$$

? ?

THE SMALLER IS $\delta \Rightarrow$ THE MORE PROBABLE THIS STATEMENT IS TRUE, THE LARGER IS THE INTERVAL BETWEEN R AND $\hat{R}$

$$\lim_{\delta \to 0} |R-\hat{R}|_{1-\delta} \le \sqrt{\frac{\sigma_\ell^2}{m\delta}} = \infty$$

IF I HAVE A lot OF samples ($m \to \infty$) THE DISTANCE BECOMES smaller ($|R-\hat{R}| \to 0$)

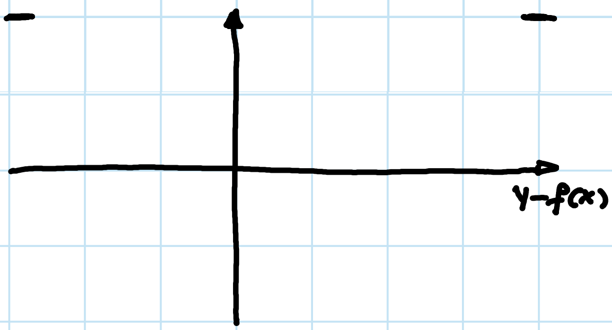


is $\ell \in [0,\infty)$? (UNBOUNDED)

THE FAR I'M FROM THE PREDICTION, THE LARGER THE ERROR,

BUT THEN AFTER A CERTAIN value IS ONly AN ERROR.

↳ usually THE loss IS BOUNDED, $\ell \in [0,1]$ IF I NORMALIZE. IN M.C.C. I JUST HAVE CORRECT OR NOT CLASSIFICATION.



IF $\ell \in [0,1] \Rightarrow \sigma_\ell^2 \le 1$

$\Downarrow$

δ CONFIDENCE

$$\underbrace{P\{|R-\hat{R}| \ge \varepsilon\}} \le \frac{\sigma_\ell^2}{m\varepsilon^2} \le \frac{1}{m\varepsilon^2}$$

PROB. OF TRUE ERROR TO BE FAR AWAY FROM THE EMPIRICAL ERROR MORE THAN ε $\le \frac{1}{m\varepsilon^2}$

$$|R-\hat{R}|_{1-\delta} \le \sqrt{\frac{\sigma_\ell^2}{m\delta}} \le \sqrt{\frac{1}{m\delta}}$$

EVERY ML ALGORITHM CAN BE RECONDUCTED TO THIS FORMULA.
THERE'S RELATION BETWEEN THE MAX. DISTANCE FROM R AND $\hat{R}$ AND THE # OF samples collected.

→ I EITHER HAVE MORE DATA ($m \to \infty$) OR LESS CONFIDENCE ($\delta \to 0$)

$P\{|R-\hat{R}| \ge \varepsilon\} \le \frac{1}{m\varepsilon^2}$ CHERNOFF BOUND → HYPOTHESIS: i.i.d. $\ell \in [0,1]$

OTHER BOUNDS ARE Also USED: $P\{|R-\hat{R}| \ge \varepsilon\} \le e^{-2m\varepsilon^2}$ HOEFFDIG BOUND

CASE $|F| = m_f$

INDEPENDENT FROM THE DATA D_m!!

ASSUMPTIONS: D_m is i.i.D, $\ell \in [0,1]$, $|F| = m_f$

$\Rightarrow$ f IS LEARNED, FROM D_m, $F \rightarrow f \in F$

THE CHOSEN FUNCTION CAN DEPEND FROM THE DATA D_m

THE FUNCTION SPACE F IS CHOSEN BEFORE OBSERVING THE DATASET $\Rightarrow$ F IS INDEPENDENT FROM THE DATA $f \in F$ IS CHOSEN BY LOOKING AT THE DATA.

$$\forall f \in F \quad P\{|R(f) - \hat{R}(f)| \geq \varepsilon\} \leq \frac{1}{m\varepsilon^2}$$

BECAUSE EACH FUNCTION IS INDEPENDENTLY FROM THE DATASET.

WORST CASE SCENARIO: $\forall f \in F$ THE TRUE ERROR IS FAR AWAY FROM THE OBSERVED.

$$\hat{f}: \quad P\{|R(\hat{f}) - \hat{R}(\hat{f})| \geq \varepsilon\} \leq \sum_{i=1}^{m_f} \frac{1}{m\varepsilon^2} = \frac{m_f}{m\varepsilon^2}$$

$\rightarrow$ UNDERSTAND THE FORMULA

$$\frac{m_f}{m\varepsilon^2} = \delta \Rightarrow \varepsilon = \sqrt{\frac{m_f}{m\delta}}$$

BY PITAGORA

$$|R(\hat{f}) - \hat{R}(\hat{f})|_{1-\delta} \leq \sqrt{\frac{m_f}{m\delta}} \leq \sqrt{\frac{m_f + 1 - 1}{m\delta}} \leq \sqrt{\frac{m_f - 1}{m\delta}} + \sqrt{\frac{1}{m\delta}}$$

STATISTICAL PRICE: THE MORE DATA I HAVE THE LESS RISK I HAVE.

DEPENDS FROM THE LEARNING (m_f)

- IF $m_f = 1$ $\Rightarrow$ CLASSICAL STATISTICS $\rightarrow$ $|R(\hat{f}) - \hat{R}(\hat{f})|_{1-\delta} \leq \sqrt{\frac{1}{m\delta}}$ (NO LEARNING)
- IF $m_f > 1$ $\Rightarrow$ LEARNING PROCESS

$\boxed{D_m}$ $\boxed{F}$

$$\hat{f} = \arg\min_{f \in F} \hat{R}(f)$$

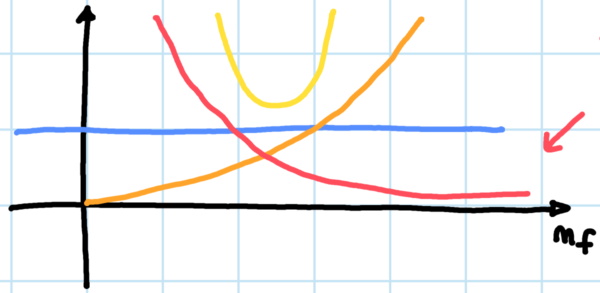
THE MINIMUM ERROR ON MY DATA

EMPIRICAL RISK MINIMIZATION (E.R.M.)

$\hat{R}(f)$ AND m_f ARE RELATED: THE MORE FUNCTION I HAVE, BIGGER IS THE RISK, BUT IT'S PROBABLE THAT MY $\hat{R}(f)$ WILL BE SMALLER.

$$|R(\hat{f}) - \hat{R}(\hat{f})| \underset{1-\delta}{\leq} \sqrt{\frac{m_f - 1}{m\delta}} + \sqrt{\frac{1}{m\delta}}$$

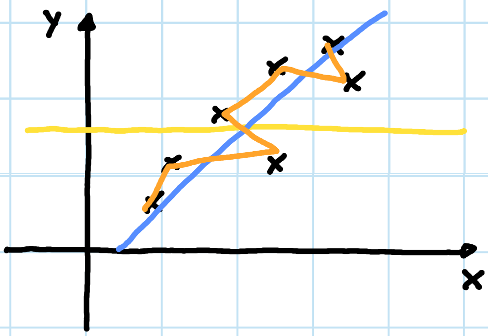
$$R(\hat{f}) \leq \hat{R}(\hat{f}) + \sqrt{\frac{m_f - 1}{m\delta}} + \sqrt{\frac{1}{m\delta}}$$



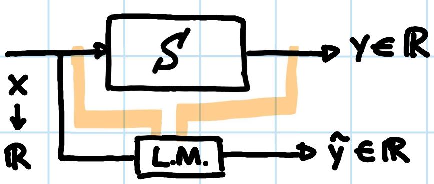
THE MORE FUNCTION I LEARN THE MORE PROB. A FUNC. WILL WORK WELL ON MY DATA

OCCAM'S RAZOR PRINCIPLE: GIVEN A PROBLEM IF YOU HAVE A LIST OF SOLUTIONS THAT MORE OR LESS BEHAVES THE SAME, THE ONE TO SELECT IS THE SIMPLEST ONE

IN M.L.: STRUCTURAL RISK MINIMIZATION



THE BLU IS THE CORRECT ONE



$$D_m = \{(x_1, y_1), \ldots, (x_N, y_N)\} \text{ i.i.d}$$

$$f(x) = \sum_{i=0}^{P} a_i x^i$$

HYPERPARAMETER

PARAMETERS

ALSO THE FUNC. FORM IS AN HYPER PARAMETER (MCLAURIN, FOURIER...)

$$\ell(f(x), y) = (y - f(x))^2 \quad \text{M.S.E.}$$

WHAT WE WANT TO FIND: $f^* = \underset{\substack{f \in F \\ a \in \mathbb{R}^P}}{\text{ARG min}}\, R(f)$ WITH $R(f) \leq \hat{R}(f) + C(m_f) + \sqrt{\frac{1}{m\delta}}$

ORACLE

COMPLEXITY OF MY MODEL

WITH $R(f) \underset{\delta}{\leq} \hat{R}(f) + \lambda C(m_f) + \sqrt{\frac{1}{m\delta}}$

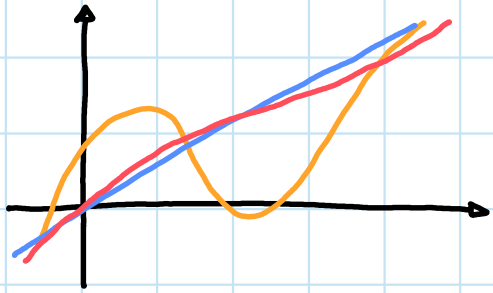
λ: REGULARIZATION PARAMETER (HYPERPARAMETER)

TRADES OFF ACCURACY AND COMPLEXITY: HOW MUCH I WANT TO LEARN FROM MY DATA AND HOW MUCH I WANT MY MODEL COMPLEX

HOW TO MEASURE COMPLEXITY

$C(f)$

$$f(x) = \sum_{i=0}^{p} a_i x^i$$



• MORE COMPLEX THAN • (DEGREE OF THE POLY)

FOR • I USE THE FLACTUATION (f'')

$\dfrac{d^2 f}{dx^2}$: NOT SO GOOD SINCE IS A FUNCTION $\Rightarrow$ I'll CALCULATE THE INTEGRAL IN THE DOMAIN (0 AND 1 SINCE I CAN ALWAYS NORMALIZE)

$\displaystyle\int_0^1 \left(\frac{d^2 f}{dx^2}\right)^2 dx$: FIRST MEASURE OF COMPLEXITY

$\displaystyle\int_0^1 \left(\frac{d^2 f}{dx^2}\right)^2 dx$ WITH $\displaystyle f(x) = \sum_{i=0}^{p} a_i x^i$

$$\frac{df}{dx} = \sum_{i=1}^{p} i a_i x^{i-1}$$

↑ DERIVATIVE OF 1 = 0

$$\frac{df^2}{dx^2} = \sum_{i=2}^{p} i(i-1) a_i \cdot x^{i-2}$$

$$\left(\frac{df^2}{dx^2}\right)^2 = \sum_{i=2}^{p} \sum_{j=2}^{p} ij(i-1)(j-1) a_i a_j x^{i-2} x^{j-2}$$

$$\int_0^1 \left(\frac{df^2}{dx^2}\right)^2 = \sum_{i=2}^{p} \sum_{j=2}^{p} ij(i-1)(j-1) a_i a_j \int_0^1 x^{i+j-4}\, dx = \sum_{i=2}^{p} \sum_{j=2}^{p} ij(i-1)(j-1) a_i a_j \left.\frac{x^{i+j-3}}{i+j-3}\right|_0^1 =$$

$$= \sum_{i=2}^{p} \sum_{j=2}^{p} \frac{(i-1)(j-1)ij}{i+j-3} a_i a_j$$

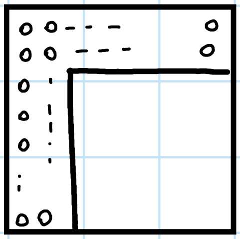
WHAT WE WANT TO MINIMIZE?

$$\hat{f} = \min_{\substack{f \in F \\ \underline{a} \in \mathbb{R}^d}} \hat{R}(f) + \lambda C(f) \Rightarrow \min_{\underline{a}} \frac{1}{m} \sum_{i=1}^{m} \Big(\underbrace{\sum_{j=0}^{p} a_j x_i^j}_{f(x_i)} - y_i \Big)^2 + \lambda \overbrace{\sum_{i=2}^{p} \sum_{j=2}^{p} \frac{ij(i-1)(j-1)}{i+j-3} a_i a_j}^{\underline{a}' M \underline{a}}$$

REWRITE THIS FORMULA INTO MATRICIAL FORM

$$\underline{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_p \end{bmatrix} \in \mathbb{R}^{p+1} \qquad \underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^{m} \qquad X = \begin{bmatrix} x_1^0 & \cdots & x_1^p \\ & \vdots & \\ x_N^0 & \cdots & x_N^p \end{bmatrix} \in \mathbb{R}^{m \times (p+1)}$$

$$M = \begin{cases} M_{ij} = 0 & i,j < 2 \\ M_{ij} = \dfrac{ij\,(i-1)(j-1)}{i+j-3} & i,j \geq 2 \end{cases}$$



$\in \mathbb{R}^{(p+1)\times(p+1)}$

$$\min_{\underline{a}} \frac{1}{m} \|X\underline{a} - \underline{y}\|_2^2 + \lambda \underline{a}' M \underline{a}$$

$$\min_{\underline{a}} \underbrace{\|X\underline{a} - \underline{y}\|_2^2}_{\hat{R}(f)} + \overbrace{\lambda \underline{a}' M \underline{a}}^{C(f)}$$

PARABOLOID

I : RIDGE REGRESSION (REGULARIZED RIDGE SQUARE)

TO FIND THE MINIMUM I COMPUTE THE GRADIENT:

$$\nabla_{\underline{a}} \left(\underline{a}' X' X \underline{a} - 2\underline{a}' X' \underline{y} + \underline{y}' \underline{y} + \lambda \underline{a}' M \underline{a} \right) = \underline{0}$$

$$\cancel{2} X' X \underline{a} - \cancel{2} X' \underline{y} + \cancel{2} M \underline{a} = 0$$

$$(X'X + \lambda M)\underline{a} = X'\underline{y}$$

LINEAR SYSTEM, GAUSS JORDAN $O(m^2)$