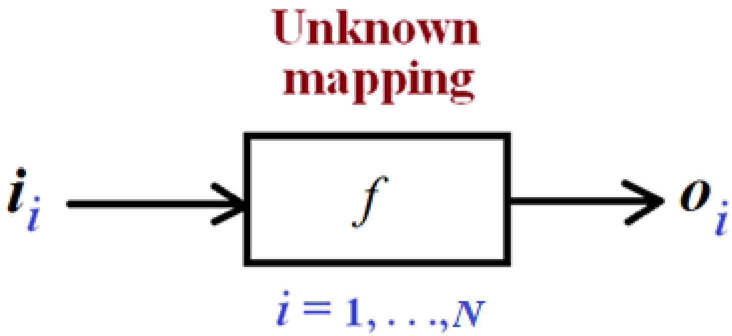
CH. 2: Supervised Learning

2.1 Introduction

Idea: Supervised learning learns an unknown mapping f through a set of (input $\boldsymbol{i}$, output $\boldsymbol{o}$) pairs.



f can be a numerical function or a classifier.

Given a training sample, $X = \{(\boldsymbol{i}_i, \boldsymbol{o}_i)\}_{i=1}^{N}$, where $(\boldsymbol{i}_i, \boldsymbol{o}_i)$ is a training example, figure out f.

Example: Given a training sample,

$$X = \{(i_i, o_i)\}_i = \{(3,15),\ (7,19),\ (5,17), \cdots\cdots\}$$

$$i_i \longrightarrow \boxed{f} \longrightarrow o_i$$

and a set of hypotheses, $H = \{h_k\}$, of f

$$(i_1, o_1) = (3,\ 15) \Rightarrow\ h_1(x) = 5x,\ h_2(x) = x + 12$$

$$h_3(x) = 2x^2 - 4x + 9, \text{ -----}$$

$$(i_2, o_2) = (7,19) \Rightarrow\ h_2(x) = x + 12, \text{ -------------}$$

$$(i_3, o_3) = (5,17) \Rightarrow\ h_2(x) = x + 12, \text{ -------------}$$

- -

$$(i_i, o_i) \text{ ---}$$

Questions:

i) Whether a hypothesis set $H = \{h_1, h_2, \cdots, h_m\}$ is provided? How about if $m \to \infty$?

ii) How many training examples are required?

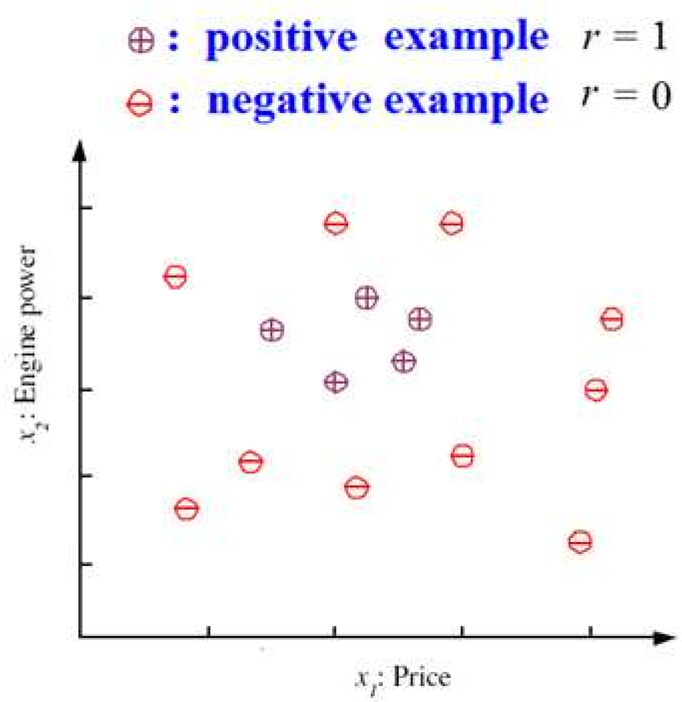
iii) How to select an h from H?

- **Example**: Learn a classifier that discriminates the class c of "family cars" from different classes of cars.

Car representation: $\boldsymbol{x} = (x_1,\ x_2)^T$ where

x_1: price,

x_2: engine power

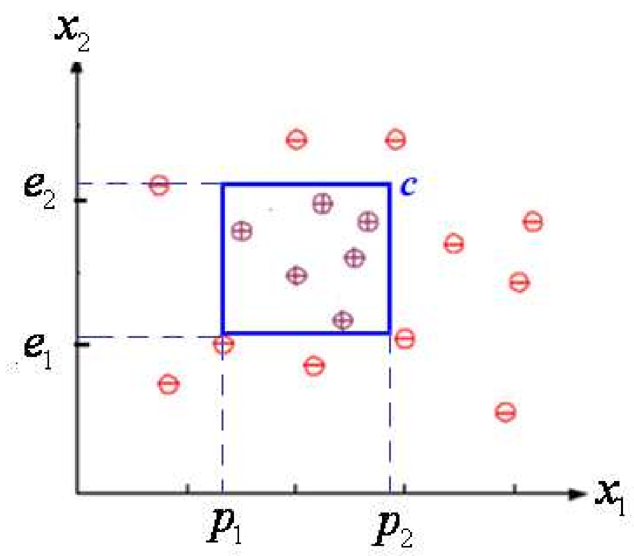


Given a training sample

$$S = \{(\boldsymbol{x}_i, r_i)\}_{i=1}^{N}$$

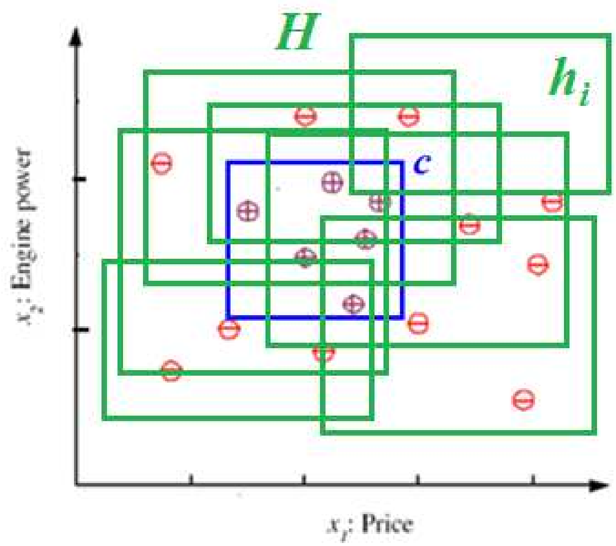
Let $$r = \begin{cases} 1 & \text{class } c \\ 0 & \text{otherwise} \end{cases}$$

Suppose a family car with price x_1 and engine power x_2 should be in a certain range $(p_1 \le x_1 \le p_2) \wedge (e_1 \le x_2 \le e_2)$.

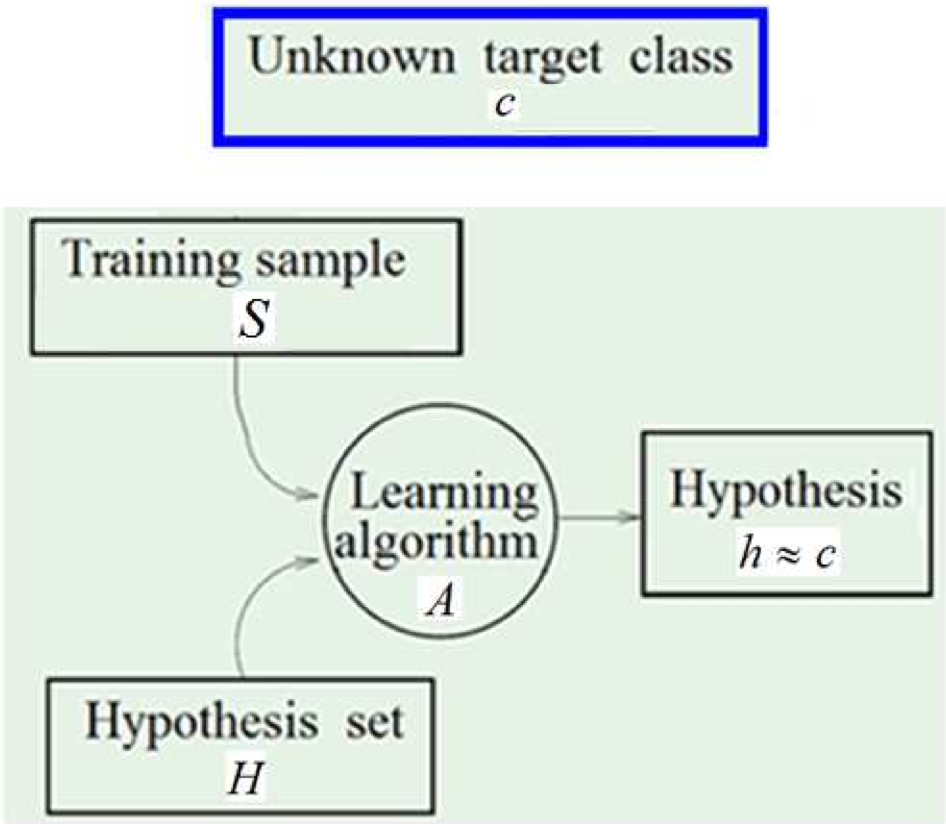


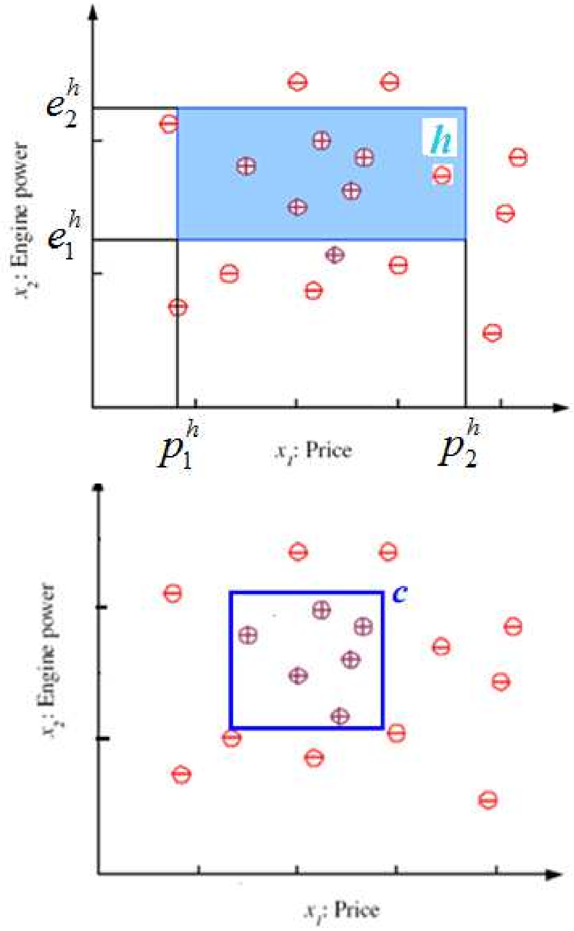
We then define the hypothesis set H as the set of rectangles.

$$H = \{h_i \mid h_i = (p_1^i \le x_1 \le p_2^i) \wedge (e_1^i \le x_2 \le e_2^i)\}$$



Learning process: Given S and H, a learning algorithm A attempts to find a hypothesis $h^* \in H$ that minimizes the empirical error $E(h) = \sum_{t=1}^{N} 1(h(\boldsymbol{x}^t) \neq r^t)$, $h^* = \arg\min_h E(h)$.





2.2 Learnability

In the supervised learning process, three components:

S (sample), *H* (hypothesis set), *A* (learning algorithm)

will determine the learnability of the process.

Probably approximately correct (**PAC**) learning

⇨ answer the complexity of sample set *S*

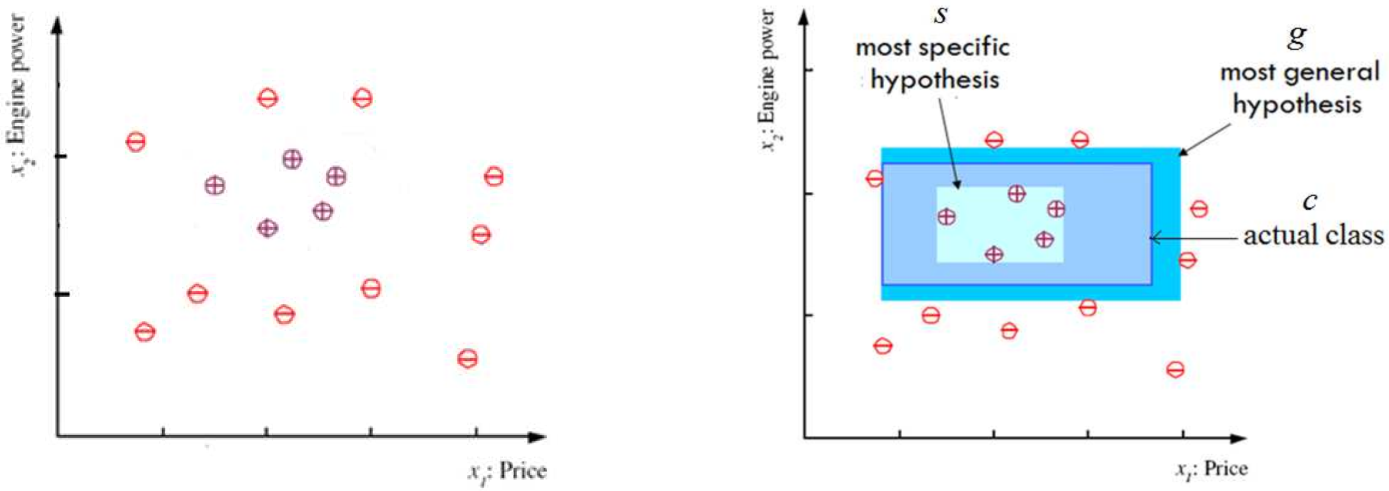
Vapnik-Chervonenkis (**VC**) dimension

⇨ answer the complexity of hypothesis set *H*

2.2.1 PAC Learning

-- Answer to the sample complexity, i.e., the number of training examples required to achieve a satisfied (PAC) answer.

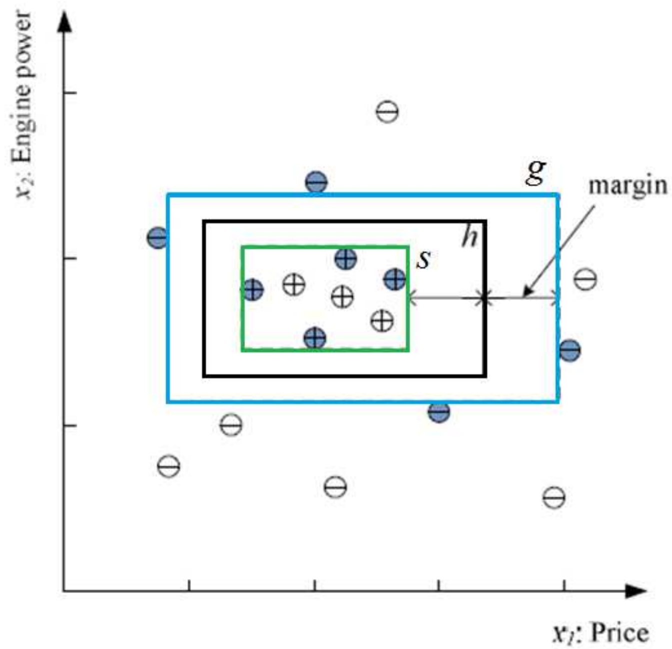
- Hypotheses s, g, and Version Space V



Version space: $V = \{h \mid h \text{ is between } s \text{ and } g\}$

In general, the hypothesis $h \in V$ that has the most margin between s and h is selected as the result.

Hypotheses s and g can be known from the given training set X.

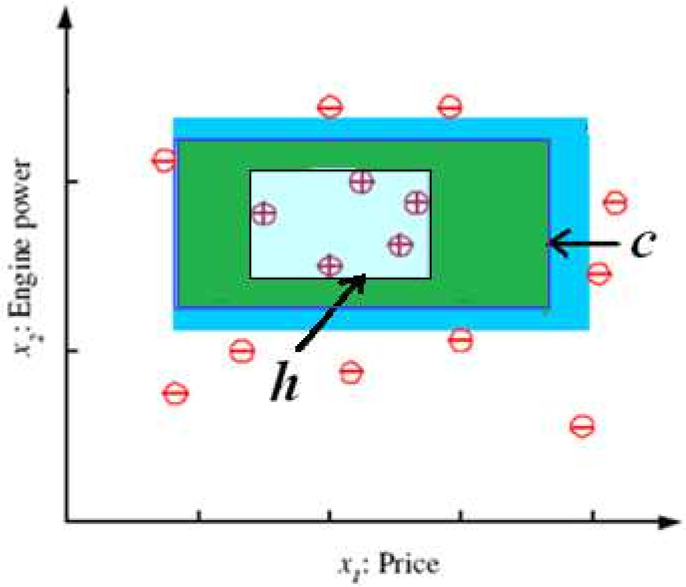


PAC learnable – How many training examples N should have s.t. the probability that the selected hypothesis h has error rate, error(h), at most ε, i.e., $\text{error}(h) \le \varepsilon$, is at least $1-\delta$? Mathematically,

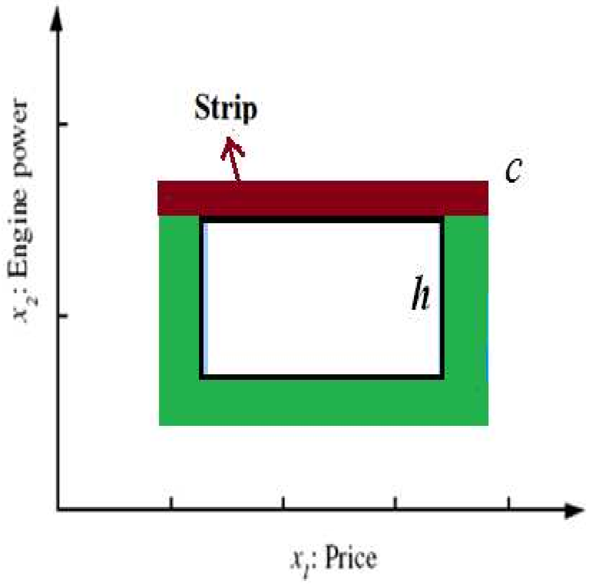
$P(\text{error}(h) \le \varepsilon) \ge 1-\delta$, where

$\text{error}(h) = c\Delta h$: the region of difference between c and h (assume $h = s$) .

ε: accuracy; δ: confidence



In order that the error rate (ER) of a positive car falling in $c\Delta h$ is at most $(\leq)\ \varepsilon$



- ER that fall in a strip is at most $(\leq)\ \varepsilon/4$
- Pr. that miss a strip at least $(>)\ 1-\varepsilon/4$
- Pr. that N instances miss a strip $(>)\ (1-\varepsilon/4)^N$
- Pr. that N instances miss 4 strips $(>)\ 4(1-\varepsilon/4)^N$

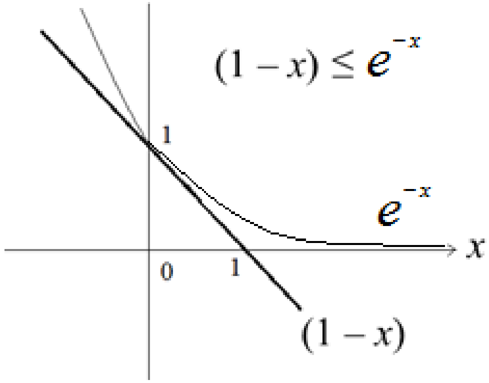
We would like this probability to be at most δ,

i.e. $4(1-\varepsilon/4)^N \leq \delta.$

$$4(1-\varepsilon/4)^N \le \delta \Rightarrow (1-\varepsilon/4)^N \le \delta/4$$

$$(\because (1-x) \le e^{-x}) \quad (1-\varepsilon/4) \le e^{-\varepsilon/4}$$

$$(1-\varepsilon/4)^N \le (e^{-\varepsilon/4})^N = e^{-\varepsilon N/4}$$



Choose N and δ s.t.

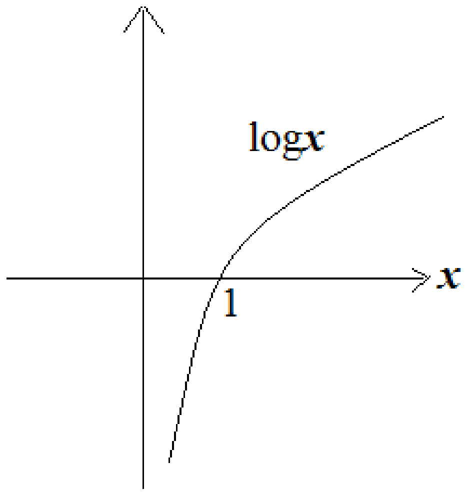
$$(1-\varepsilon/4)^N \le e^{-\varepsilon N/4} \le \delta/4$$

$$\Rightarrow -\varepsilon N/4 \le \ln \delta/4$$

$$\Rightarrow -N \le (4/\varepsilon)\ln \delta/4$$

$$\Rightarrow N \ge -(4/\varepsilon)\ln \delta/4$$

$$\Rightarrow \boxed{N \ge (4/\varepsilon)\ln(4/\delta)}$$



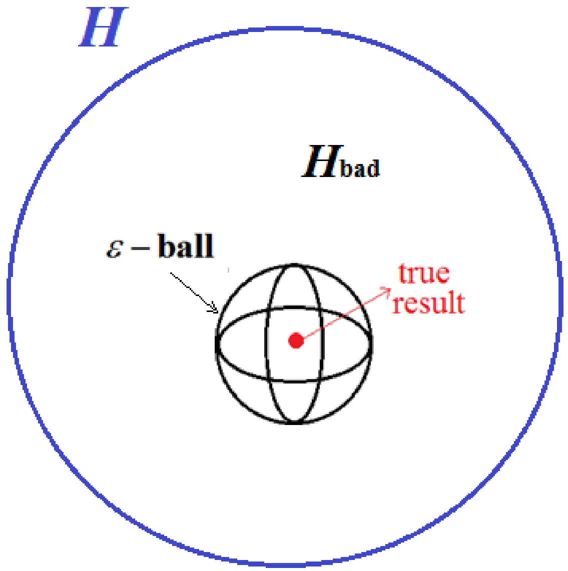
General bound: $\boxed{N \geq \frac{1}{\varepsilon}(\ln|H| + \ln(1/\delta))}$

The error rate of a hypothesis h, error(h), is the probability that h misclassifies a test example and is defined as the expected error of training examples, $\text{error}(h) = \sum_{x,y} L(y, h(x))P(x, y)$,

where $L(\ )$: 0/1 loss function

A hypothesis h is called approximately correct if $\text{error}(h) \leq \varepsilon$.

Let $\varepsilon-\mathbf{ball}$ be the small space around the true result. The hypothesis space outside it is H_{bad}.



Let $h_b \in H_{bad}$. Then, $\text{error}(h_b) > \varepsilon$.

$P(h_b$ agrees with an example$) \le 1-\varepsilon$.

$P(h_b$ agrees with N examples$) \le (1-\varepsilon)^N$.

$$P(H_{bad} \text{ contains a consistent hypothesis})$$
$$\leq |H_{bad}|(1-\varepsilon)^N \leq |H|(1-\varepsilon)^N$$
$$P(H_{bad} \text{ contains a consistent hypothesis})$$
$$\leq |H|(1-\varepsilon)^N \leq \delta \text{ for some small } \delta$$
$$\because (1-\varepsilon) \leq e^{-\varepsilon} \Rightarrow (1-\varepsilon)^N \leq e^{-N\varepsilon}$$
$$|H|(1-\varepsilon)^N \leq |H| e^{-N\varepsilon} \leq \delta$$
$$\Rightarrow \ln|H| - N\varepsilon \leq \ln\delta \Rightarrow -N\varepsilon \leq \ln\delta - \ln|H|$$
$$\Rightarrow N\varepsilon \geq \ln|H| - \ln\delta \Rightarrow N \geq \frac{1}{\varepsilon}(\ln|H| - \ln\delta)$$

How about for an infinite H ?

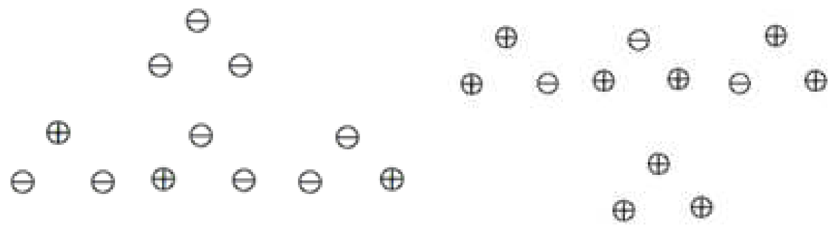
2.2.2 VC Dimension

-- A measure of hypothesis complexity.

Infinite $|H|$ may have limited VC(H).

- N points can be labeled in 2^N ways as +/–,

e.g., $N = 3,\ 2^N = 8$



$\forall$ labeling, if $\exists\, h \in H$ that separates + from – examples, H **shatters** N points.

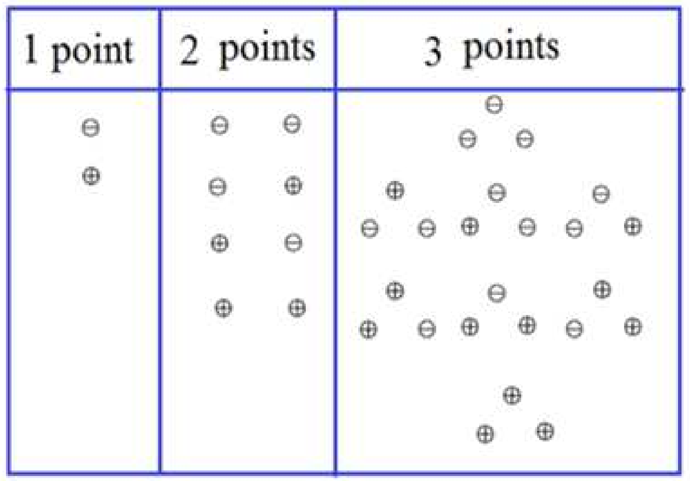
e.g., "Line" hypothesis class can shatter 3 points in the 2D space.

VC(*H*): the maximum number of points that can be shattered by *H*

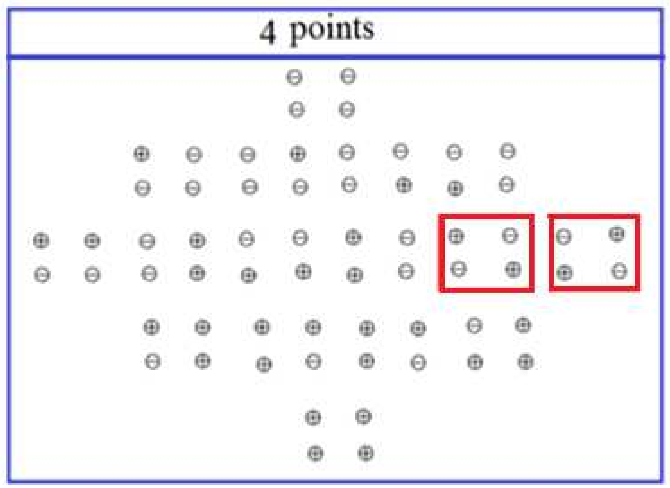
Examples:

i) The VC dimension of the “line” hypothesis class is 3 in 2D space, i.e., VC(line) = 3

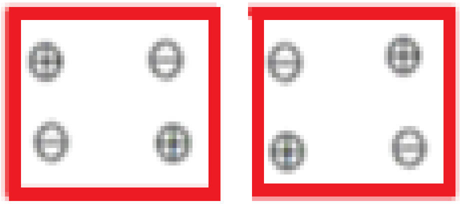
Can be shattered by “line” hypothesis class



Cannot be shattered by “line” hypothesis class

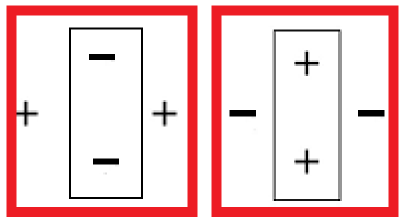


ii) The VC dimension of the "axis-aligned (AA) rectangle" hypothesis class is 4 in 2D space, i.e., VC(AA-rectangle) = 4

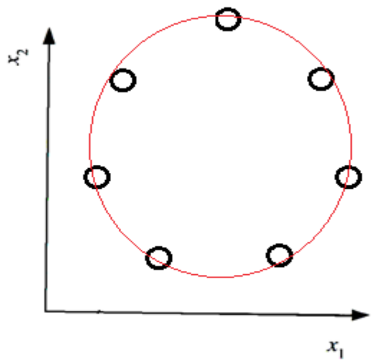


can not be shattered by "line" hypothesis class

They can be shattered by "AA-rectangle" class



iii) VC(triangle) = 7



2.3 Examples of Supervised Learning

2.3.1 Regression

Find $f(\cdot)$, s.t. $r = f(x),\ r \in R$

Training set: $X = \{x^t, r^t\}_{t=1}^{N}$

Let g be the estimate of f.

Expected total error: $E(g \mid X) = \dfrac{1}{N}\sum_{t=1}^{N}\left[r^t - g(x^t)\right]^2$

For linear model: $g(x \mid \boldsymbol{\theta}) = w_1 x + w_0,\ \boldsymbol{\theta} = (w_0, w_1)$

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

Problem: $\arg\min_{w_0, w_1} E(w_0, w_1 \mid X)$

$$\text{Let}\quad \frac{\partial E(w_0, w_1 \mid X)}{\partial w_1} = \frac{1}{N} \frac{\partial \sum_{t=1}^{N} \left[r^t - (w_1 x^t + w_0) \right]^2}{\partial w_1}$$

$$= \frac{-2}{N} \sum_{t=1}^{N} \left[r^t - (w_1 x^t + w_0) \right] x^t = 0$$

$$\sum_{t=1}^{N} \left[r^t - (w_1 x^t + w_0) \right] x^t = 0$$

where

$$\sum_{t=1}^{N} r^t x^t - w_1 \sum_{t=1}^{N} (x^t)^2 + w_0 \sum_{t=1}^{N} x^t = 0 \qquad \overline{x} = \sum_t x^t / N$$

$$\sum_{t=1}^{N} r^t x^t - w_1 \sum_{t=1}^{N} (x^t)^2 + w_0 N \overline{x} = 0 \qquad \text{------ (1)}$$

$$\frac{\partial E(w_0, w_1 \mid X)}{\partial w_0} = \frac{1}{N} \frac{\partial \sum_{t=1}^{N} \left[r^t - (w_1 x^t + w_0) \right]^2}{\partial w_0} = 0$$

$$\sum_{t=1}^{N}\left[r^t-(w_1x^t+w_0)\right]=0$$

$$\sum_{t=1}^{N}r^t-w_1\sum_{t=1}^{N}x^t-Nw_0=0 \qquad \text{where } \bar{r}=\sum_t r^t/N$$

$$N\bar{r}-w_1N\bar{x}-Nw_0=0 \qquad \text{-----}\ (2)$$

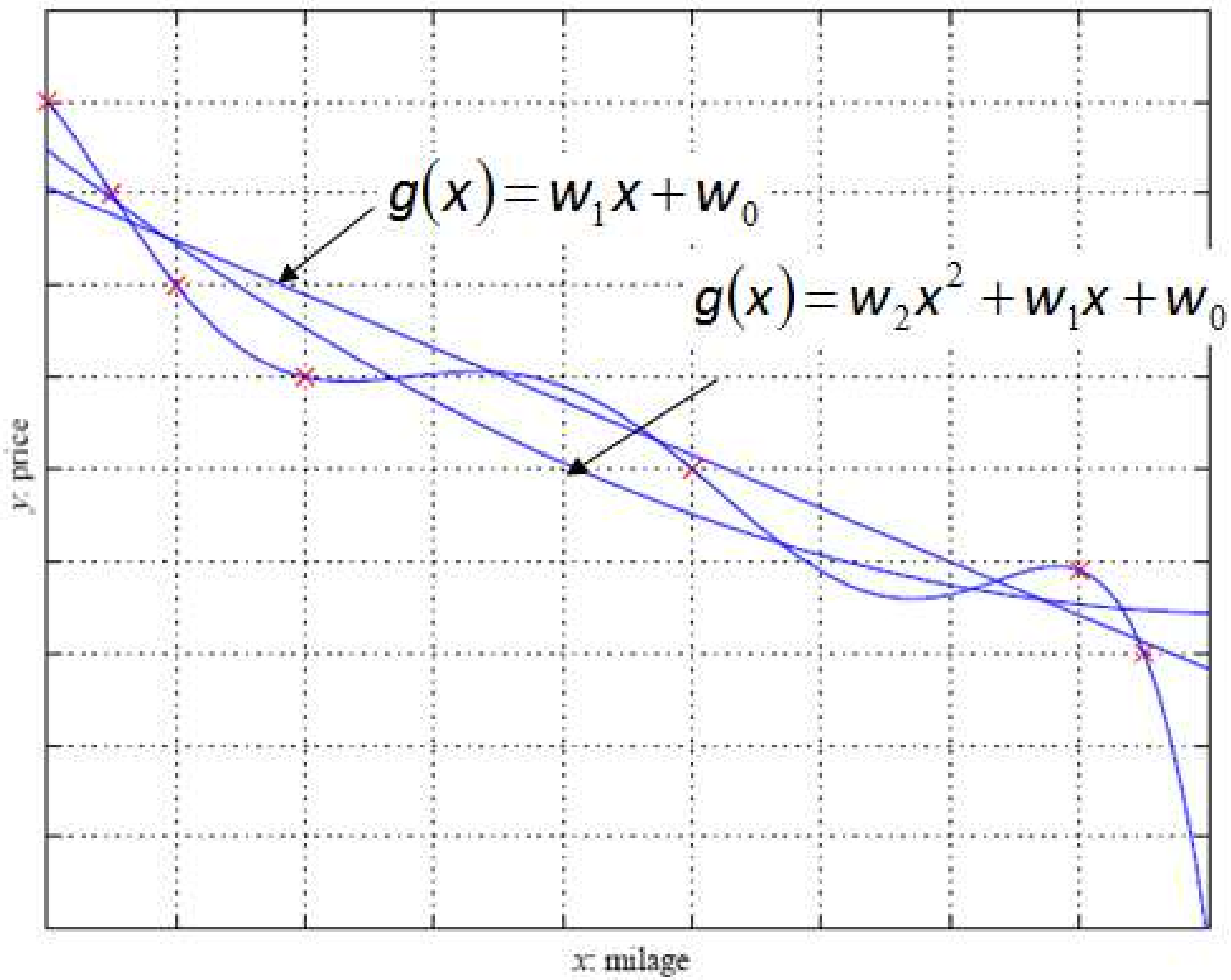
Solve (1) and (2) for w_1 and w_0

$$w_1=\frac{\sum_t x^t r^t-\overline{xr}N}{\sum_t (x^t)^2-N\bar{x}^2},\quad w_0=\bar{r}-w_1\bar{x}$$

For quadratic model:

$$g(x\,|\,\boldsymbol{\theta})=w_2x^2+w_1x+w_0,\ \boldsymbol{\theta}=(w_0,w_1,w_2)$$

Solve for w_2, w_1 and w_0


$g(x) = w_1 x + w_0$
$g(x) = w_2 x^2 + w_1 x + w_0$
y: price
x: milage

2.3.2 Classification

Multiple Classes, C_i $i = 1, ..., K$

Training set: $X = \{\boldsymbol{x}^t, \boldsymbol{r}^t\}_{t=1}^N$, $r_i^t = \begin{cases} 1 & \boldsymbol{x}^t \in C_i \\ 0 & \boldsymbol{x}^t \in C_{j \neq i} \end{cases}$

A. Treat a K-class classification problem as K

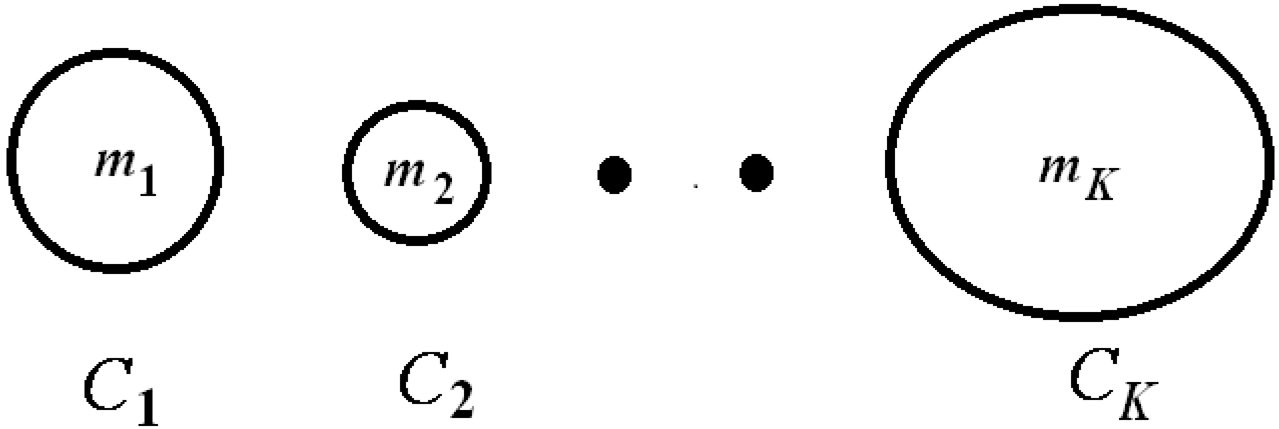
2-class problems, i.e., train hypotheses

$h_i(\boldsymbol{x}^t) = \begin{cases} 1 & \boldsymbol{x}^t \in C_i \\ 0 & \boldsymbol{x}^t \in C_{j \neq i} \end{cases}$, $i = 1, \cdots, K$ that minimize

the total error $E = \sum_{t=1}^{N} \sum_{i=1}^{K} 1(h_i(\boldsymbol{x}^t) \neq r_i^t)$.

Problem: $\arg \min_{h_i,\ i=1,\cdots,K,} E(h_1, \cdots, h_K)$

B. Training set: $X = \{\boldsymbol{x}^t, \boldsymbol{r}^t\}_{t=1}^N,\ \ r_i^t = \begin{cases} 1 & \boldsymbol{x}^t \in C_i \\ 0 & \boldsymbol{x}^t \in C_{j \neq i} \end{cases}$



$$\boldsymbol{m}_i = \sum_t r_i^t \boldsymbol{x}^t \Big/ \sum_t r_i^t$$

Classification rule:

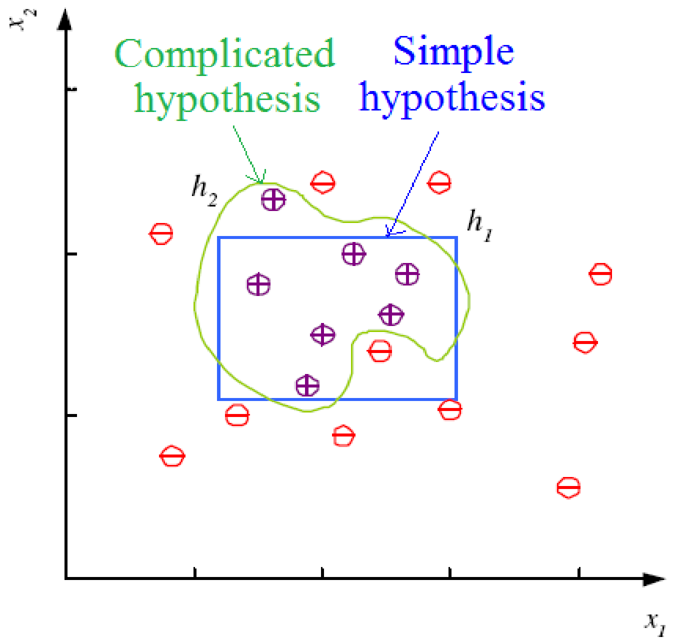
Assign $\boldsymbol{x}$ to C_k if $k = \arg\min_{1 \le i \le K} \left\| \boldsymbol{x} - \boldsymbol{m}_i \right\|^2$

2.4 Noise

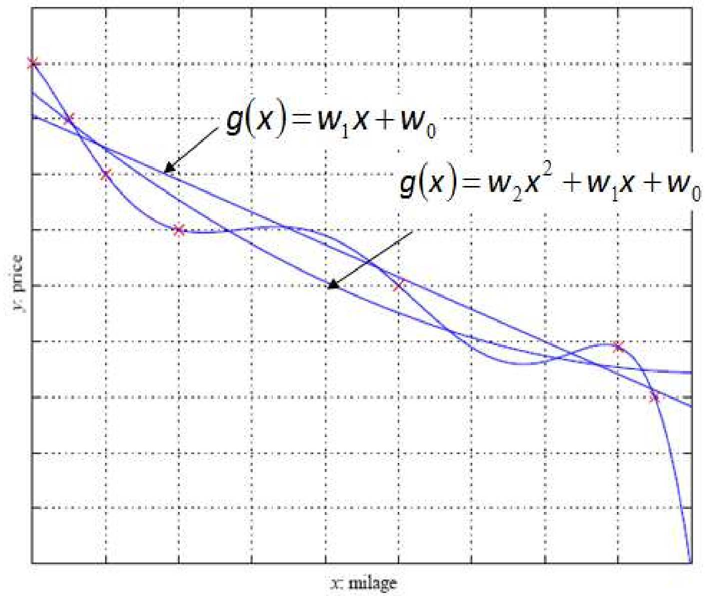
Noise due to error, imprecision, uncertainty, etc..

complicated hypotheses are generally necessary to cope with noise.

1) Classification



2) Regression



2.5 Dimensions of a Supervised Learning Algorithm

Given a sample: $X = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$

1. Model $g(\boldsymbol{x} \mid \theta)$: $g(\cdot)$ defines the hypothesis set H

2. Error function $E(\cdot)$: $E(\theta \mid X) = \sum_t L(r^t, g(\boldsymbol{x}^t \mid \theta))$

   where $L(\cdot)$: loss functin expresses the difference between r^t and $g(\boldsymbol{x}^t \mid \theta)$

3. Optimization procedure: $\theta^* = \arg\min_\theta E(\theta \mid X)$

2.6 Model Selection

Training set: to train candidate models

Validation set: to evaluate the candidate models and choose the best one

Test set: to provide the generalization error of the chosen model

K-fold cross-validation: Data are divided into *k* equal subsets; perform *k* rounds of learning; on each round, one subset serves the validation set and the remaining data serve the test set.

Ill-posed problem: training examples are not sufficient to lead to a unique solution

Inductive bias: additional information, prior knowledge, assumption, etc. for making learning possible

Model selection: chooses a good hypothesis set

Underfitting: hypothesis set H is less complex than the function underlying the data, e.g., fit a line to data sampled from a 3rd order polynomial

Overfitting: hypothesis set H is more complex than the function, e.g., fit a 3rd order polynomial to data sampled from a line

Triple trade-off:

1. The size of training set, N
2. The complexity of H, $C(H)$
3. The error on new data, E

As $N \Uparrow$, $E \Downarrow$; As $C(H) \Uparrow$, E first $\Downarrow$, then $\Uparrow$