Principles of Machine Learning

CSCI-B455

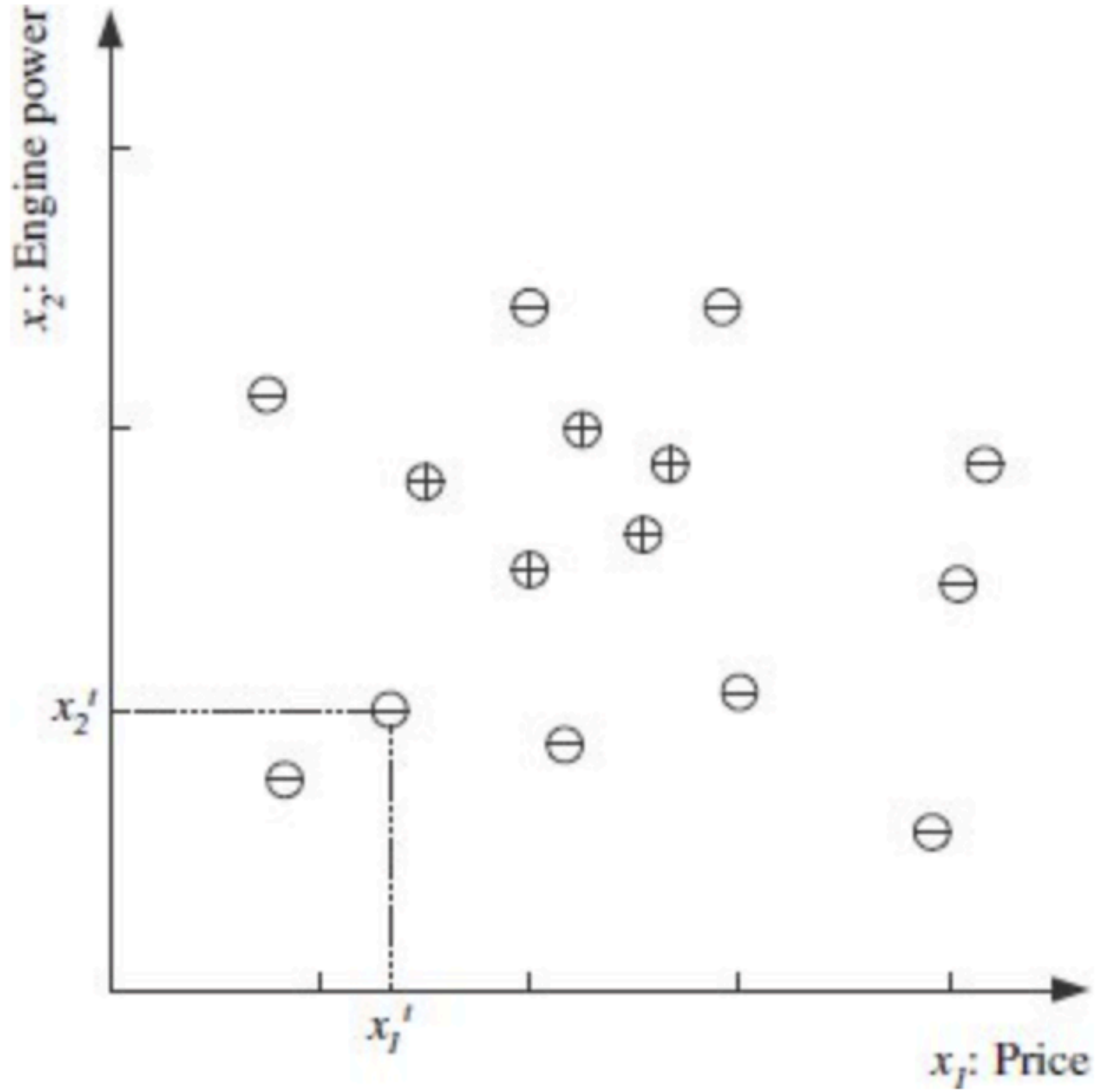
Supervised Learning — I

M. Oğuzhan Kulekci

A simple classification task

- Class **C** represents the **family cars**, and we aim to learn this class.
- Given a car, can we decide whether it is a family car ?
- Output is **yes (positive)** or **no (negative)**, a **binary classification**
- The purpose might be the
  - **Prediction:** When we see a new car, we want to answer the query
  - **Knowledge extraction:** The car manufacturer wants to get a good definition of a `family car’
- Assume we decided to use the **price** and **engine power** as the input representation, the attributes or dimensions of the input
- We have previously **labeled** data, the cars in class C and outside of it, to use in our learning

Two-Class or Binary Classification



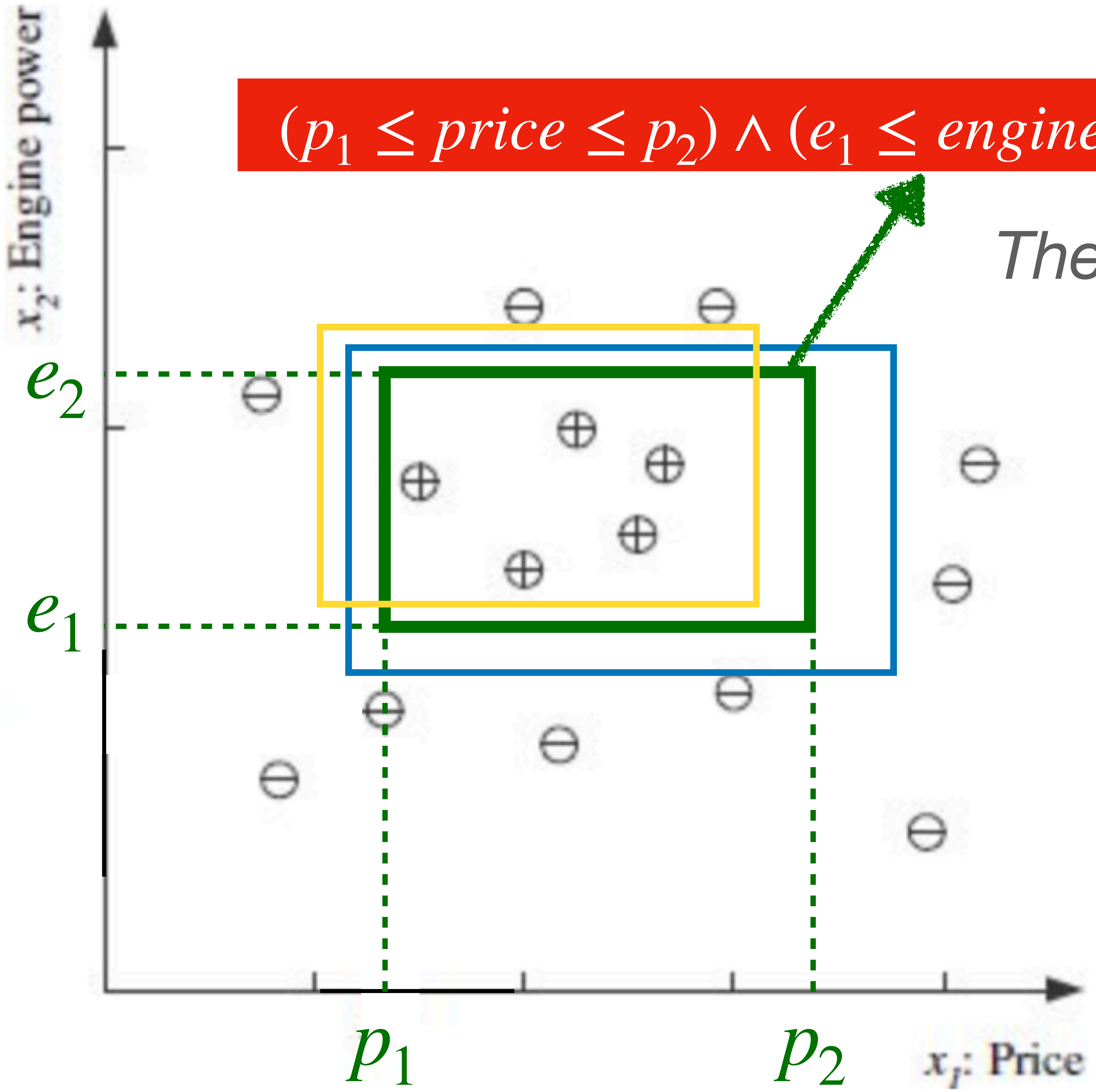
$$\chi = \{x_1^t, x_2^t, r^t\}_{t=1}^{N}$$

N **distinct** car samples the training set, where

- x_1^t is the price of the car t .
- x_2^t is the engine power of the car t .
- r^t is 1, if the car t is a family car, otherwise 0.

How would you mark the family car area on the figure left ?

Binary Classification



There can be many other guesses, right ?

Hypothesis class $\mathcal{H} = \{h_1, h_2, \ldots\}$

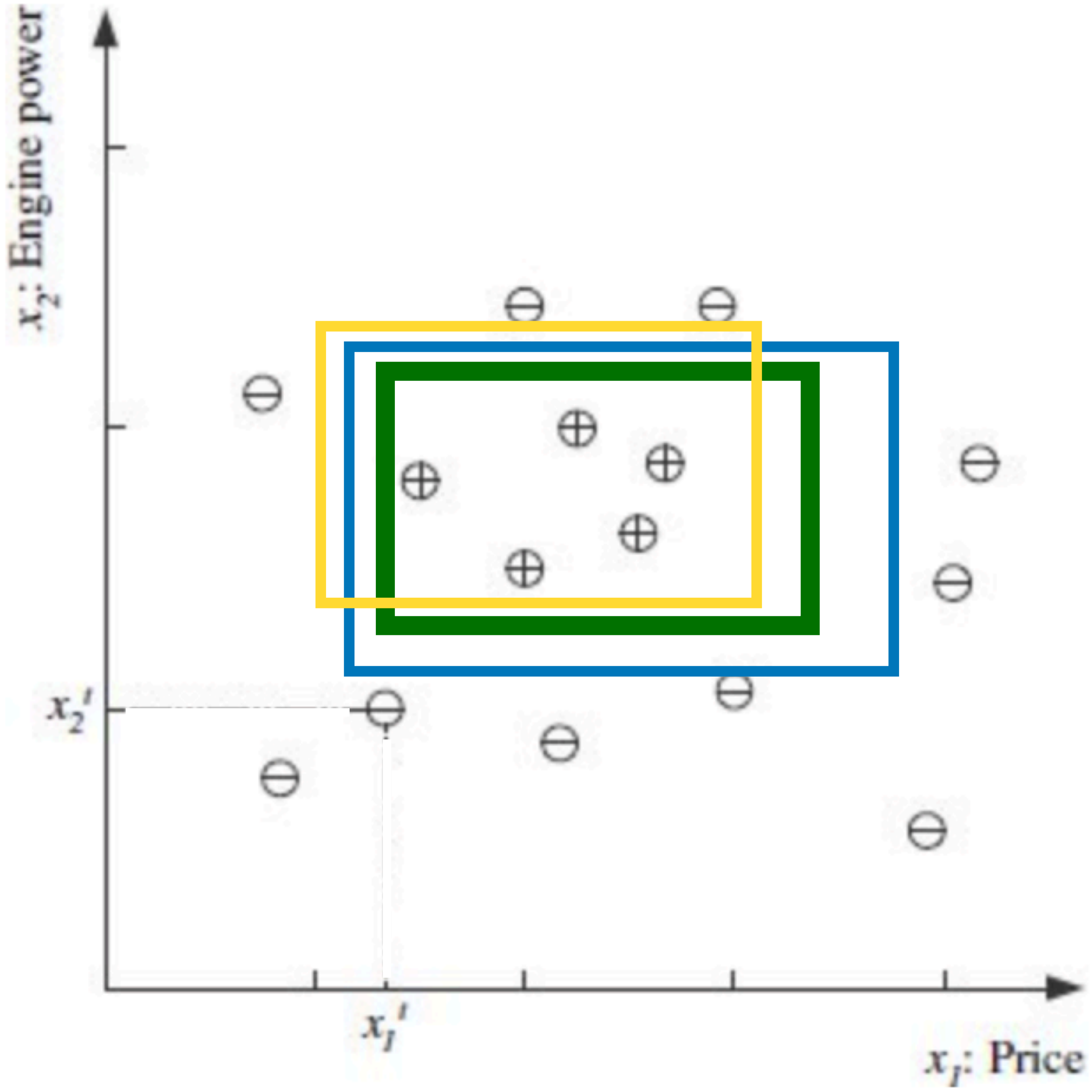
- $h_i \in \mathcal{H}$ is defined with $\langle p_1^i, p_2^i, e_1^i, e_2^i \rangle$.
- The learning algorithm just returns a $h_i \in \mathcal{H}$
- Is it the best one ? We never know !
- We can only try to minimize the **empirical error** on **training** set.

Binary Classification

Empirical error of a hypothesis h on training set $\mathscr{X}$ is

$$E(h\,|\,\mathscr{X}) = \sum_{a=1}^{N} 1 \cdot (\ h(x^t) \neq r^t\ )$$

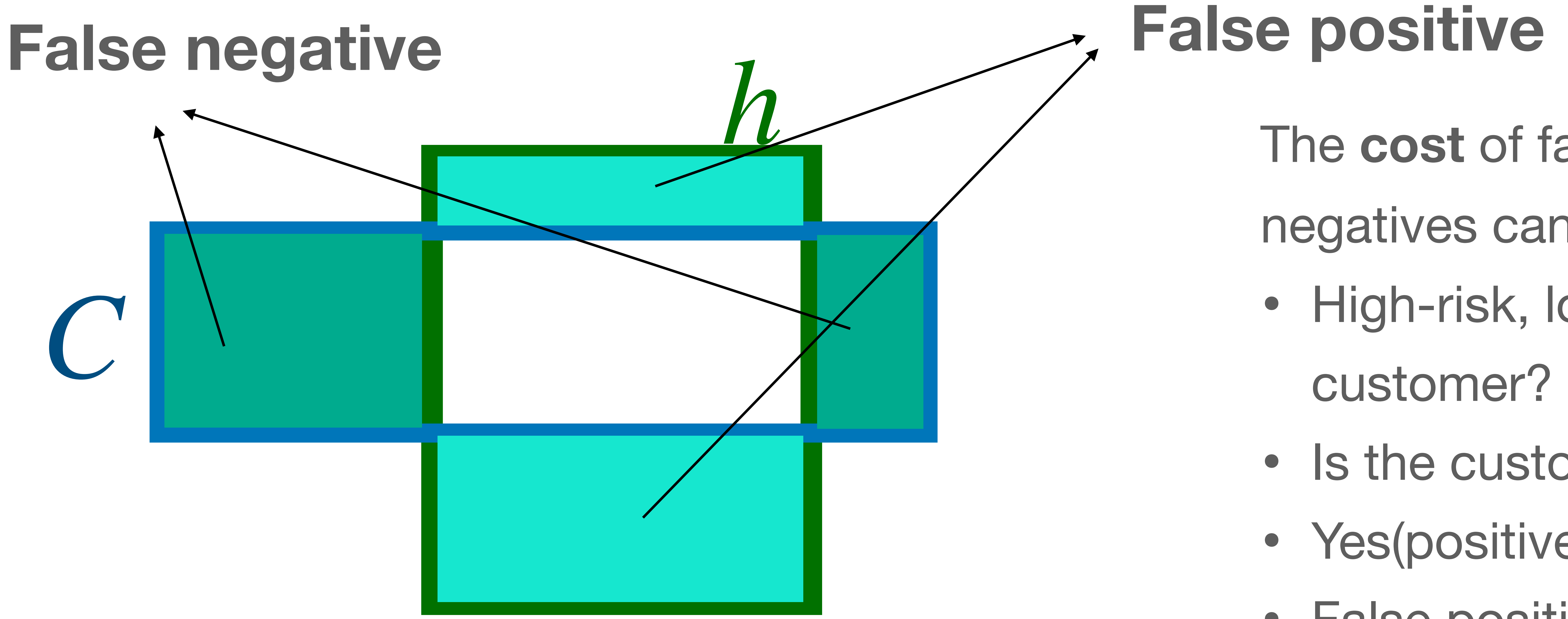
$(\ h(x^t) \neq r^t\ )$ is 1 if the output of h on x^t is not equal to r^t, else 0.



- Possibly many hypotheses h without error, $E(h\,|\,\mathscr{X}) = 0$
- **How will you choose among them ?**
- The future performance of chosen h cannot be known.
- This is the **generalization problem** in learning.

What can be the future implications of hypothesis h?

Binary Classification



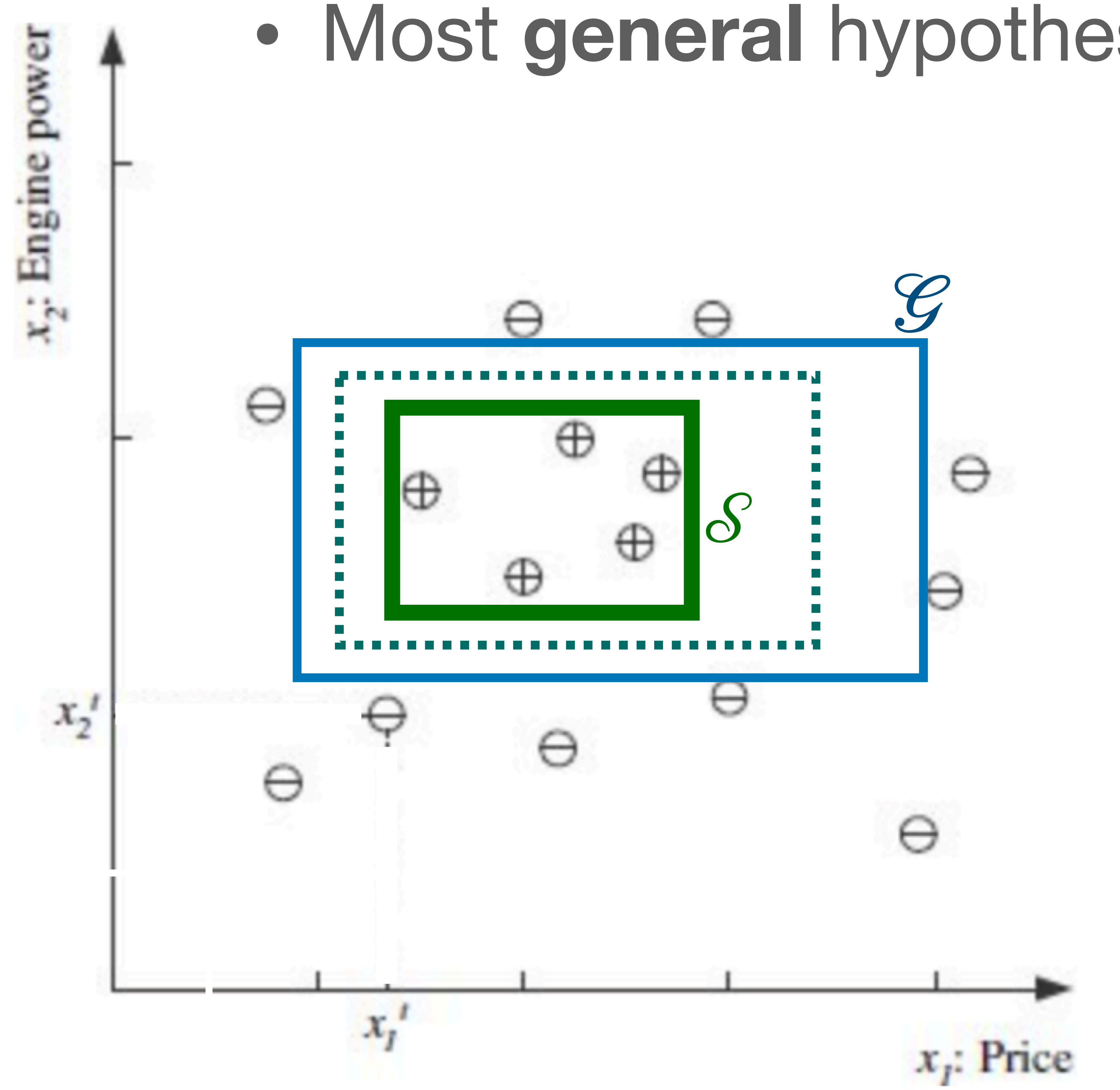
h is the learned hypothesis, and C is the actual class (ground truth)

The **cost** of false positive and false negatives can be **different**.

- High-risk, low-risk analysis for a credit customer?
- Is the customer a low-risk one ?
- Yes(positive), No (negative)
- False positive: A high-risk customer label as a low-risk. Credit granted and lost
- False negative: A low-risk customer labeled high-risk. Credit not granted, bank lost the opportunity of profit.

Binary Classification

- Most **specific** hypothesis $\mathcal{S}$: The smallest rectangle including all positives
- Most **general** hypothesis $\mathcal{G}$: The largest rectangle excluding all negatives

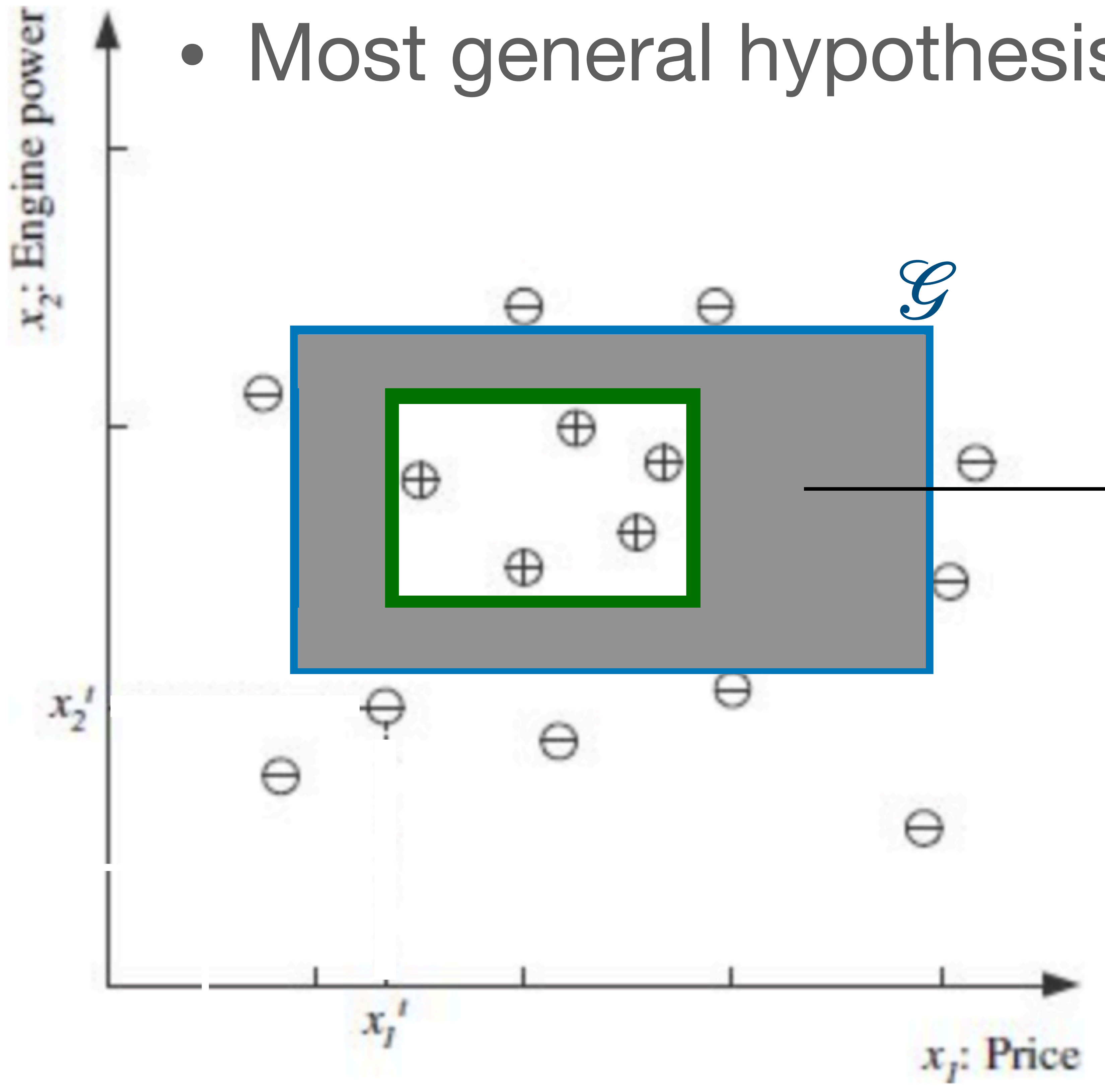


- All h in between $\mathcal{S}$ and $\mathcal{G}$ is called the **version space.**
- Any such h is **consistent** without error on $\mathcal{X}$.
- Having h in half-way between $\mathcal{S}$ and $\mathcal{G}$ seems intuitive.
- Why? Because, it increases the **margin, distance between boundary and its closest instances.**
- *To learn such an h, it needs a fix on the empirical error calculation !*

- $h(x)$ should return a distance value rather than 0/1
- This return value should be used in a loss function for the optimization
- **SVM**, support-vector-machine, is the typical example

Binary Classification

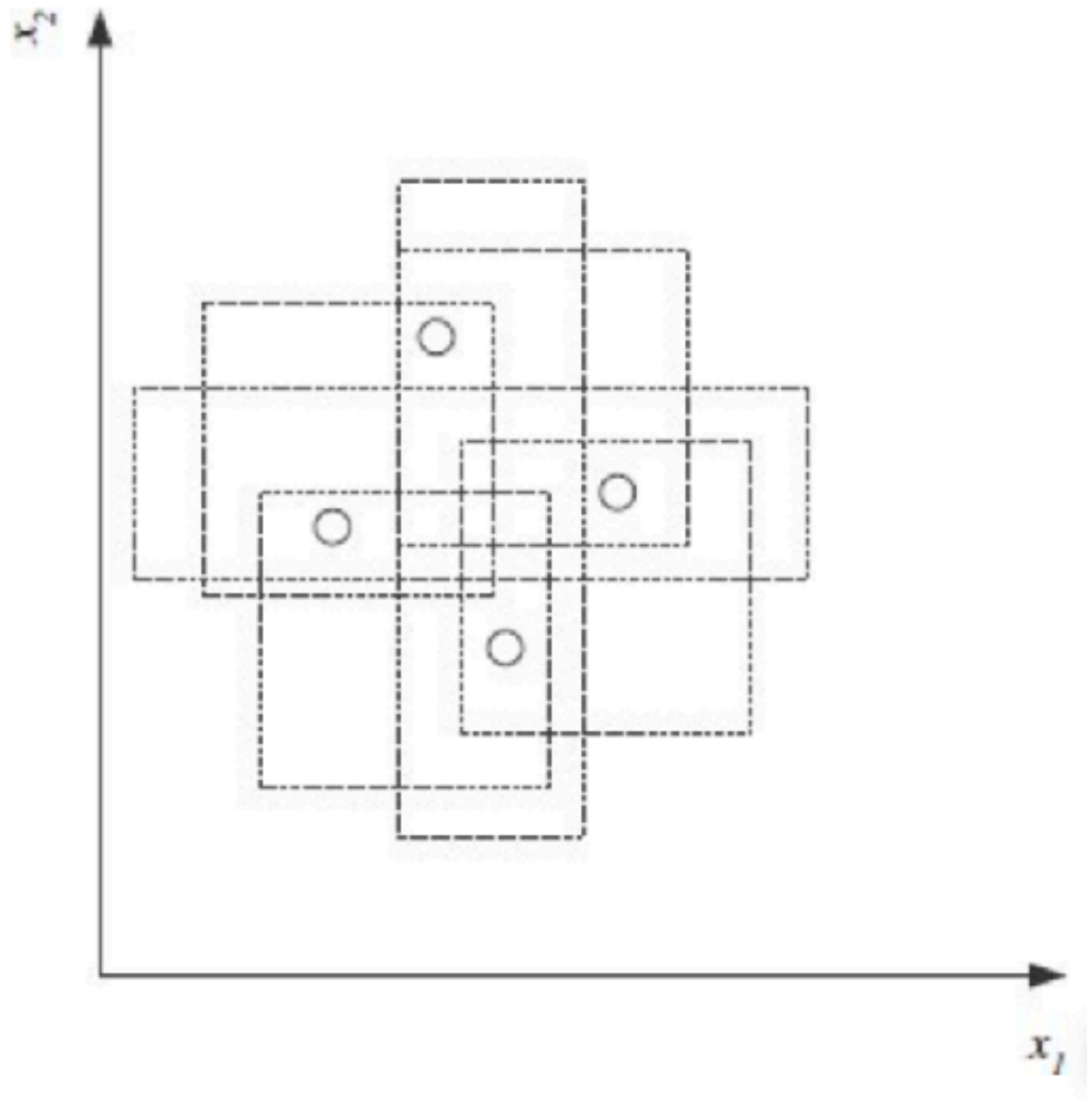
- Most specific hypothesis $\mathscr{S}$: The smallest rectangle including all positives
- Most general hypothesis $\mathscr{G}$: The largest rectangle excluding all negatives



In case, the consequences of false positives or negatives are very serious, the area in between $\mathscr{S}$ and $\mathscr{G}$ is assumed the **doubt** region, and forwarded to <u>human judgement</u>

Vapnik-Chervonenkis (VC) Dimension

- A measure of the capacity or the expressive power of a hypothesis set $\mathscr{H}$
- N points define 2^N different +/- attribution, each of which is a learning problem
- If there's an $h \in \mathscr{H}$ that solves all possible attributions, then $\mathscr{H}$ **shatters** N points.
- This means the learning problem defined by N points can be solved without error.



Example:

Assume $\mathscr{H}$ consists of axis-aligned rectangles, i.e. we draw an axis-aligned rectangle to separate positives from negatives or vice versa.

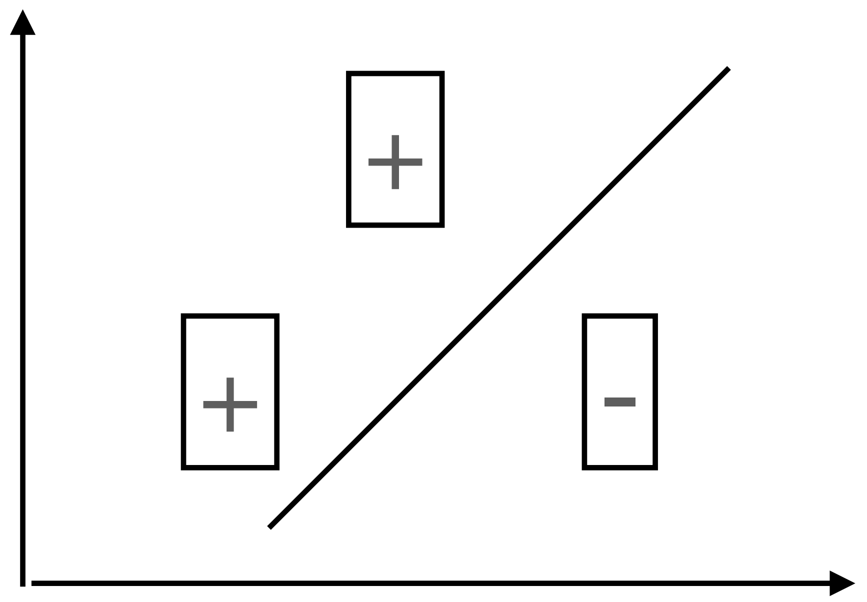
What is the VC-dimension of $\mathscr{H}$?

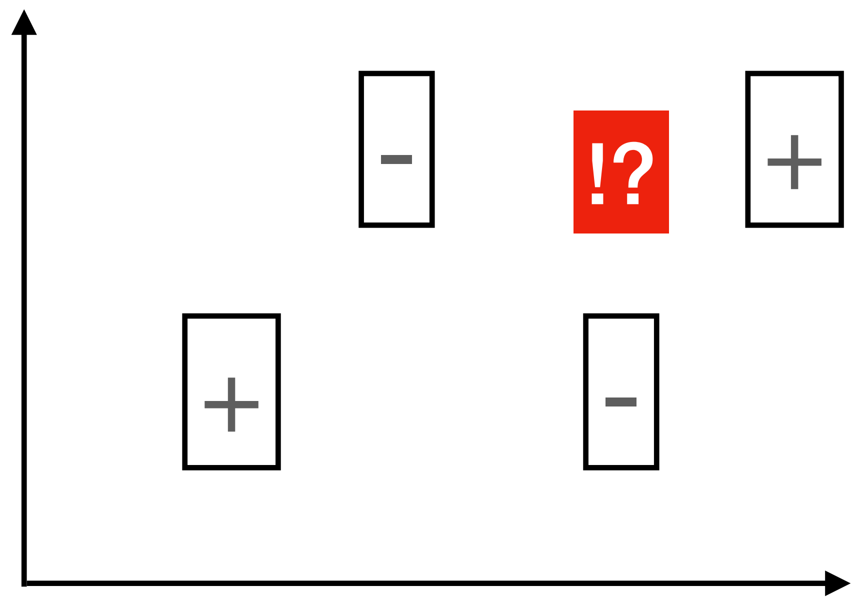
For the 4-points in 2D space, there is an axis-aligned rectangle that does the job for every possible +/- labeling of the points.
But, it does not hold for 5 points, so VC-dimension is 4.

Very pessimistic, it says you can only learn 4-point data sets with $\mathscr{H}$. Does it really that limiting in real-world?

Vapnik-Chervonenkis (VC) Dimension

- Assume $\mathscr{H}'$ is consist of drawing lines in 2D space, instead of axis-aligned rectangles.
- What is the VC dimension of $\mathscr{H}'$?
- There are $N = 3$ points all possible +/- assignments can be separated by a single line.
- Not possible to *shatter* for $N = 4$. Thus, VC dimension of $\mathscr{H}'$ is 3.



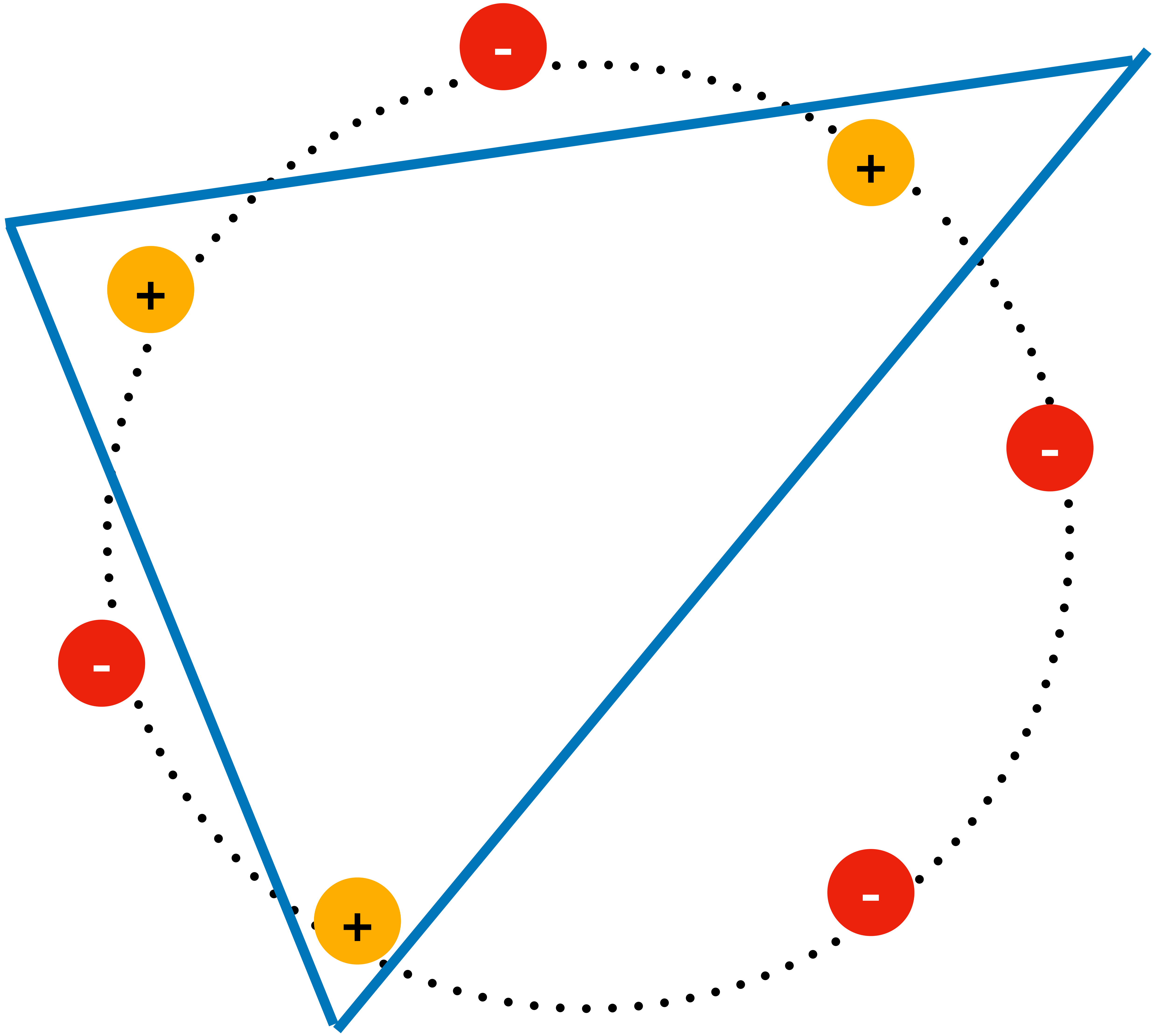


The VC dimension of $\mathscr{H}$ is 4 while $\mathscr{H}'$ is 3. What does it mean ?

The expressive power, the capacity of learning, of $\mathscr{H}$ is superior to $\mathscr{H}'$. $\mathscr{H}$ can learn more complex data sets compared to $\mathscr{H}'$.

Vapnik-Chervonenkis (VC) Dimension

- What is the VC dimension of a triangle on 2D space (or plane) ?
- 7 points in a plane can be shattered by a triangle. Thus, VC dimension of triangle on 2D is 7.



- *However, VC is still very pessimistic. With axis-aligned rectangles you can only learn 4-point data sets and with lines only 3-point data sets !*
- *Real-world practice is different, data points are not random, we are free to have errors on the training data, etc…*

Probably Approximately Correct (PAC) Learning

- **Accuracy:** We want the hypothesis h to be close to the true target class C.
- A **probability** that a given point is misclassified is **at most ϵ, i.e., the error is at most ϵ**
- The **level of desired accuracy** is denoted by ϵ.

	h : If $(weight > 115)$ then Orange, else Apple	C : If $(weight > 125)$ then Orange else Apple
Training Data	Learned	True Class
1. Apple 120g	**1. Orange**	**1. Apple**
2. Orange: 150g	2. Orange	2. Orange
3. Apple: 130g	3. Orange	3. Orange
4. Orange: 140g	4. Orange	4. Orange
5. Apple:110g	5. Apple	5. Apple

h deviates from C by 20%, accuracy is 80%.

Error rate is then $\epsilon = 0.2$.

Does h provide a confidence ?

- **Confidence:** We want to be confident that the h provides the desired level of accuracy.
- The level of desired accuracy is maintained with at least $(1 - \delta)$ probability.