CHAPTER 1:

SUPERVİSED LEARNİNG

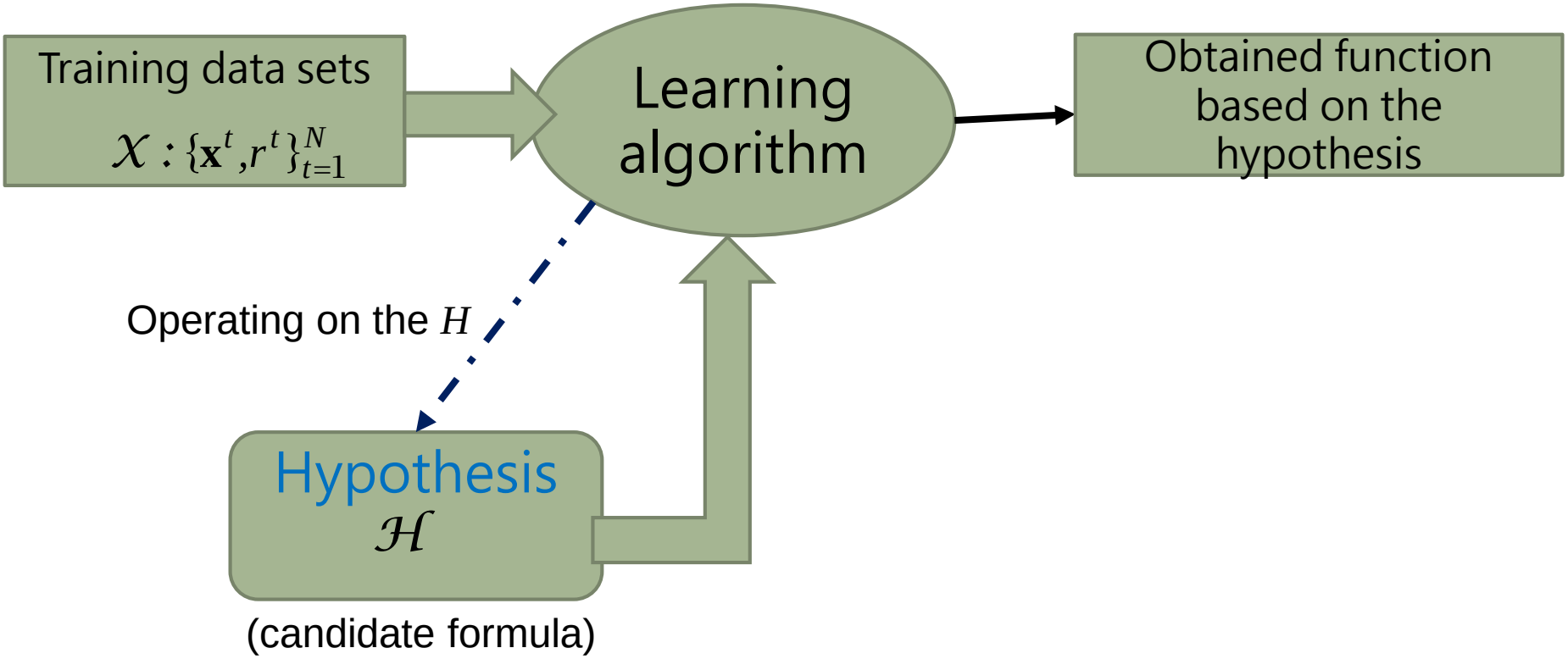
國立雲林科技大學 資訊工程研究所
張傳育(Chuan-Yu Chang) 博士
Office: EB 212
TEL: 05-5342601 ext. 4516
E-mail: chuanyu @yuntech.edu.tw
Website: http://MIPL.yuntech.edu.tw

Supervised vs Unsupervised learning

- Supervised learning:
  - Target of training for each training datum is provided
  - Usually the target is used to build cost function
  - Equations are derived to minimize the cost function
  - E.g. Classification,

- Unsupervised learning
  - No target is provided
  - A specific rule of updating is used but no target is provided
  - Equations are derived from the patterns for optimization
  - E.g. clustering

How machine learns?



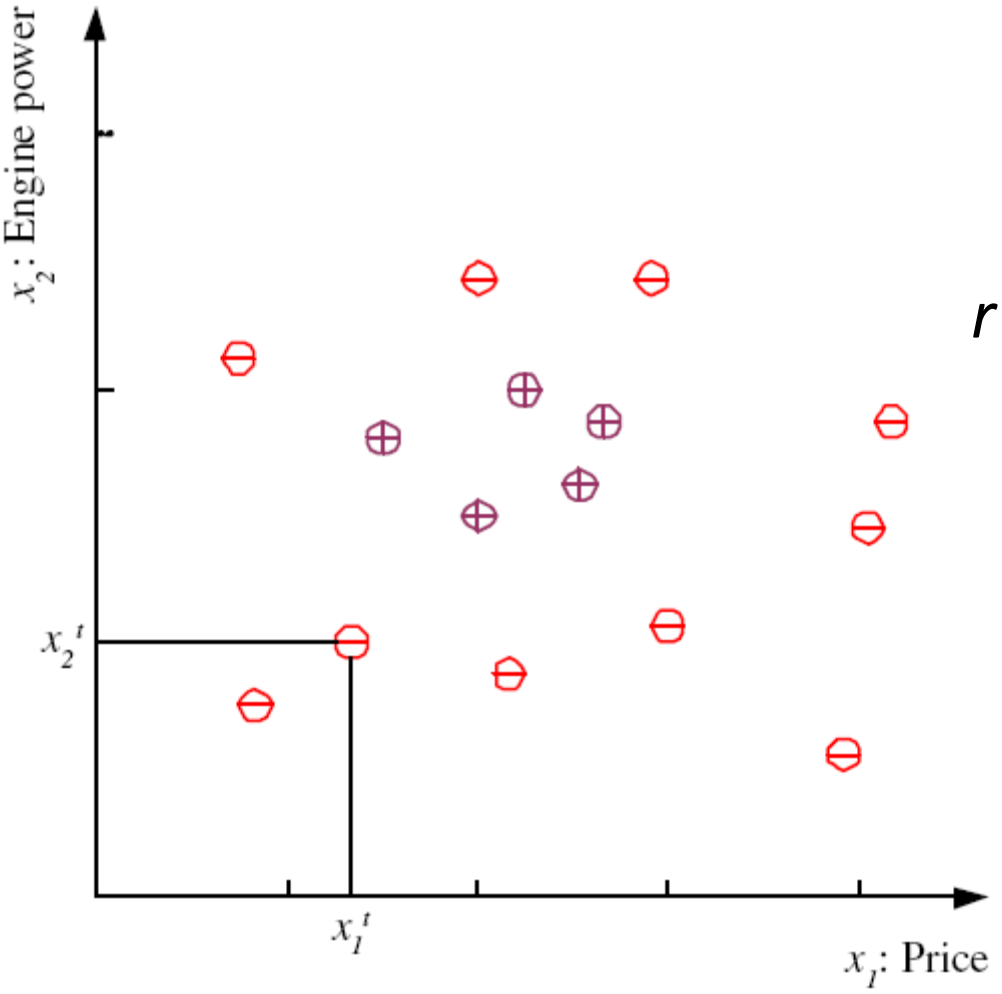
Learning a Class from Examples

- Class C of a “family car”
  - Prediction: Is car $\mathbf{x}$ 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

Training set $\mathcal{X}$



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

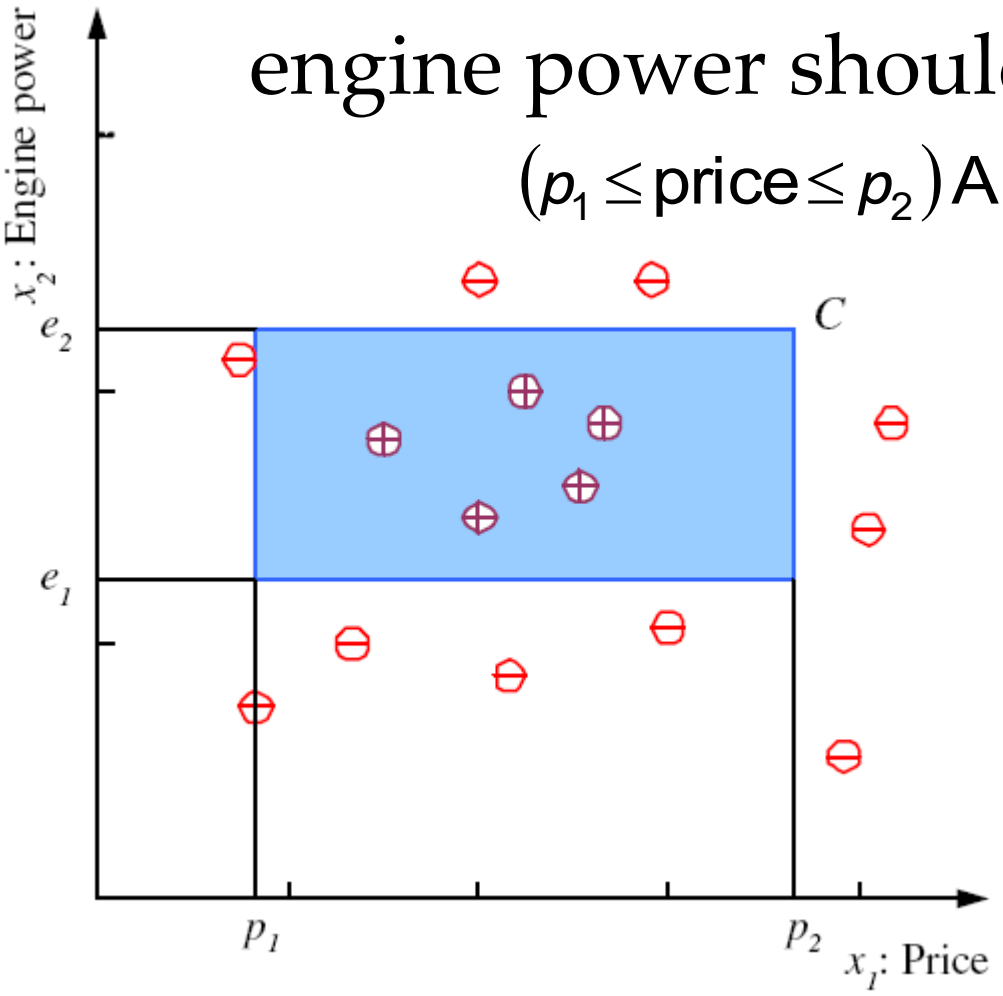
$$r = \begin{cases} 1 \text{ if } \mathbf{x} \text{ is positive} \\ 0 \text{ if } \mathbf{x} \text{ is negative} \end{cases}$$

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

Class C

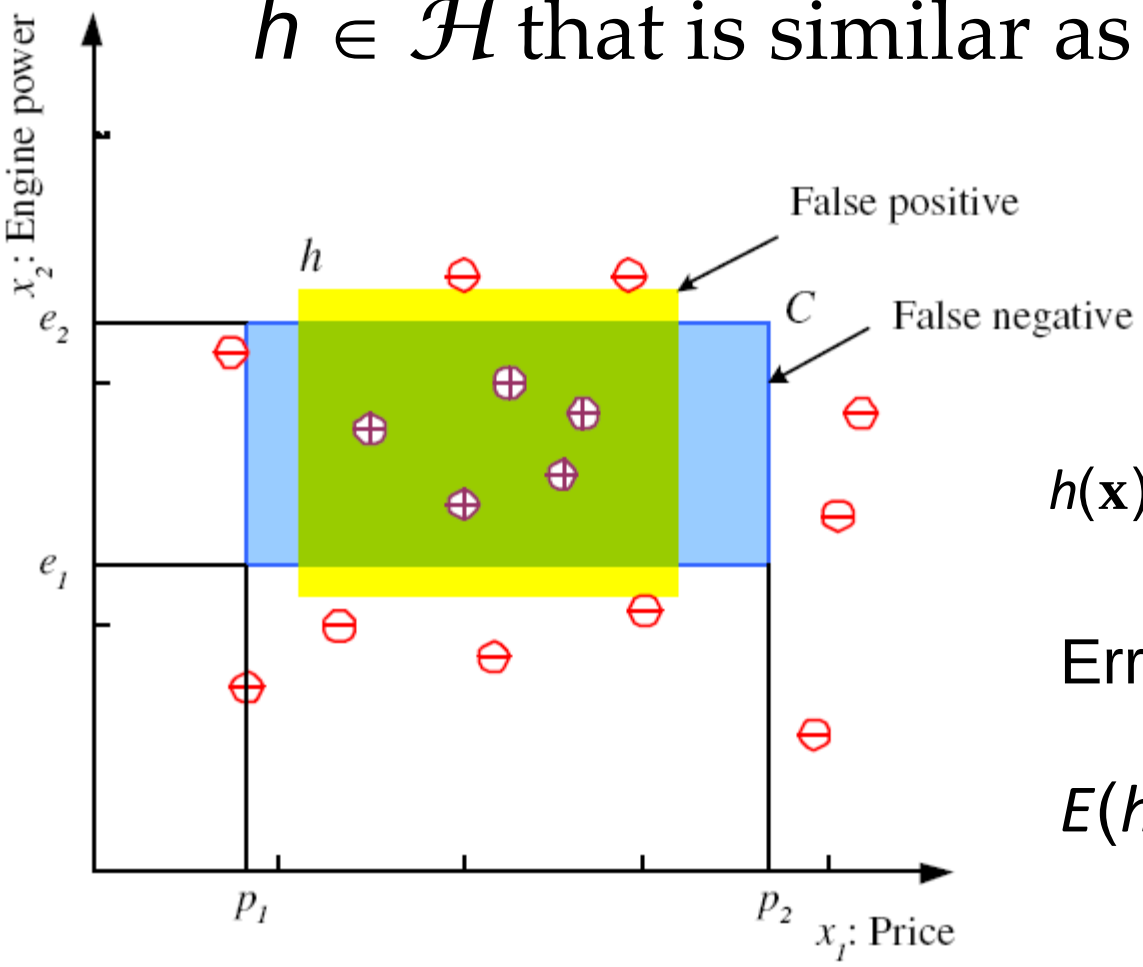
For a car to be a family car, its price and engine power should be in a certain range

$$(p_1 \leq \text{price} \leq p_2) \text{ AND } (e_1 \leq \text{engine power} \leq e_2)$$



Hypothesis class $\mathcal{H}$

The aim of learning algorithm is to find $h \in \mathcal{H}$ that is similar as possible to C.



$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ says } \mathbf{x} \text{ is positive} \\ 0 \text{ if } h \text{ says } \mathbf{x} \text{ is negative} \end{cases}$$

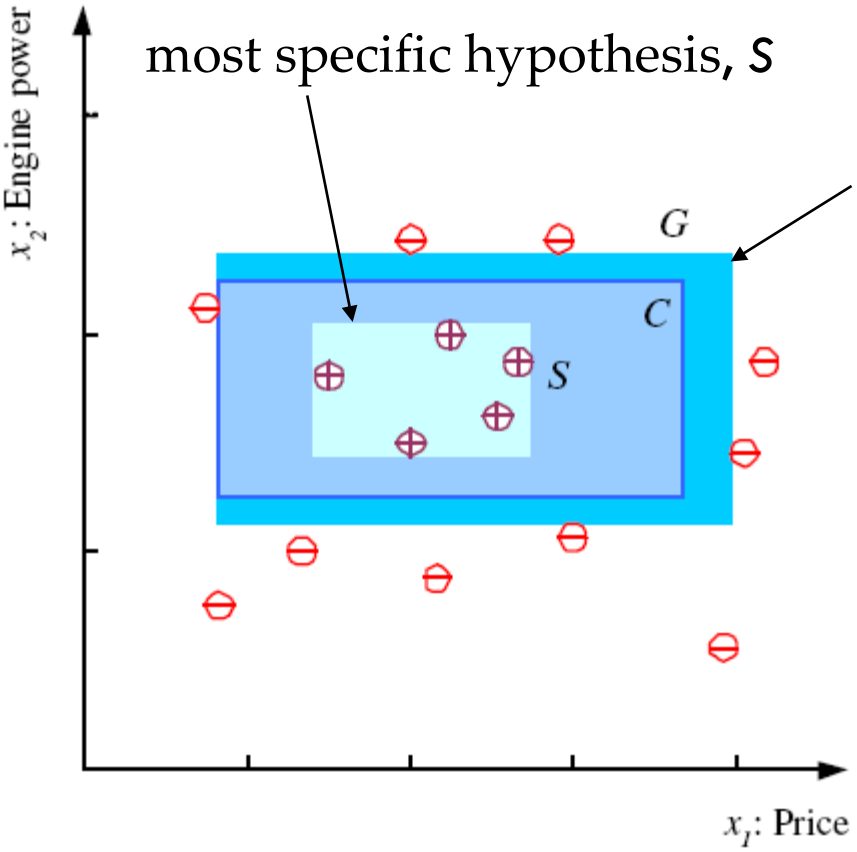
Error of h on $\mathcal{H}$

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

Generalization

- Generalization
  - How well the hypothesis will correctly classify future examples that are not part of the training set.
- To find the most specific hypothesis, *S*, that is the tightest rectangle that includes all the positive examples and none of the negative examples.
- The most general hypothesis, *G*, is the largest rectangle we can draw that includes all the positive examples and none of the negative examples.

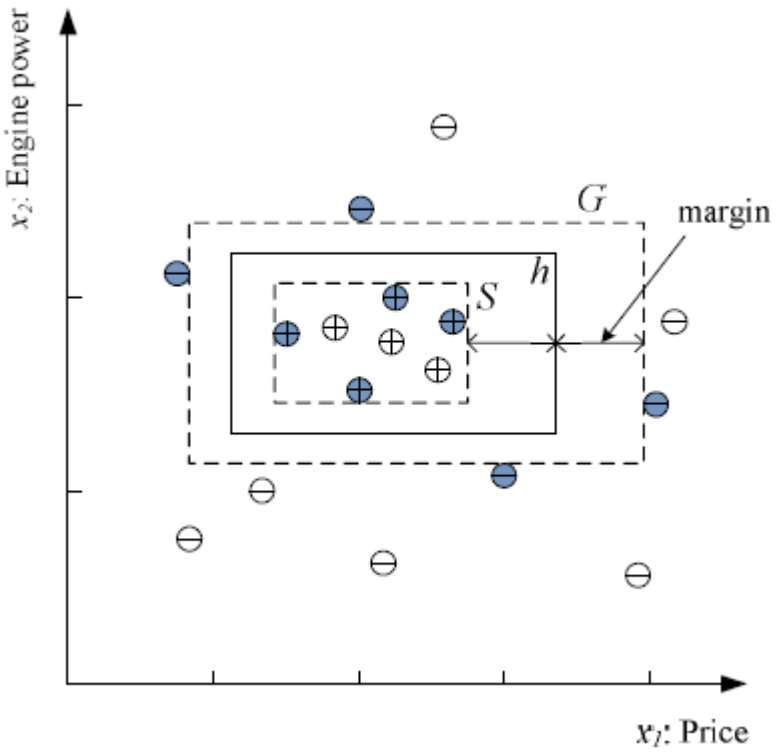
S, G, and the Version Space



Any $h \in \mathcal{H}$, between S and G is a valid hypothesis with no error, said to be consistent with the training set, and such h make up the version space (Mitchell, 1997)

Margin

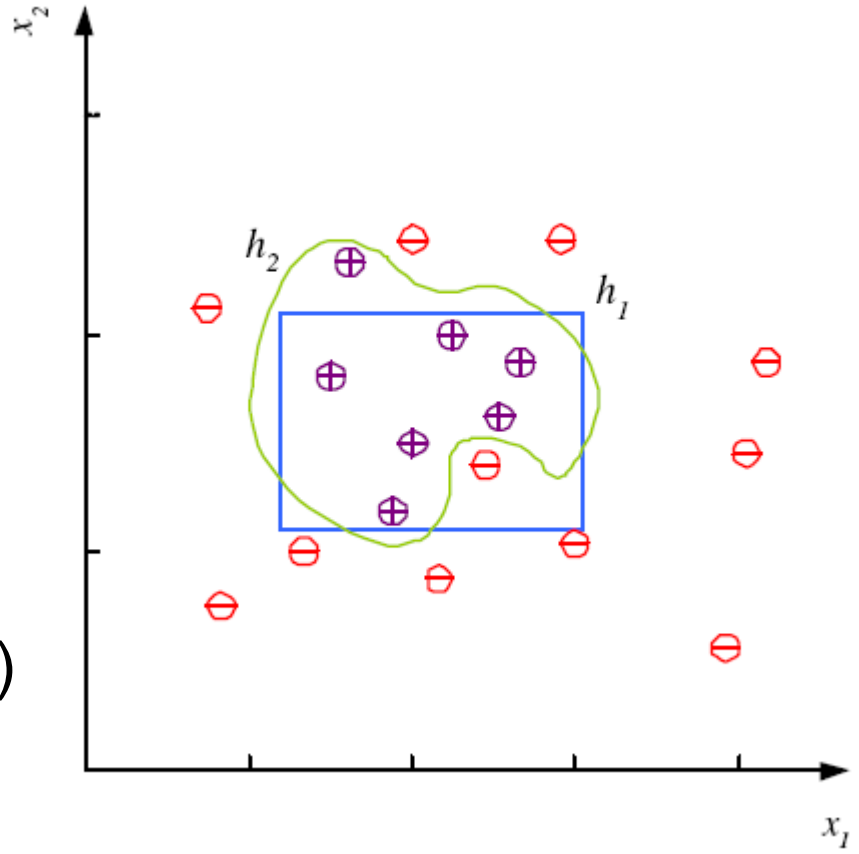
- Choose h with largest margin and minimum error function.
  - Any instance that falls in between S and G is a case of doubt, which we cannot label with certainty due to lack of data.
  - The system rejects the instance and defers the decision to a human expert.



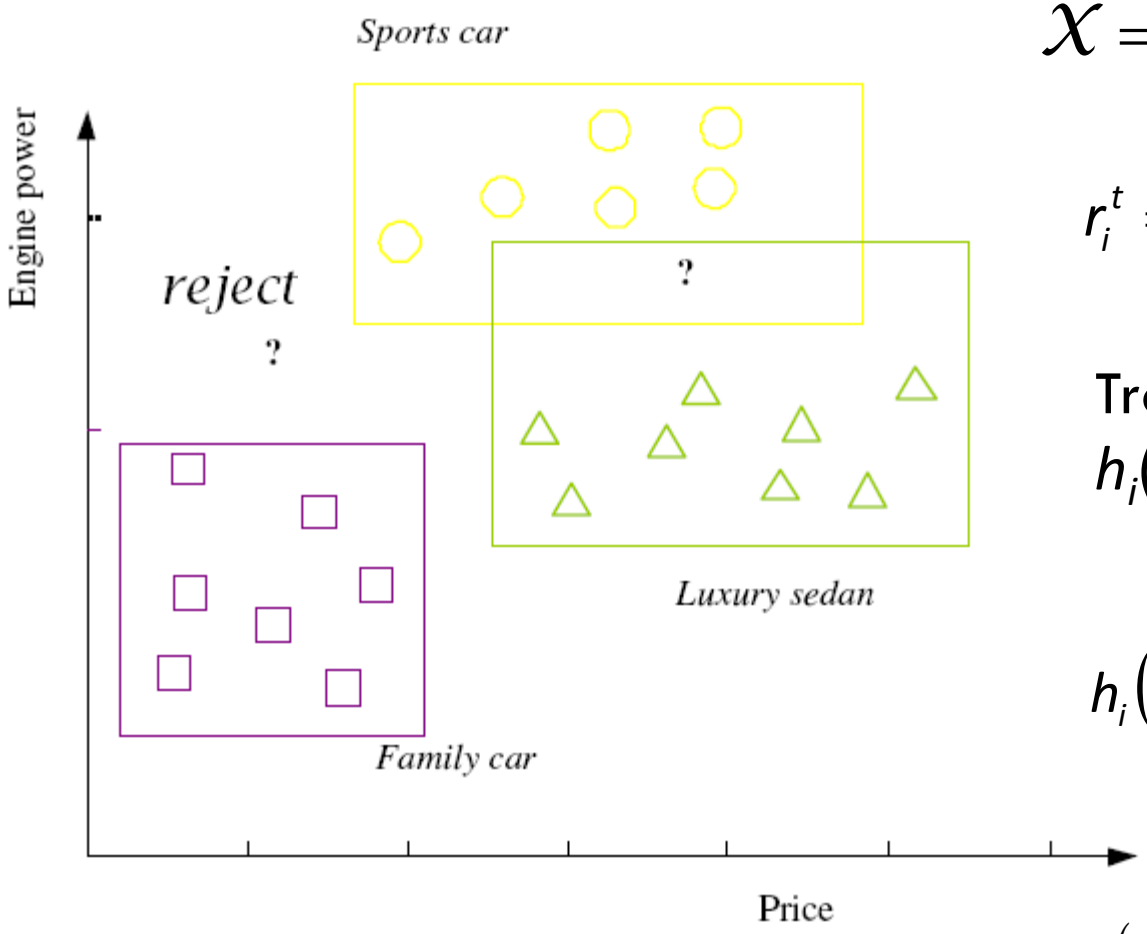
Noise and Model Complexity

Use the simpler hypothesis 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



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

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

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

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

$$E\left(\{h_i\}_{i=1}^k \mid \mathcal{X}\right) = \sum_{t=1}^{N} \sum_{i=1}^{K} 1\left(h_i\left(x^t\right) \neq r_i^t\right)$$

Regression

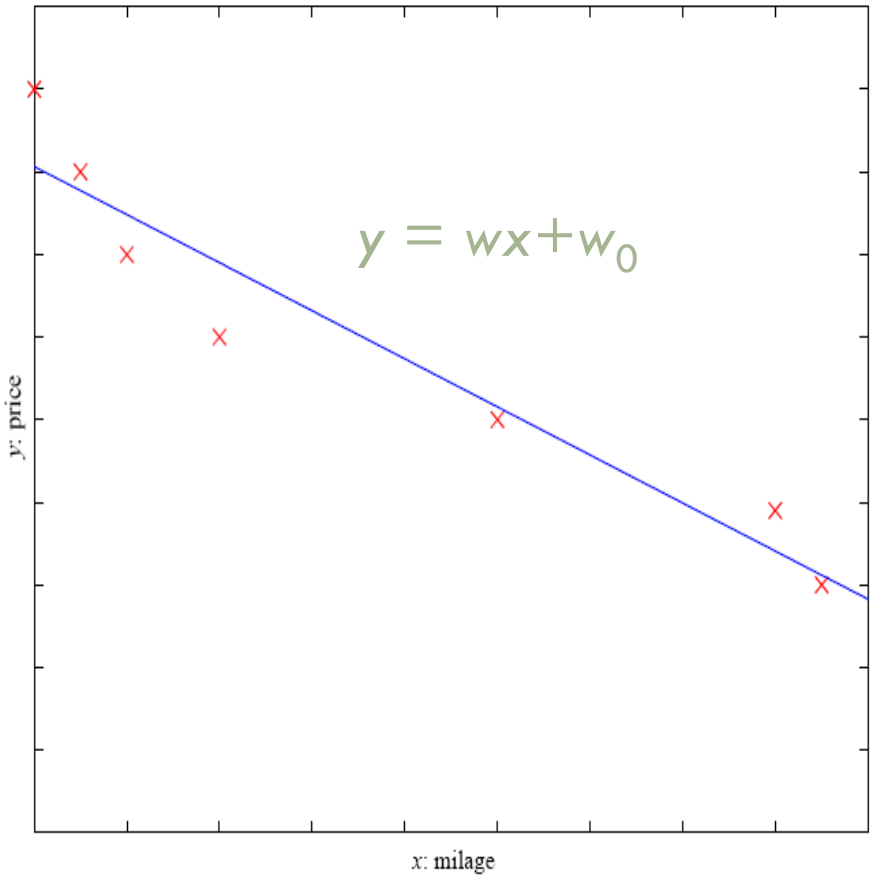
- Example: Price of a used car
- x : car attributes

  y : price

  $y = g(x \mid \theta)$

  g (.) : model,

  θ : parameters



Regression

- Assume that we have a training set of examples

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

where $r^t \in \Re$

- We would like to find the function $f(x)$ that passes through these points such that we have

$$r^t = f\left(x^t\right)$$

- In regression, there is noise added to the output of the unknown function

$$r^t = f\left(x^t\right) + \varepsilon$$

Regression

- The empirical error on the training set $\mathcal{X}$ is

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

- Our aim is to find g(.) that minimizes empirical error.
- If we assume that $g(x)$ is linear, we have

$$g(x) = w_1 x_1 + \cdots + w_d x_d + w_0 = \sum_{j=1}^{d} w_j x_j + w_0$$

Regression

- For a single input linear model

$$g(x) = w_1 x + w_0$$

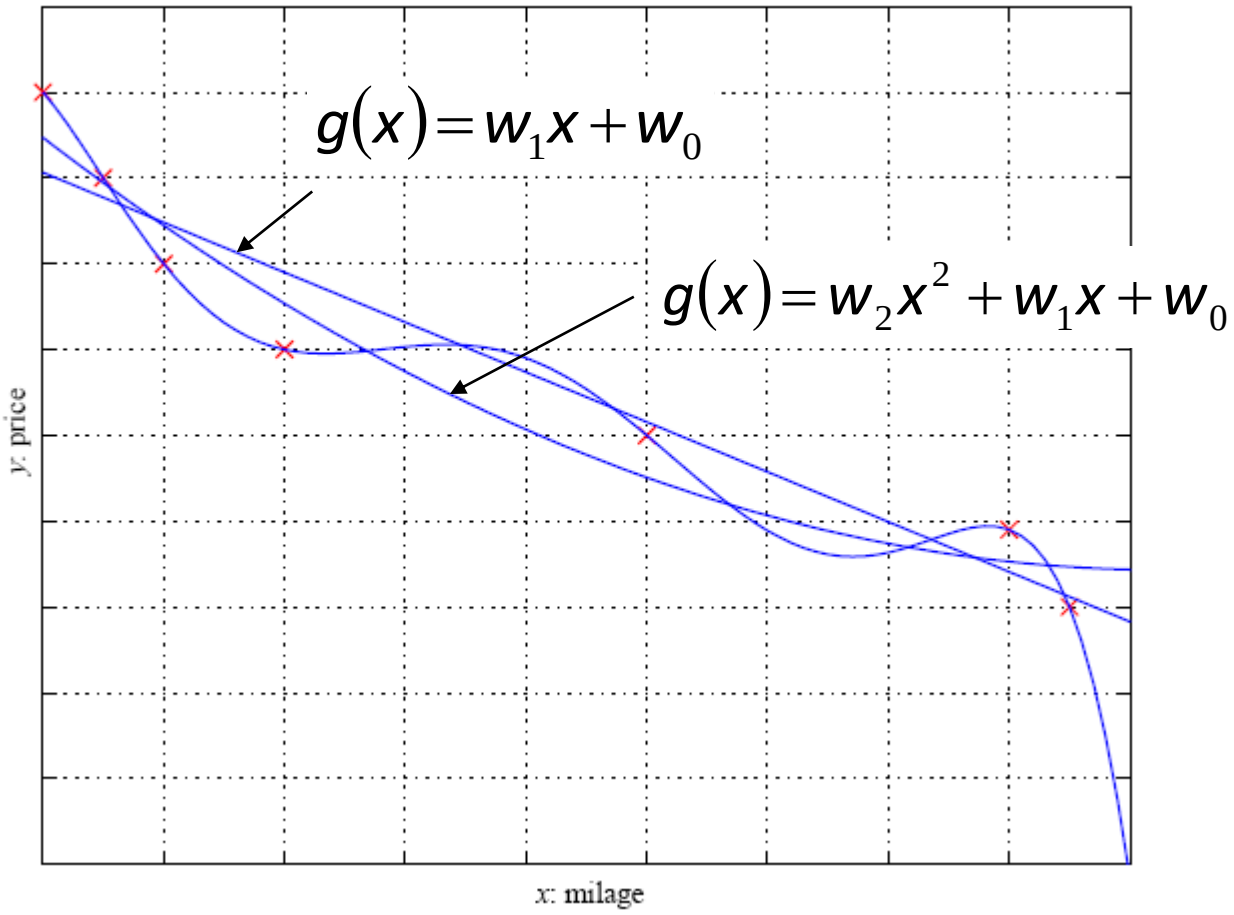
- The w_1 and w_0 values should minimize

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

- Its minimum point can be calculated by taking partial derivatives of E with respect to w_1 and w_0 , setting them equal to 0.

$$w_1 = \frac{\sum_t x^t r^t - \bar{x}\bar{r}N}{\sum_t \left(x^t\right)^2 - N\bar{x}^2} \qquad w_0 = \bar{r} - w_1\bar{x}$$

Regression

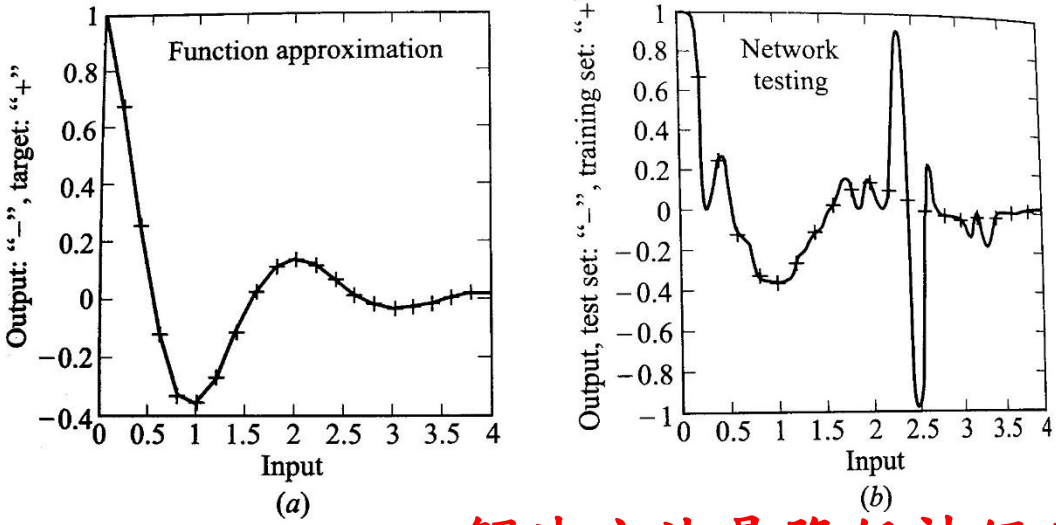


Model Selection & Generalization

- Learning is an ill-posed problem; data is not sufficient to find a unique solution
- The need for inductive bias, assumptions about $\mathcal{H}$.
  - The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered.
- Generalization: How well a model performs on new data
- Overfitting: $\mathcal{H}$ more complex than C or f
- Underfitting: $\mathcal{H}$ less complex than C or f

Model Selection & Generalization

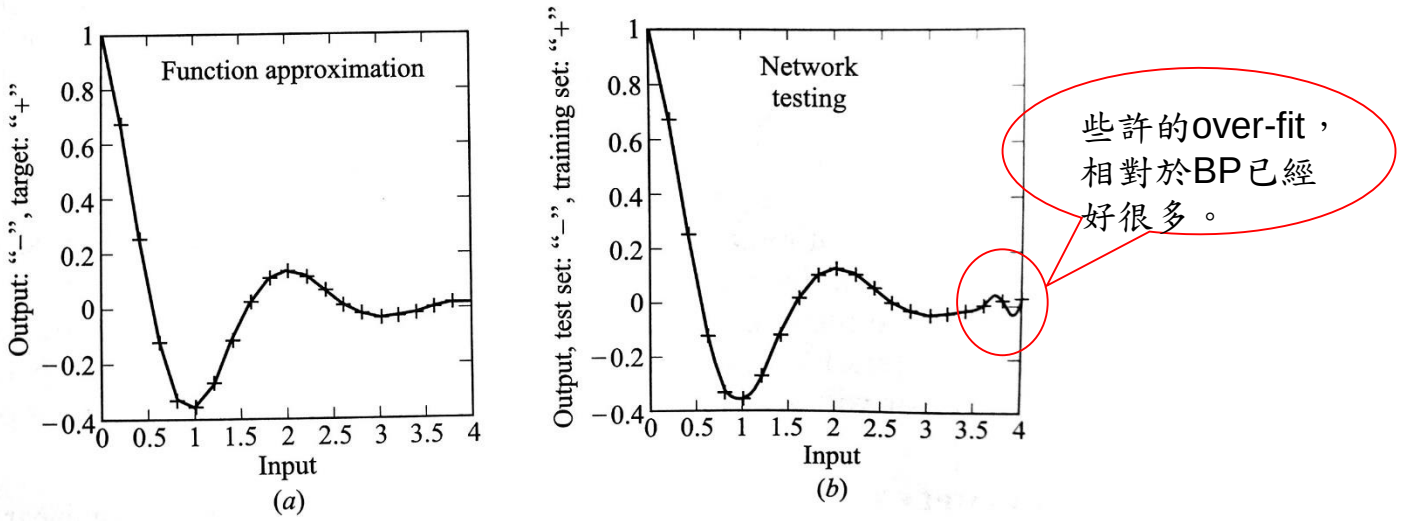
- Example: Backpropagation Learning Algorithm
  - 此MLP網路具有一個hidden layer，50個神經元。
  - 要訓練一個非線性方程式 $y = e^{-x}\sin(3x)$
  - (a)Training: 在[0,4]之間，每0.2取樣一點，共21點，target mean square error=0.01。The network converaged in only 5 epochs.
  - (b)Testing: 在[0,4]之間，每0.01取樣一點，共401點。
    - Training data造成overfitting



解決方法是降低神經元的個數

Model Selection & Generalization

- Example: Radial Basis Function Neural Networks
  - 要訓練一個RBFNN來近似非線性方程式 $y = e^{-x}\sin(3x)$
  - Interval [0,4] ，hidden layer: 21 neurons。
  - Fig. (a) :training：取樣間距0.2，所以隱藏層有21個神經元。(中心點即為此21個取樣值，Gaussian RBF, σ=0.2)
  - Fig. (b) :testing：取樣間距0.01，有401個取樣點。



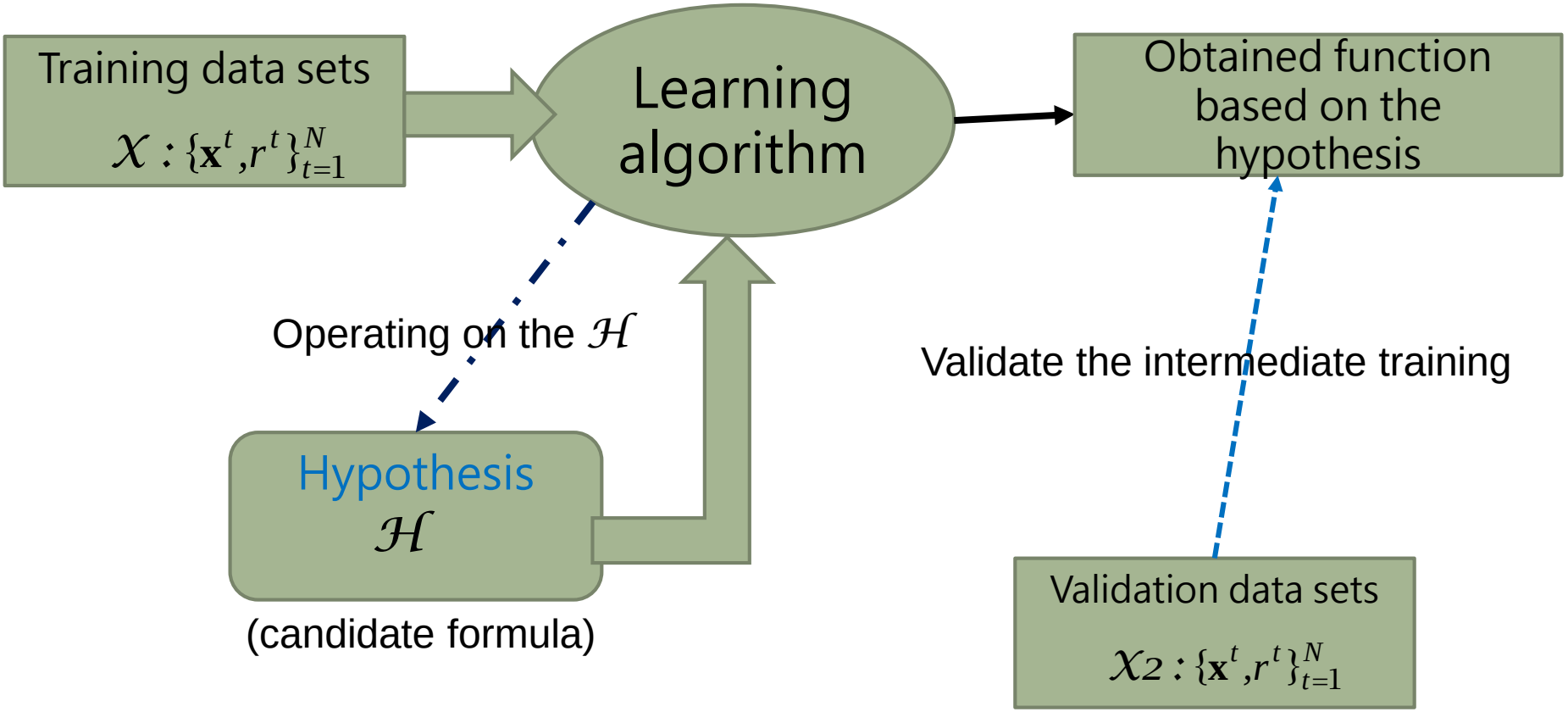
Triple Trade-Off

- There is a trade-off between three factors (Dietterich, 2003):
  1. Complexity of $\mathcal{H}$, c ($\mathcal{H}$),
  2. Training set size, N,
  3. Generalization error, E, on new data
- As $N \uparrow$, $E \downarrow$
- As c ($\mathcal{H}$) $\uparrow$, first $E \downarrow$ and then $E \uparrow$

Cross-Validation

- To estimate generalization error, we need data unseen during training. We split the data as
  - Training set (50%)
  - Validation set (25%)
  - Test (publication) set (25%)
- Resampling when there is few data

Cross validation



Dimensions of a Supervised Learner

- Assume we have a sample $\mathcal{X} = \left\{x^t, r^t\right\}_{t=1}^{N}$
- The aim is to build a good and useful approximation to r^t using the model.

1. Model: $g(\mathbf{x}|\theta)$

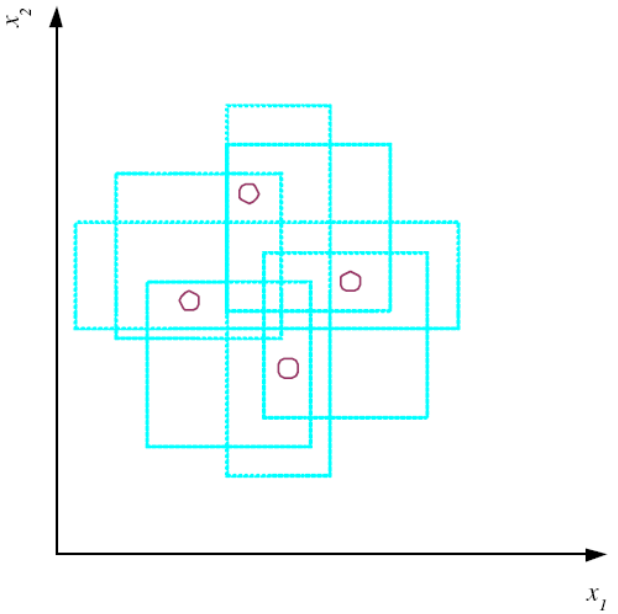
2. Loss function: $E(\theta|\mathcal{X}) = \sum_t L\left(r^t, g\left(\mathbf{x}^t|\theta\right)\right)$

3. Optimization procedure:

$$\theta^* = \arg\min_{\theta} E(\theta|\mathcal{X})$$

VC Dimension

- N points can be labeled in 2^N ways as +/–
- $\mathcal{H}$ shatters N if there exists $h \in \mathcal{H}$ consistent for any of these:
  $VC(\mathcal{H}) = N$



An axis-aligned rectangle shatters 4 points only !

Probably Approximately Correct (PAC) Learning

- How many training examples N should we have, such that with probability at least $1 - \delta$, h has error at most ε ?

  (Blumer et al., 1989)

- Each strip is at most $\varepsilon/4$
- Pr that we miss a strip $1 - \varepsilon/4$
- Pr that N instances miss a strip $(1 - \varepsilon/4)^N$
- Pr that N instances miss 4 strips $4(1 - \varepsilon/4)^N$
- $4(1 - \varepsilon/4)^N \leq \delta$ and $(1 - x) \leq \exp(-x)$
- $4\exp(-\varepsilon N/4) \leq \delta$ and $N \geq (4/\varepsilon)\log(4/\delta)$