OPTIMAL LECTURE 1, 18/1/2021

ML converts experience in expertise / knowledge.

DATASET

supervised learning $S = \{(x_i, y_i) \; i = 1, \dots, n\}$

$x \to ?$ → classification
→ regression

unsupervised learning $S = \{z_i, \; i = 1, \dots n\}$

clustering

reinforcement learning

$h : x \to y$

$\ell(h, (x,y)) = (h(x) - y)^2$
regr.

$\ell(h, (x,y)) = \mathbb{1}(h(x) \neq y)$
class.

$= \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{if } h(x) = y \end{cases}$

HOW TO EVALUATE A PREDICTOR

tomorrow $(x,y) \sim \mathcal{D}$

minimize $\mathbb{E}_{(x,y)\sim\mathcal{D}} [\ell(h, (x,y))]$
$h \in H$

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right] \;\begin{cases} \to L_{\mathcal{D}}(h) & \text{expected loss} \\ \to R(h) & \text{expected risk} \end{cases}$$

FIND A GOOD PREDICTOR → SOLVE AN OPTIMIZ. PROBLEM

$$\text{minimize}_{h\in\mathcal{H}} \; R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right]$$

$\mathcal{D}$ is unknown

ML is "distribution-free"

ML assumes S is drawn from $\mathcal{D}$ (independently)

$$\frac{1}{n}\sum_{i=1}^{n}\ell(h,(x_i,y_i)) \;\approx\; \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right]$$

EMPIRICAL LOSS/RISK

$L_S(h)$ / $R_S(h)$

$(R_n(h))$

CLOSER AS $n\to\infty$

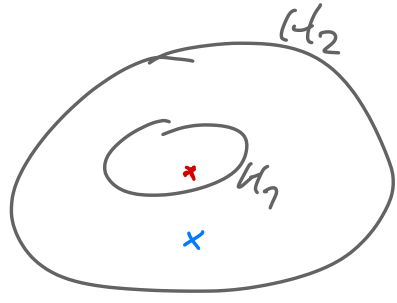
THE PROBLEM ML CONSIDERS IS

$$\underset{h \in H}{\text{minimize}} \; R_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i))$$

Choose between H_1, H_2 $\quad H_1 \subset H_2$

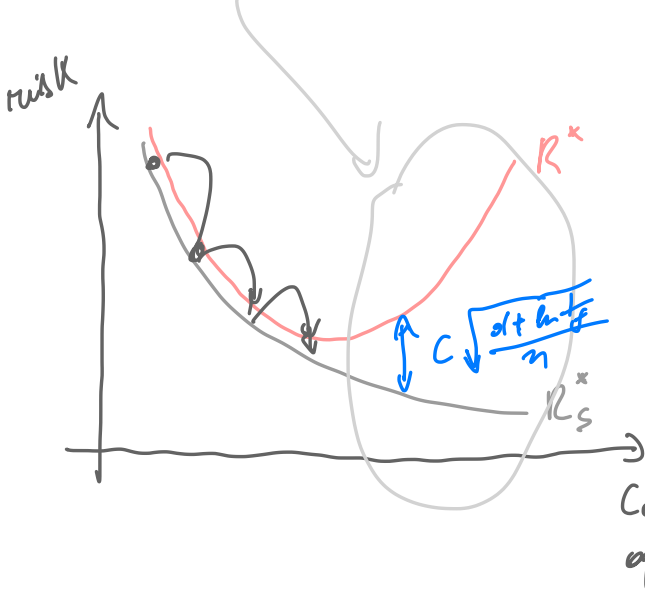
$h_1^* \in \underset{h \in H_1}{\operatorname{argmin}} R_S(h)$

$h_2^* \in \underset{h \in H_2}{\operatorname{argmin}} R_S(h)$



$$R_S(h_2^*) \leq R_S(h_1^*)$$

OVERFITTING PROBLEM : IT MAY BE $R(h_2^*) > R_S(h_1^*)$



statistics vs VC

asymptotic results

finite sample size

$R_S(h) \simeq R(h)$

$|S| = n \to \infty$

$\forall h \in H$

$$|R_S(h) - R(h)| \leq C \sqrt{\frac{d + \ln \frac{1}{\delta}}{n}}$$

CONSTANT

with prob. $1 - \delta$

$d = VC\text{-}dim(H)$

VC-dim

Vapnik Chervonenkis dimension

quantifies the complexity of the class of predictors

for uni-dimensional polynomials

d = # of coefficients of the polynomial

$$R(h) \simeq R_S(h) + C \sqrt{\frac{d + \ln \frac{1}{\delta}}{n}}$$

if small it is ok to minimize $R_S(h)$ to minimize $R(h)$

REGULARIZATION TERM

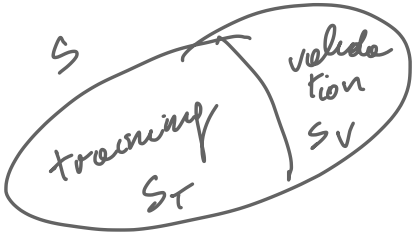
minimize $R_S(h) + \lambda \|h\|^2$

WE COULD

minimize $R_S(h) + C\sqrt{\frac{d + \ln\frac{1}{\delta}}{n}}$

BUT IT IS A LOOSE BOUND

IN PRACTICE:



$h_1^* \in \underset{h \in H_1}{\operatorname{argmin}} R_{S_T}(h)$

$h_2^* \in \underset{h \in H_2}{\operatorname{argmin}} R_{S_T}(h)$

$\left(R_{S_T}(h_1^*) \geq R_{S_T}(h_2^*)\right)$

$R_{S_V}(h_1^*) < R_{S_V}(h_2^*) \Rightarrow$ SELECT h_1^*

$R_{S_V}(h_1^*) > R_{S_V}(h_2^*) \Rightarrow$ SELECT h_2^*

$R(h_1^*) \overset{?}{\gtrless} R(h_2^*)$

$R(h_1^*) \simeq R_{S_V}(h_1^*)$

$R(h_2^*) \simeq R_{S_V}(h_2^*)$

$$\left| R(h) - R_{S_T}(h) \right| \leq C \sqrt{\frac{VC\text{-}dim(H) + \ln \frac{1}{\delta}}{|S_T|}}$$

$\forall h \in H$

VC-dim (H) BIG!

I only care of h_1^* & h_2^*

$$\left| R(h) - R_{S_V}(h) \right| \leq C \sqrt{\frac{VC\text{-}dim(\{h_1^*, h_2^*\}) + \ln \frac{1}{\delta}}{|S_V|}}$$

$h \in \{h_1^*, h_2^*\}$

$$VC\text{-}dim(\{h_1^*, h_2^*\}) \ll\ll VC\text{-}dim(H)$$

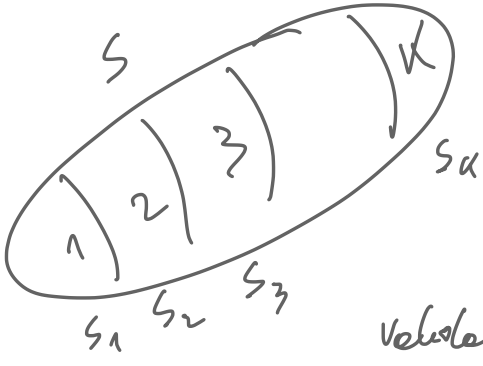
$\hookrightarrow \leq 1$

IT CAN BE

$R_{S_T}(h_1^*)$ is not close to $R(h_1^*)$

but $R_{S_V}(h_1^*)$ is close to $R(h_1^*)$

K-fold (cross) validation



$h_1 \in \text{argmin } R_{S_2 \cup \dots S_K}(h)$

$h_2 \in \text{argmin } R_{S_1 \cup S_3 \cup \dots S_K}(h)$

$h_K \in \text{argmin } R_{S_1 \cup S_2 \dots S_{K-1}}(h)$

validation $= \frac{1}{K} \sum_{i=1}^{K} R_{S_i}(h_i)$

K cross fold validation → solving K optimization problem

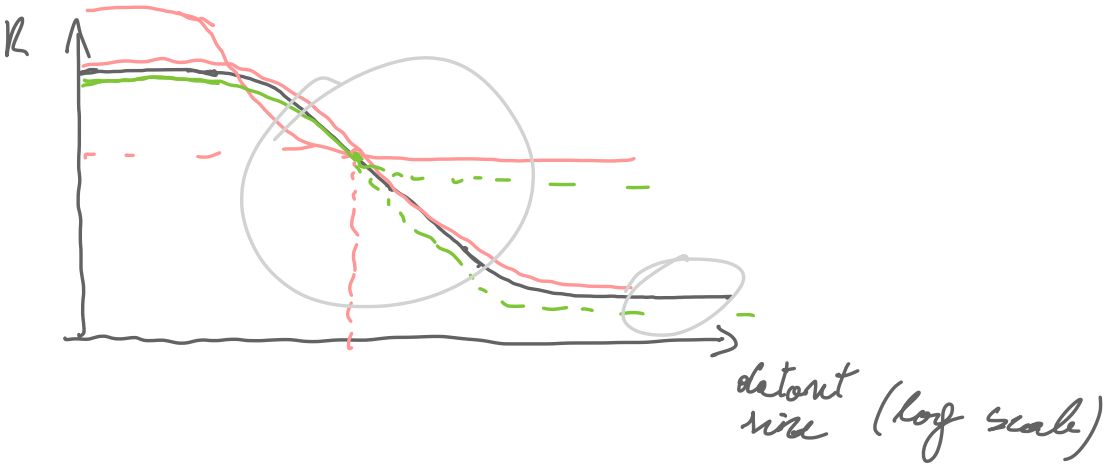
HYPER PARAMETERS
= PARAMETERS OF THE ALGORITHM USED TO LEARN THE MODEL
≠ PARAMETERS OF THE MODEL

- learning rate
- parameters of the grid search
- batch size

EXPERIENCE: 3 ÷ 10 VALUES PER HYPER PARAMETER

LEARNING PROBLEM

$$\#\text{models} \times \underbrace{K}_{(K\text{-fold})} \times (3 \div 10)^{\#\text{hyperparameters}}$$


R
dataset size (log scale)