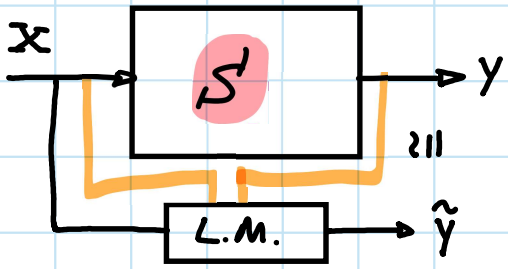
ML AND DATA ANALYSIS	HOW TO USE THE DATA	Set. 29. 2023



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

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

$y \in \begin{cases} \{0,1\} \text{ B.C.} \\ \{1,2,..,C\} \text{ M.C.C.} \\ \mathbb{R} \text{ REGR.} \end{cases}$

↳ i.i.d. THE MORE DATA I HAVE THE MORE I KNOW ABOUT THE SYSTEM (INDEP.), THE DATA DOESN'T CHANGE IN TIME (IDENTICALLY DISTR.)

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

LOSS FUNC. IS BOUNDED (NEW HYPOTHESIS)

$\ell(f(x), y) = (y - f(x))^2$ SQUARED ERROR (LOSS FUNCTION) $\in [a,b]$

$\hat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(x), y)$ EMPIRICAL ERROR

$R(f) = E_{x,y}\, \ell(f(x), y)$ TRUE ERROR

IF THE FUNC. IS INDEPENDENT FROM THE DATA: WE CAN PROVE:

STATISTICS → $|R(f) - \hat{R}(f)| < \sqrt{\frac{1}{m\delta}}$ with $1-\delta$

Dm IND f

WE WANT TO FIND THE BEST FUNCTION:

WE DON'T KNOW THE DISTRIBUTION THAT GENERATED MY DATA

$f^* = \underset{f \in F}{\text{ARG min}}\, R(f)$

I KNOW THAT IS BOUNDED!

IF THE SET OF FUNC. IS INDEPENDENT FROM THE DATA:

STATISTICAL LEARNING → $|R(f) - \hat{R}(f)| < \sqrt{\frac{1}{m\delta}} + \sqrt{\frac{M_f - 1}{m\delta}}$ with $1-\delta$

$|F| = m_f$ IND f

$f^* = \underset{f \in F}{\text{ARG min}}\, \hat{R}(f) + \lambda C(f)$

$\sqrt{\frac{m_f - 1}{m\delta}}$: STILL WE DON'T KNOW m_f ⇒ I CAN USE $\int_0^1 \left(\frac{d^2 f}{dx^2}\right)^2 dx$

IN • THERE ARE THE CHOICES, IN • THE PARAMETERS I CAN CHANGE.

WHAT IS THE NOISE?

- MEASURE
- KNOWLEDGE

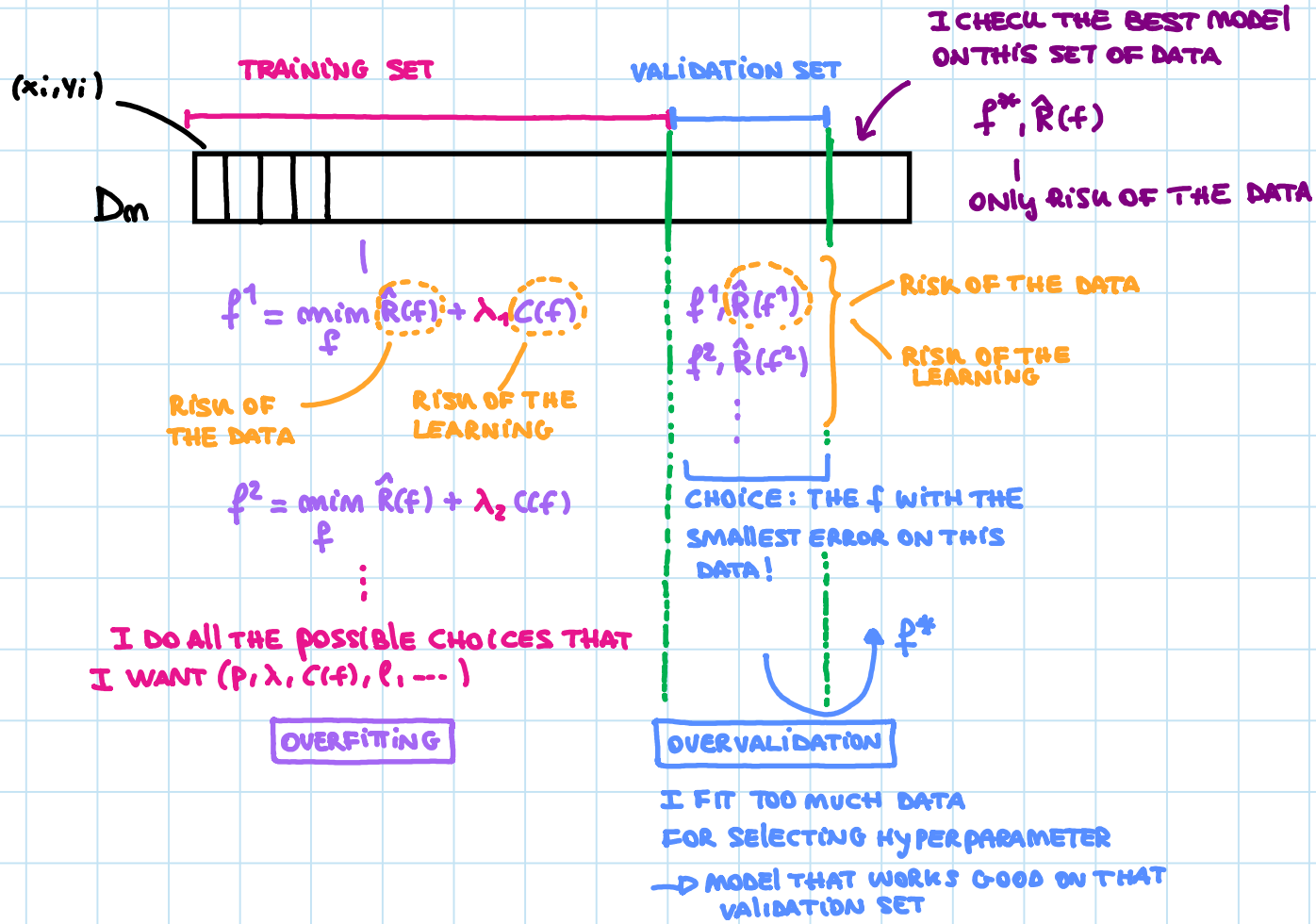
NO FREE LUNCH THEOREM: IT ALWAYS EXIST AN ALGORITHM THAT CAN PERFORM BETTER ON A SPECIFIC TASK WITH RESPECT TO ALL THE ALGORITHMS THAT EXIST IN THE REAL WORLD.

⇒ THERE WILL BE ALWAYS A MODEL BETTER THAN THE ONE YOU DEVELOPED.

SHORTCUT: I DON'T TRY TO UNDERSTAND WHAT REGULATES REAL WORLD PHENOMENA, BUT I SEARCH ONLY FOR RELATIONS BETWEEN EVENTS ⇒ SOONER OR LATER YOU'LL PAY THE PRICE

HOW TO CHOOSE $\lambda, p, (\ell, f, \dots)$ IN A GROUNDED WAY? — WE DON'T KNOW ANY OPTIMAL WAY.

① CAN I REMOVE THE LEARNING PART FROM THE EQUATION? ⇒ THE RISK DUE TO THE DATA.



THE TRAINING SET IS BIG BECAUSE WE HAVE A LOT OF RISK TO MINIMIZE.

I CAN SPLIT MY DATA INTO 3 PARTS BECAUSE IT'S I.I.D.

BEFORE DEPLOYING YOUR MODEL THE ONLY WAY TO BE SAFE IS TOO KEEP SOME FRESH DATA IN THE POCKET!

WHEN I TEST MY MODEL I HAVE TO DO IT WITH ON A SET OF DATA THAT I DIDN'T USED BEFORE.

IN THE VALIDATION SET I TAKE A FUNCTION THAT I LEARNED FROM THE DATA ON THE LEFT AND I TEST IT ON A SET OF DATA THAT IT HASN'T SEEN BEFORE.

SO f^1 IS INDEPENDENT FROM THE DATA.