DATASET

CLASSIFICATION

$x \xrightarrow{h} y$ LABEL OF A CLASS

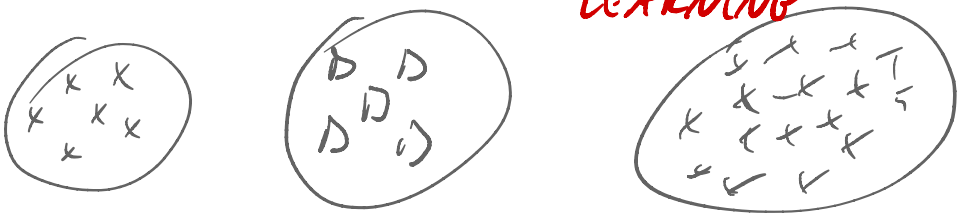
INPUT (DOG/CAT)

REGRESSION

$x \xrightarrow{h} y \in \mathbb{R}$

INPUT

CLUSTERING → UNSUPERVISED LEARNING



DIMENSION REDUCTION

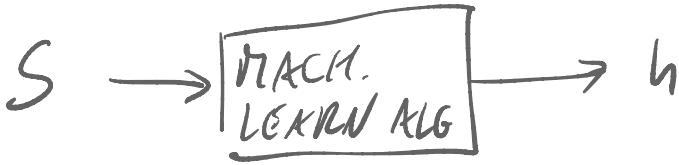
REINFORCEMENT LEARNING

CLASSIF / REGR PROBLEMS

$\Rightarrow S = \{ (x_i, f_i),\ i = 1, 2, \ldots, n \}$

SUPERVISED LEARNING

$x \xrightarrow[h]{} y$



hypothesis
predictor
classifier

How good/bad h?

$h(x) \longleftrightarrow y$

(x, y)

CLASS

$\mathbb{1}(h(x) \neq y)$

indicator function

LOSS FUNCTION

$$\ell(h, (x, y)) = \begin{cases} \mathbb{1}(h(x) \neq y) & \text{CLASSIF.} \\ (h(x) - y)^2 & \text{REGR.} \end{cases}$$

TOMORROW (x, y) FROM DISTRIBUTION $\mathcal{D}$

$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h,(x,y))]$ $\triangleq$ expected loss/risk

$L_{\mathcal{D}}(h)$ $R(h)$ $R_{\mathcal{D}}(h)$

FIND A GOOD PREDICTOR $\rightarrow$ MINIMIZE THE EXPECTED LOSS

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

IN ML $\mathcal{D}$ IS UNKNOWN

$\mathcal{H}$: class of hypotheses / predictors.

ML IS "DISTRIBUTION-FREE"

ML $\neq$ STATISTICS

ML ASSUMPTION:

WE DO NOT KNOW $\mathcal{D}$

BUT S IS DRAWN ACCORDING TO $\mathcal{D}$

EMPIRICAL LOSS

$$L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i)) \simeq \mathbb{E}_{(x,y)\sim D} [\ell(h, (x, y))]$$

MORE CORRECT THE LARGER IS n

Empirical Risk Minimization (ERM)

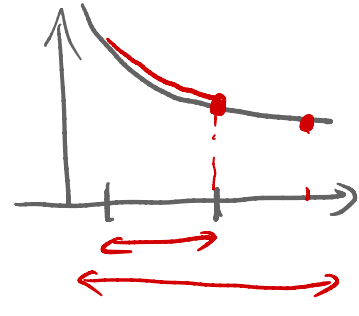
$$\underset{h \in H}{\text{minimize}} \; L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i))$$

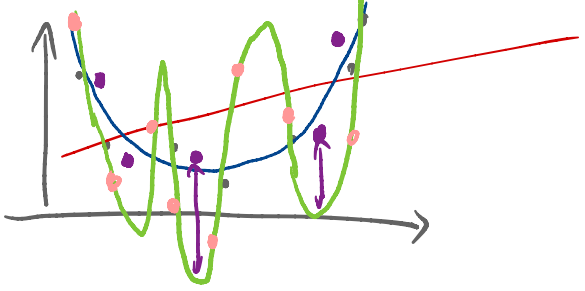
$H_1 = \{ax + b, a, b \in \mathbb{R}\}$ $\quad H_2 = \{ax^2 + bx + c, a, b, c \in \mathbb{R}\}$

$H_1 \subset H_2$

$h_1^* \in \underset{h \in H_1}{\operatorname{argmin}} L_S(h)$ $\quad h_2^* \in \underset{h \in H_2}{\operatorname{argmin}} L_S(h)$

$L_S(h_1^*) \geqslant L_S(h_2^*)$





GREEN IS OVERFITTING

GREEN IS FINE

BLUE IS UNDERFITTING

STATISTICAL LEARNING

$$|L_S(h) - L_S(h)| > C\sqrt{\frac{d + \ln\frac{1}{\delta}}{n}}$$

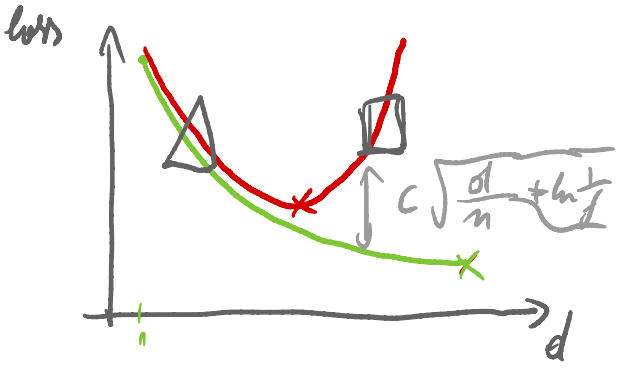
with probability δ

d : VC-dimension
measure of the complexity of the class H

Vapnik
Chervonenkis

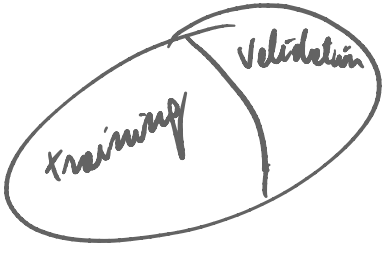
for the class of polynomials in one variable with degree m :

$$d = m + 1$$



L_S

L_D



S

$S = S_T \cup S_V$

$h_1^* \in \operatorname{argmin}_{h \in H_1} L_{S_T}(h)$

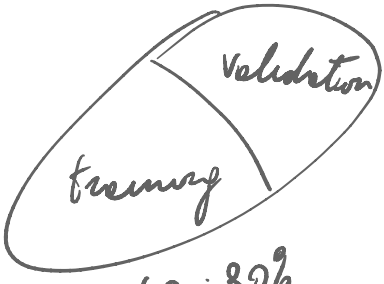
$h_2^* \in \operatorname{argmin}_{h \in H_2} L_{S_T}(h)$

$L_{S_T}(h_1^*) \geq L_{S_T}(h_2^*)$

$L_{S_V}(h_1^*) \lessgtr L_{S_V}(h_2^*)$

□ pick h_1^*
h_2^* is OVERFITTING

△ pick h_2^*
h_1^* is UNDERFITTING



60 ÷ 80%

K - fold cross validation

$$h_1^* \in \operatorname*{argmin}_{h \in H_1} L_{S_1}(h)$$

$$h_1^{*(1)} \in \operatorname*{argmin}_{h \in H_1} L_{S_2 \cup \dots S_K}(h)$$

$$h_1^{*(2)} \in \operatorname*{argmin}_{h \in H_1} L_{S_1 \cup S_3 \dots \cup S_K}(h)$$

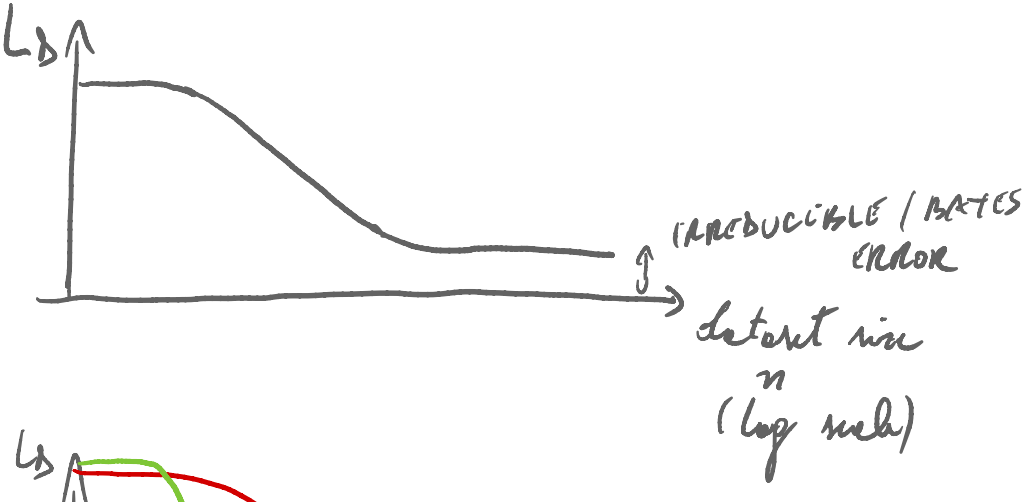
$$h_1^{*(K)} \in \operatorname*{argmin}_{h \in H_1} L_{S_1 \cup S_2 \dots \cup S_{K-1}}(h)$$

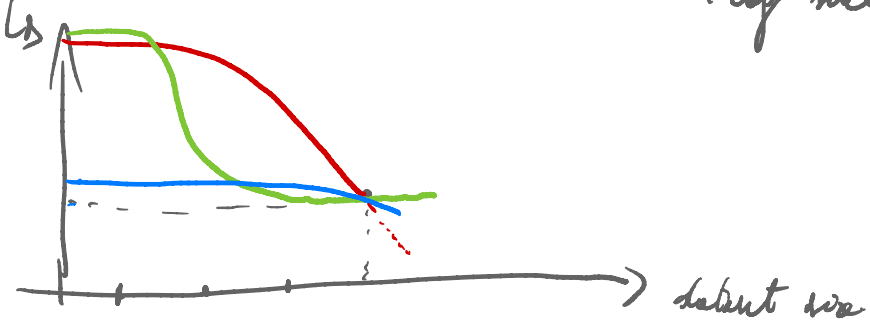
$$\text{Validation error} = \frac{1}{K} \sum_{i=1}^{K} L_{S_i}(h_1^{*(i)})$$

- learning step / rate
- batch size
- regularization coefficient

} HYPER PARAMETERS

LEARNING PROBLEM = SOLVING MANY OPTIMIZATION PROBLEMS





$$\min_{h \in H} L_S(h)$$

h is parameterized by a vector of parameters $w \in \mathbb{R}^d$

$$h(x) = \sum_i w_i x_i$$

$$L_S(h) = \sum_{i=1}^{n} \ell(h, (x_i, y_i)) =$$

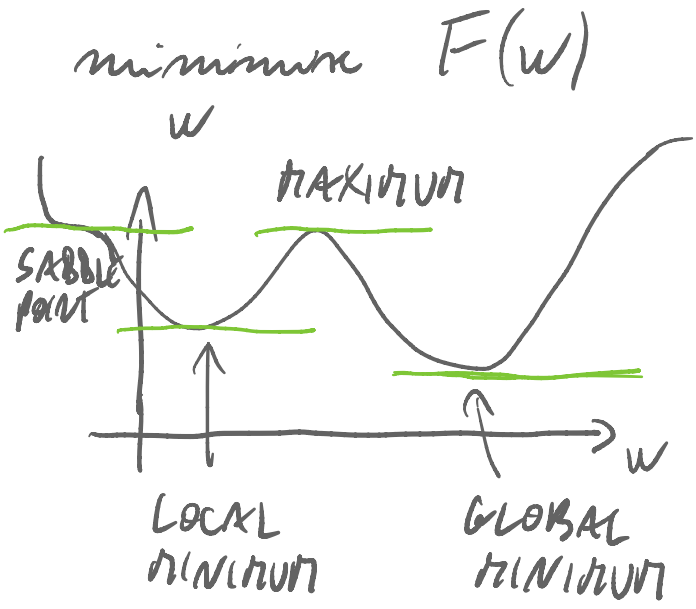
ex quadratic loss

$$= \sum_{i=1}^{n} \left(h(w, x_i) - y_i \right)^2$$

$$\ell(h(x_i, y_i)) = f(w, i)$$

$$L_S(h) \longleftrightarrow F(w)$$

$$\min_{w \in W} F(w) \longrightarrow \min_{w \in \mathbb{R}^d} F(w)$$



$F'(w)$

$$\nabla F(w) = \begin{bmatrix} \frac{\partial F}{\partial w_1} \\ \vdots \\ \frac{\partial F}{\partial w_d} \end{bmatrix} \in \mathbb{R}^d$$

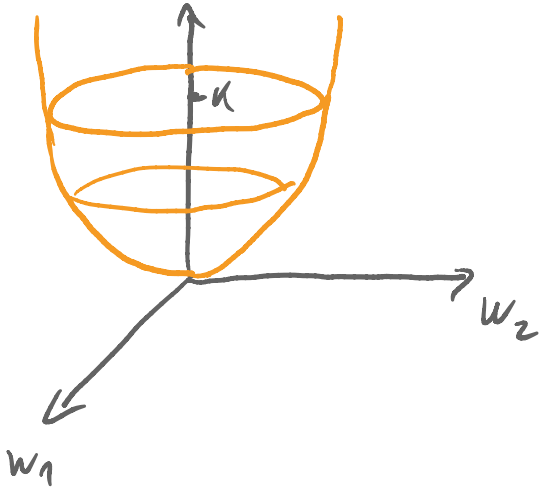
$$F(w) = w_1^2 + w_2^2$$

$$\frac{\partial F}{\partial w_1} = 2w_1 + 0 = 2w_1 \qquad \frac{\partial F}{\partial w_2} = 2w_2$$

$$\nabla F(w) = \begin{bmatrix} 2w_1 \\ 2w_2 \end{bmatrix}$$

AT A CRITICAL POINT $\iff \nabla F = 0$

$$\begin{bmatrix} 2w_1 \\ 2w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow w_1 = w_2 = 0$$



$$F(w_1, w_2) = K$$

$$w_1^2 + w_2^2 = K$$