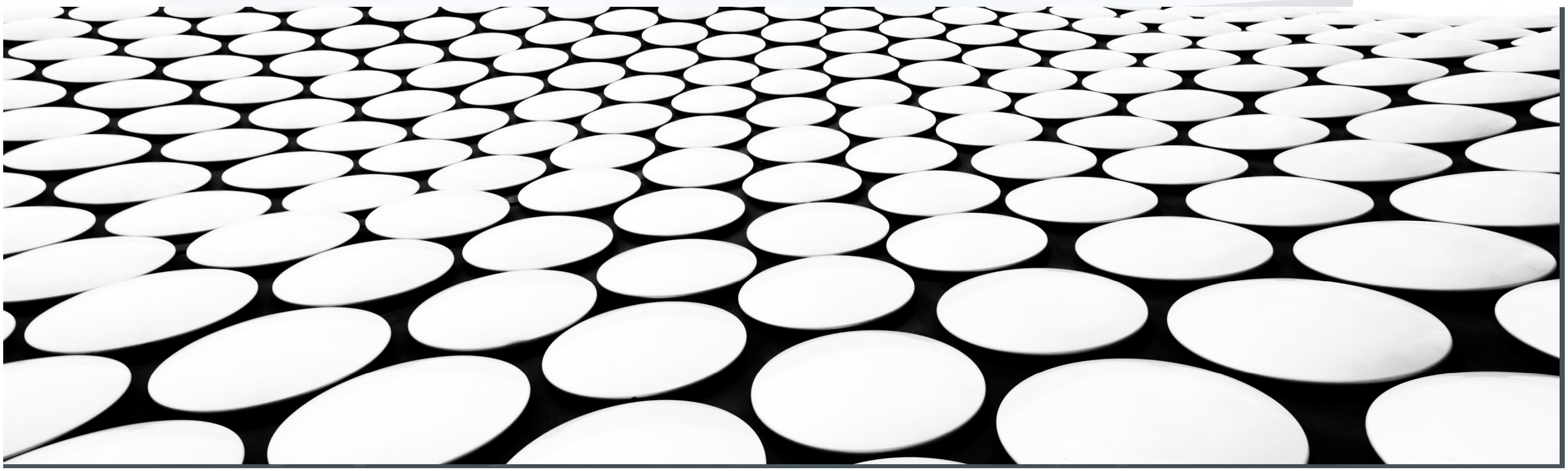
INTRODUCTION TO STATISTICAL LEARNING THEORY

ROBERTO BASILI, UNIVERSITÀ DI ROMA, TOR VERGATA

DEEP LEARNING, MARCH 2024



OUTLINE

- Statistical Learning Theory
  - PAC learnability
  - VC dimension
  - Learning Machines
  - Model Optimization and Concept Class complexity
  - Model Optimization via Cross-Validation
  - Towards perceptrons and SVMs
-

FROM STATISTICAL LEARNING THEORY TO SVMS

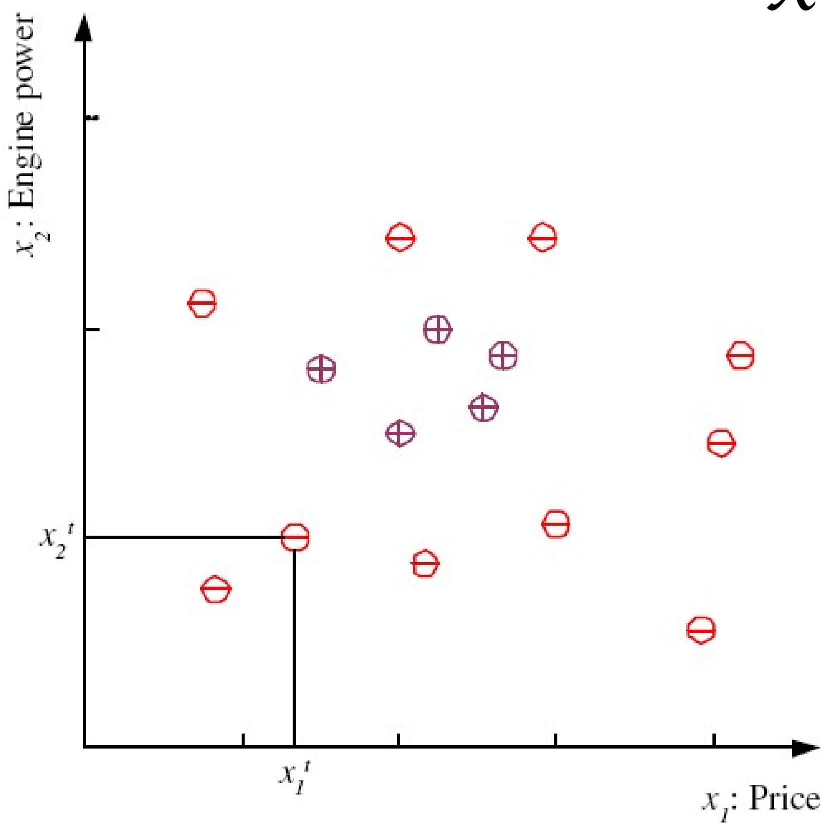
LEARNİNG A CLASS FROM EXAMPLES

- Class C of a "family car"
  - Prediction Is car x a "family car"?
  - Knowledge extraction What do people expect from a family car?
- Output:

  Positive (+) and negative (–) examples
- Input representation:

  x_1: price, x_2 : engine power

TRAİNİNG SET $\mathcal{X}$



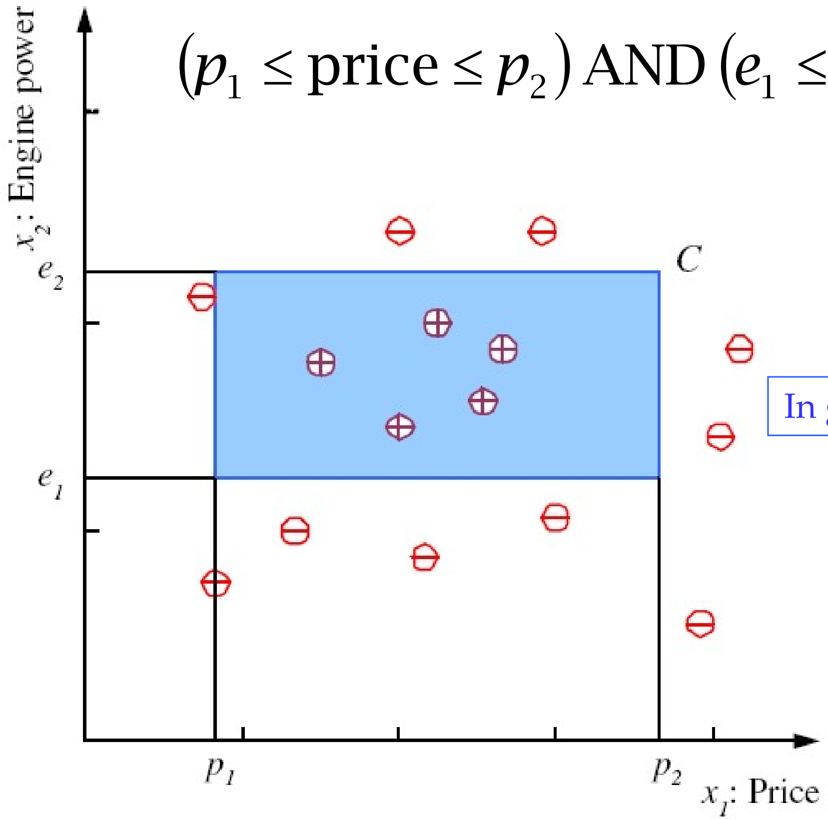
$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$$

label

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$r = \begin{cases} 1 \text{ if } \boldsymbol{x} \text{ is positive} \\ 0 \text{ if } \boldsymbol{x} \text{ is negative} \end{cases}$$

CLASS C

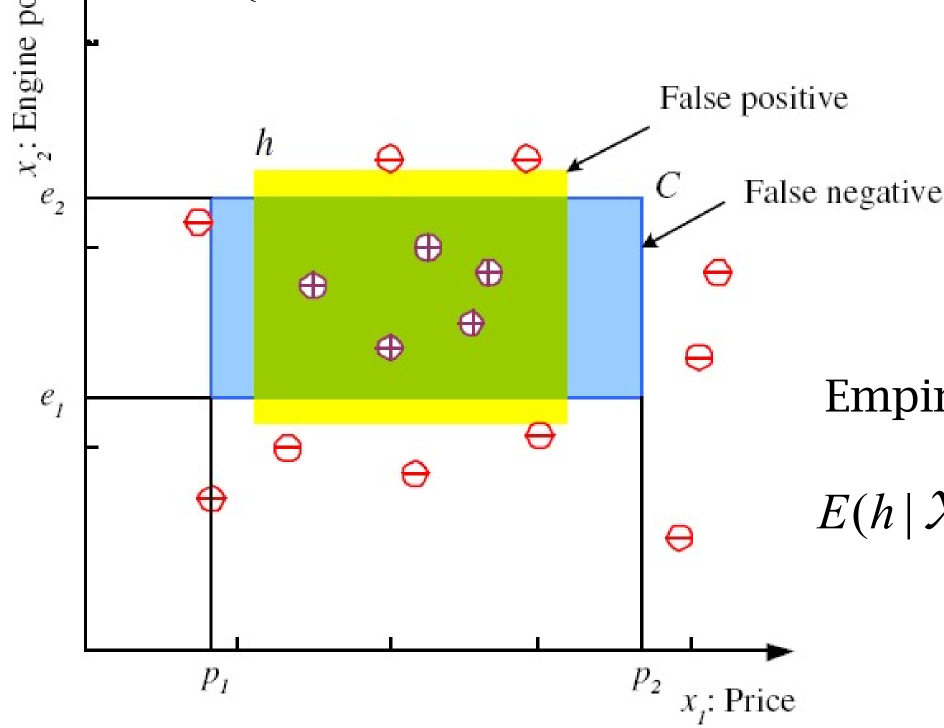


$$\left(p_1 \le \text{price} \le p_2\right) \text{ AND } \left(e_1 \le \text{engine power} \le e_2\right)$$

In general we do not know C(x).

HYPOTHESİS CLASS $\mathcal{H}$

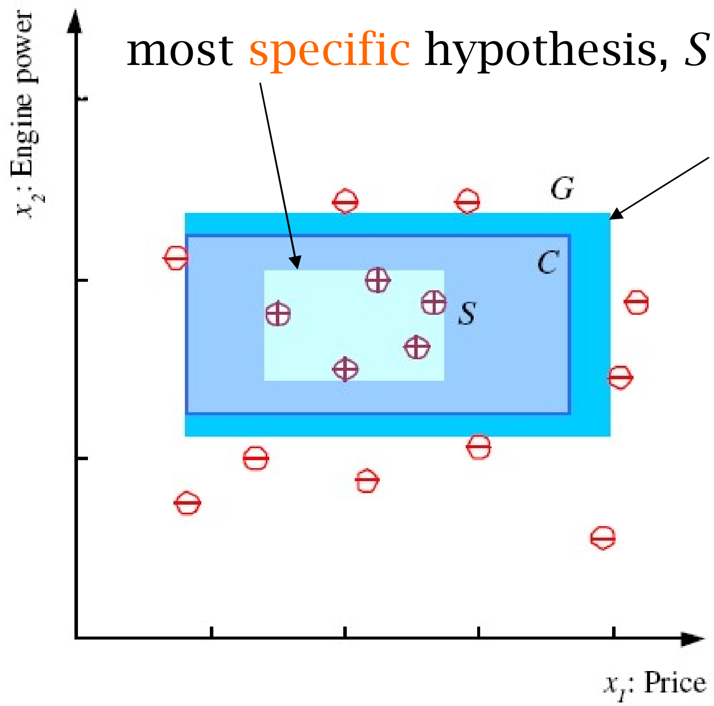
$$h(\boldsymbol{x}) = \begin{cases} 1 \text{ if } h \text{ classifies } \boldsymbol{x} \text{ as positive} \\ 0 \text{ if } h \text{ classifies } \boldsymbol{x} \text{ as negative} \end{cases}$$



Empirical error:

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

S, G, AND THE VERSİON SPACE



most general hypothesis, G

$h \in \mathcal{H}$, between S and G is consistent and make up the version space

(Mitchell, 1997)

PROBABLY APPROXİMATELY CORRECT (PAC) LEARNİNG

- How many training examples are needed so that the tightest rectangle S which will constitute our hypothesis, will probably be approximately correct?
  - We want to be *confident* (*above a level*) that $1\text{-}\delta$
  - ... the *error probability is bounded* by some value ε

- A concept class *C* is called *PAC-learnable* if there exists a PAC-learning algorithm such that, for any $\varepsilon>0$ and $\delta>0$, there exists a fixed sample size such that, for any concept $c \in C$ and for any probability distribution on X, the learning algorithm produces a probably-approximately-correct hypothesis *h*
- a (PAC) *probably-approximately-correct hypothesis h* is one that has error at most ε with probability at least $1\text{-}\delta$.

PROBABLY APPROXİMATELY CORRECT (PAC) LEARNİNG

- In PAC learning, given a class C and examples drawn from some unknown but fixed distribution p(x), we want to find the number of examples N, such that with probability at least 1-δ, *h* has error at most ε ? (Blumer et al., 1989)

- $P(C\Delta h \leq \varepsilon) \geq 1- \delta$

- where $C\Delta h$ is (area of the) “the region of difference between *C* and *h*”, and $\delta>0$, $\varepsilon>0$.

PAC LEARNİNG

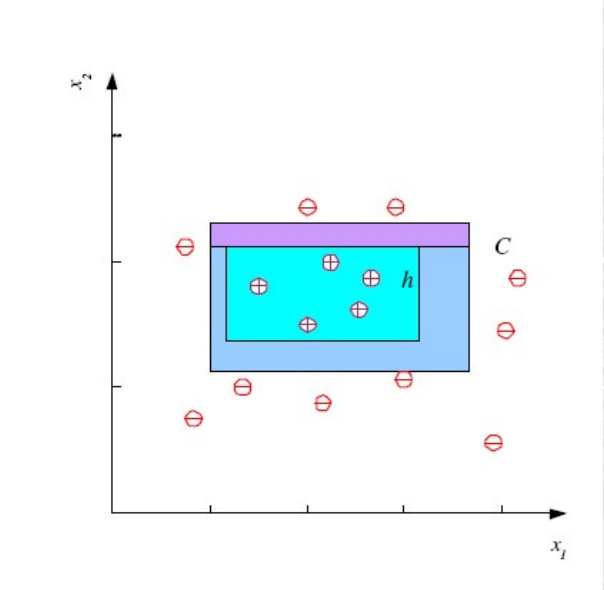
How many training examples m should we have, such that with probability at least $1 - \delta$, h has error at most ε ? (Blumer et al., 1989)

- Let prob. of a + ex. in each strip be at most $\varepsilon/4$
- Pr that a random ex. misses a strip: $1- \varepsilon/4$
- Pr that m random instances miss a strip:
  $(1 - \varepsilon/4)^m$
- Pr that m random instances instances miss 4 strips:
  $4(1 - \varepsilon/4)^m$
- We want $1-4(1 - \varepsilon/4)^m \geq 1-\delta$ or $4(1 - \varepsilon/4)^m \leq \delta$
- Using $1-x \leq e^{-x}$ an even stronger condition is:

[$(1-\varepsilon/4) \leq exp(-\varepsilon/4)$ so $(1-\varepsilon/4)^m \leq exp(-\varepsilon/4)^m = exp(-\varepsilon m/4)$]

$4e^{-\varepsilon m/4} \leq \delta$ OR

- Divide by 4, take ln... and show that $m \geq (4/\varepsilon)ln(4/\delta)$



PROBABLY APPROXİMATELY CORRECT (PAC) LEARNİNG

How many training examples m should we have, such that with probability at least 1 - δ, our hypothesis *h* has error at most ε ?   (Blumer et al., 1989)

$$m \geq (4/\varepsilon)\ln(4/\delta)$$

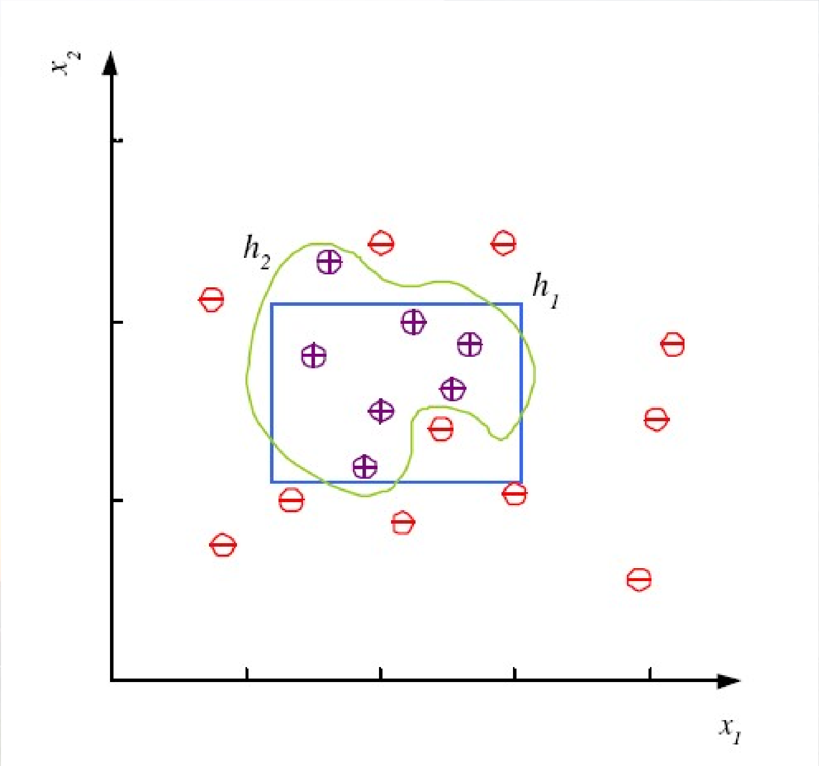
- *m* increases slowly with $1/\varepsilon$ and $1/\delta$

- Say ε=1% with confidence 95%, pick $m \geq 1752$

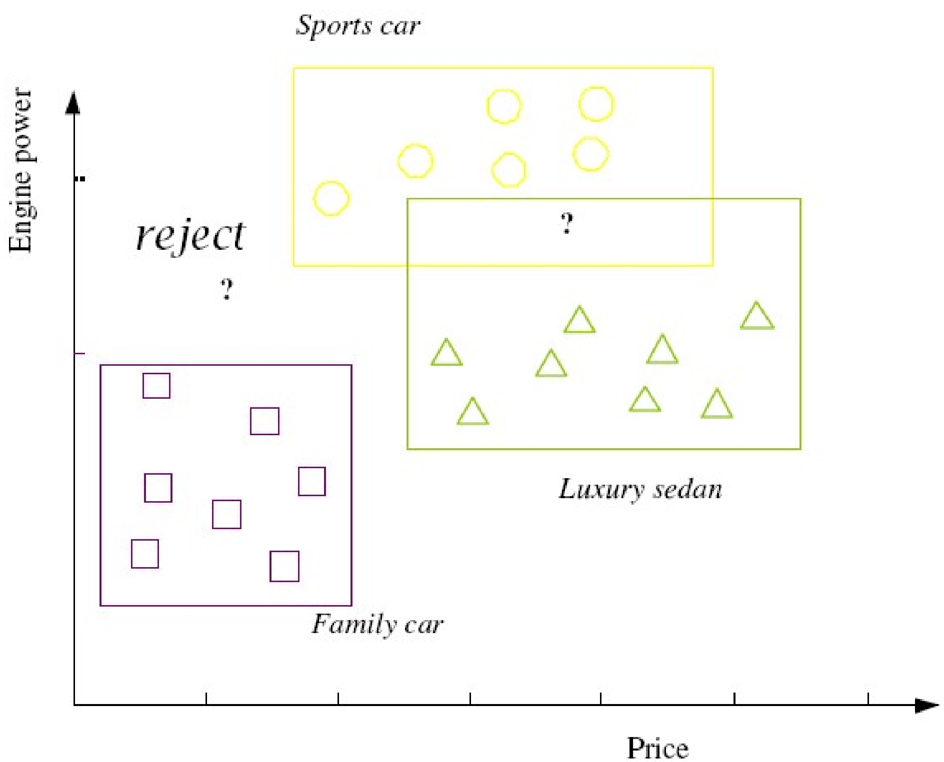
- Say ε=10% with confidence 95%, pick $m \geq 175$

MODEL COMPLEXİTY VS. NOISE

- Use the simpler one because
- Simpler to use (lower computational complexity)
- Easier to train (lower space complexity)
- Easier to explain (more interpretable)
- Generalizes better (lower variance – Occam's razor)



MULTIPLE CLASSES, C_i i=1,...,K



$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$$

$$r_i^t = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j,\ j \neq i \end{cases}$$

Train hypotheses
$h_i(\boldsymbol{x})$, $i = 1,...,K$:

$$h_i(\boldsymbol{x}^t) = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j,\ j \neq i \end{cases}$$

REGRESSİON

$$X = \{x^t, r^t\}_{t=1}^{N}$$
$$r^t \in \Re$$
$$r^t = f(x^t) + \varepsilon$$

$$E(h'|X) = \frac{1}{N}\sum_{t=1}^{N}[r^t - h'(x^t)]^2$$

$$E(h|X) = E(w_1, w_0|X) = \frac{1}{N}\sum_{t=1}^{N}[r^t - (w_1 x^t + w_0)]^2$$