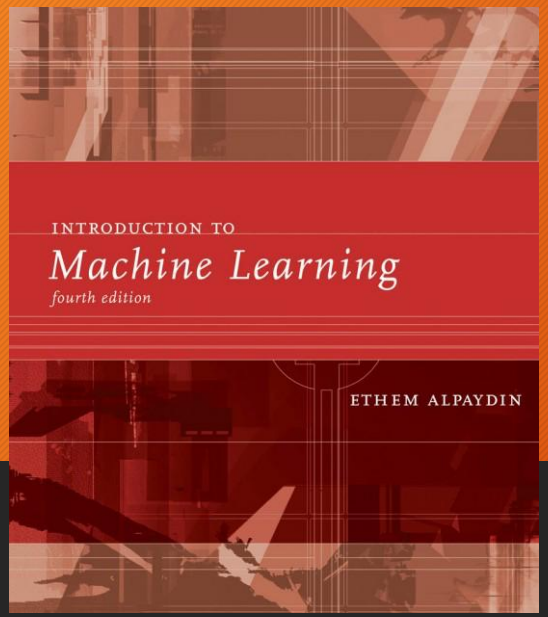


Lecture Slides for

INTRODUCTION TO Machine Learning 4e

ETHEM ALPAYDIN
The MIT Press, 2020

ethem.alpaydin@gmail.com

Troy University Resources

CHAPTER 2:

Supervised Learning

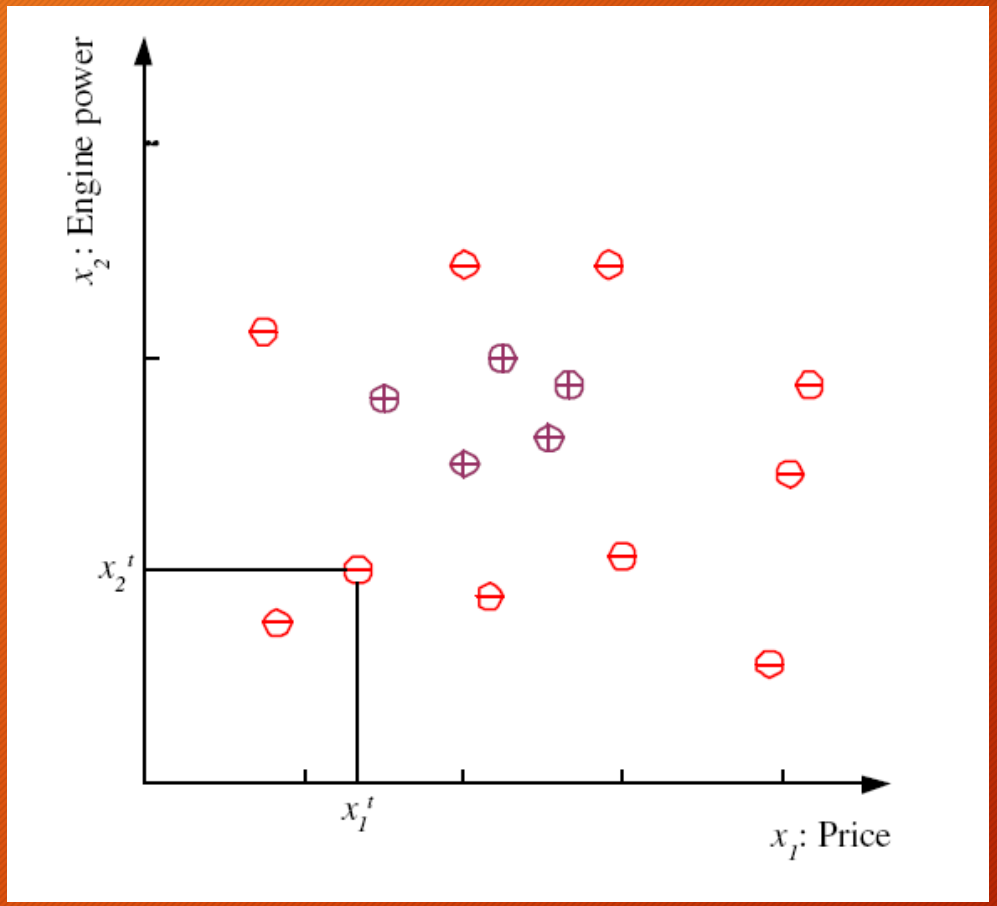
Learning a Class from Examples

- 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

Training set $\mathcal{X}$

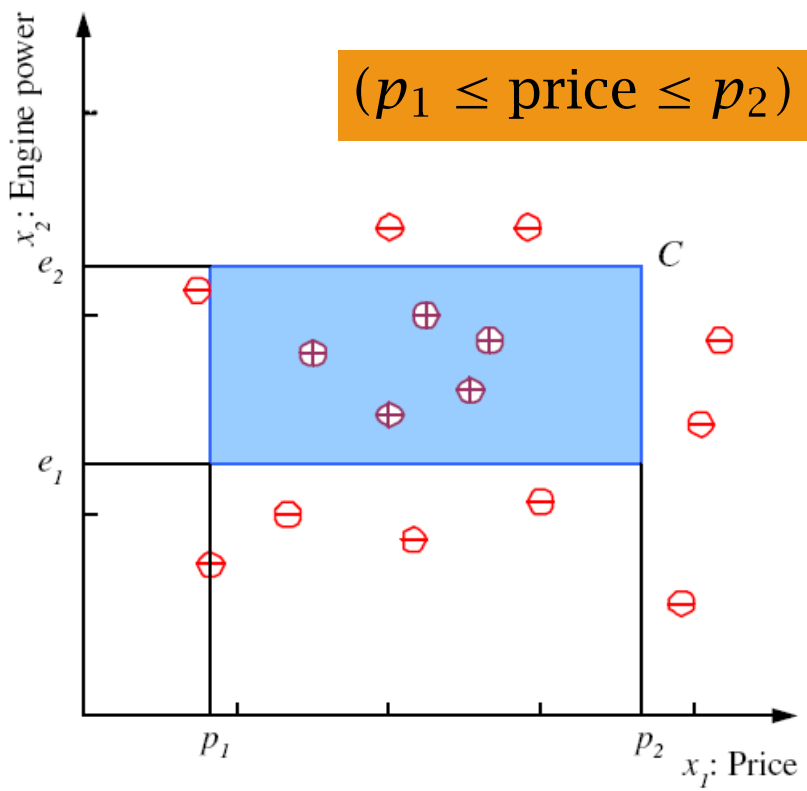


$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^{N}$$

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

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

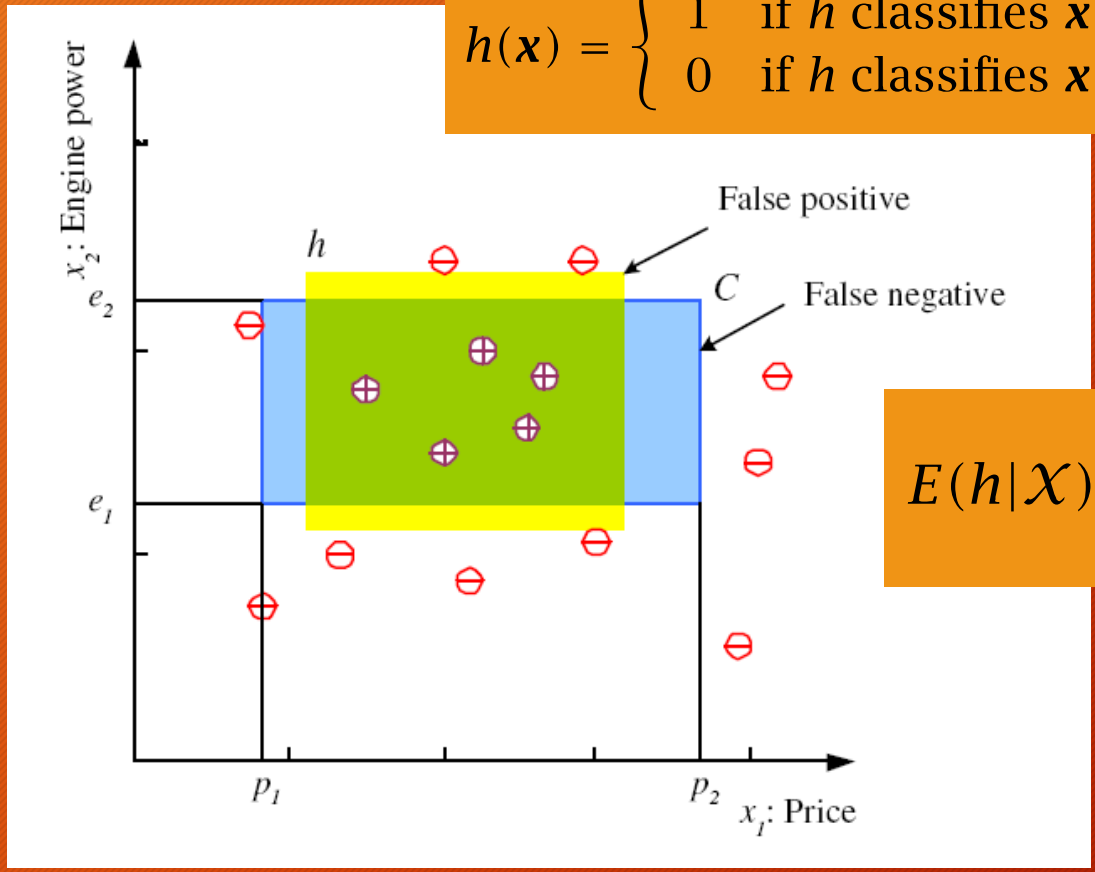
Class C



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

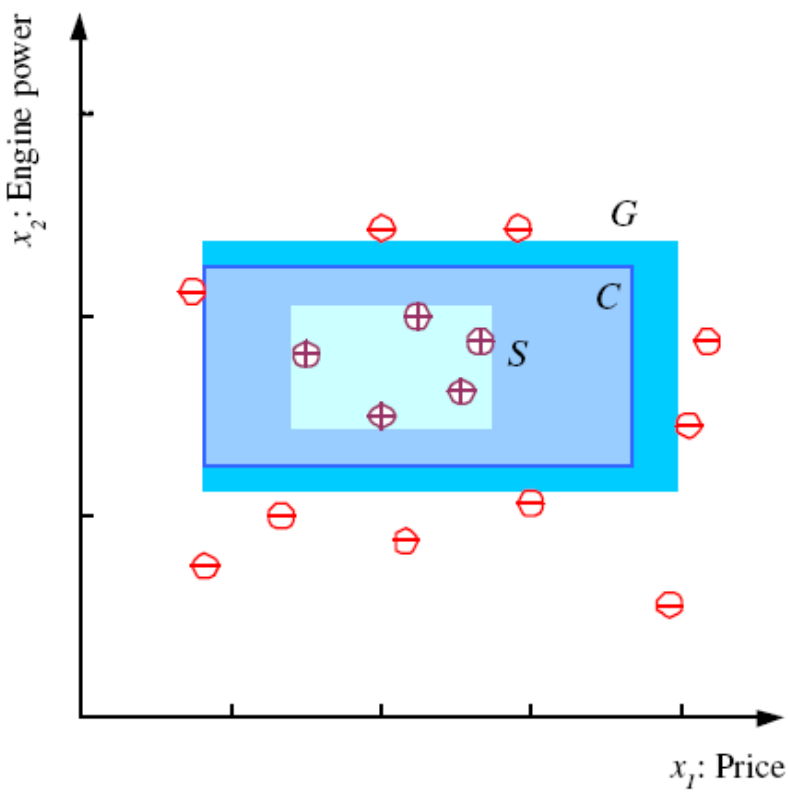
Hypothesis class $\mathcal{H}$

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



$$E(h|\mathcal{X}) = \sum_{t=1}^{N} 1(h(\boldsymbol{x}^t) \neq r^t)$$

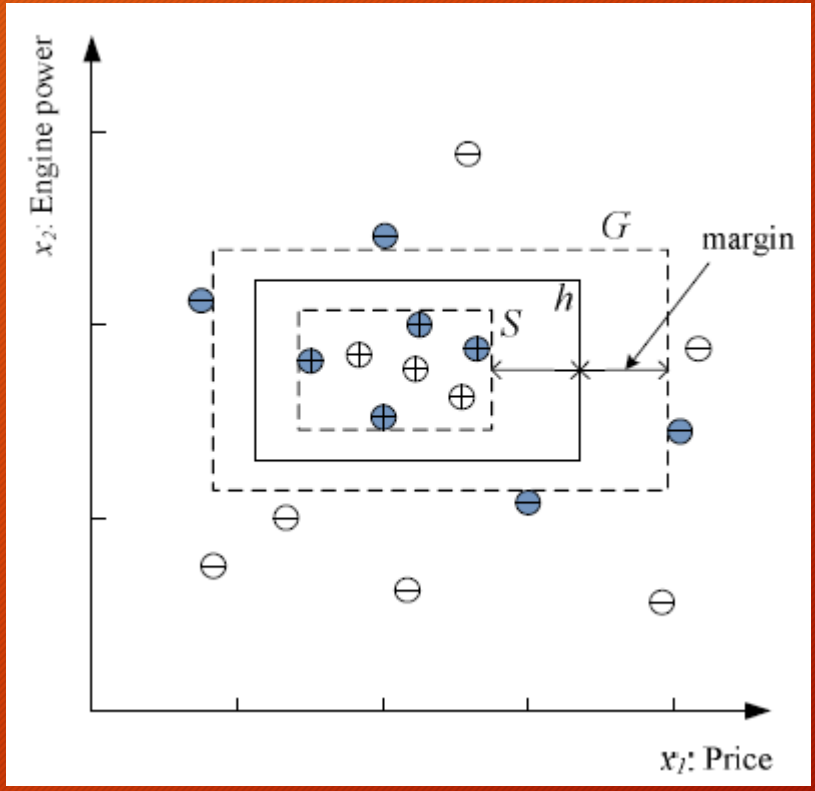
S, G, and the Version Space



S, the most specific hypothesis, *G*, the most general hypothesis, and any $h \in H$ between *S* and *G* is consistent and make up the version space
(Mitchell, 1997)

Margin

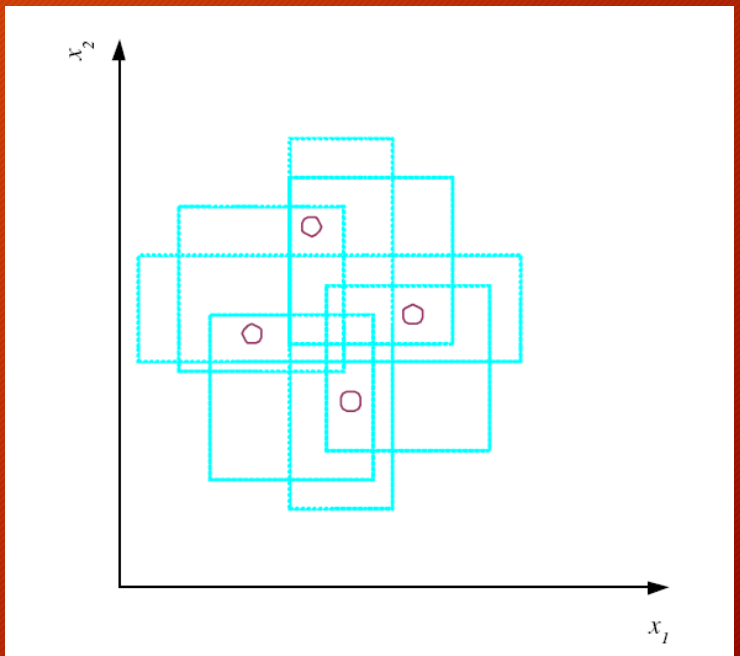
- Choose *h* with largest margin



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 !

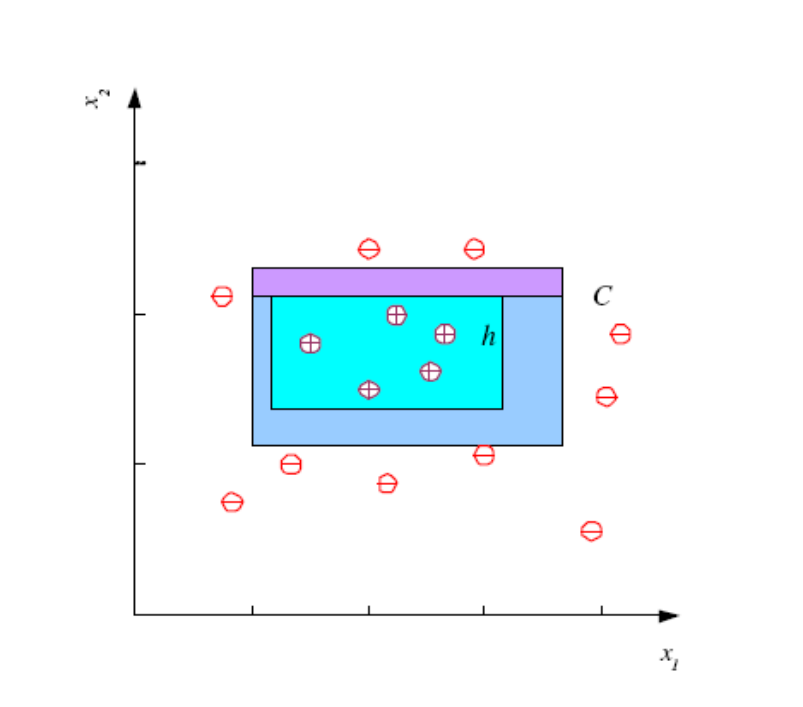


Probably Approximately Correct (PAC) Learning

- 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)

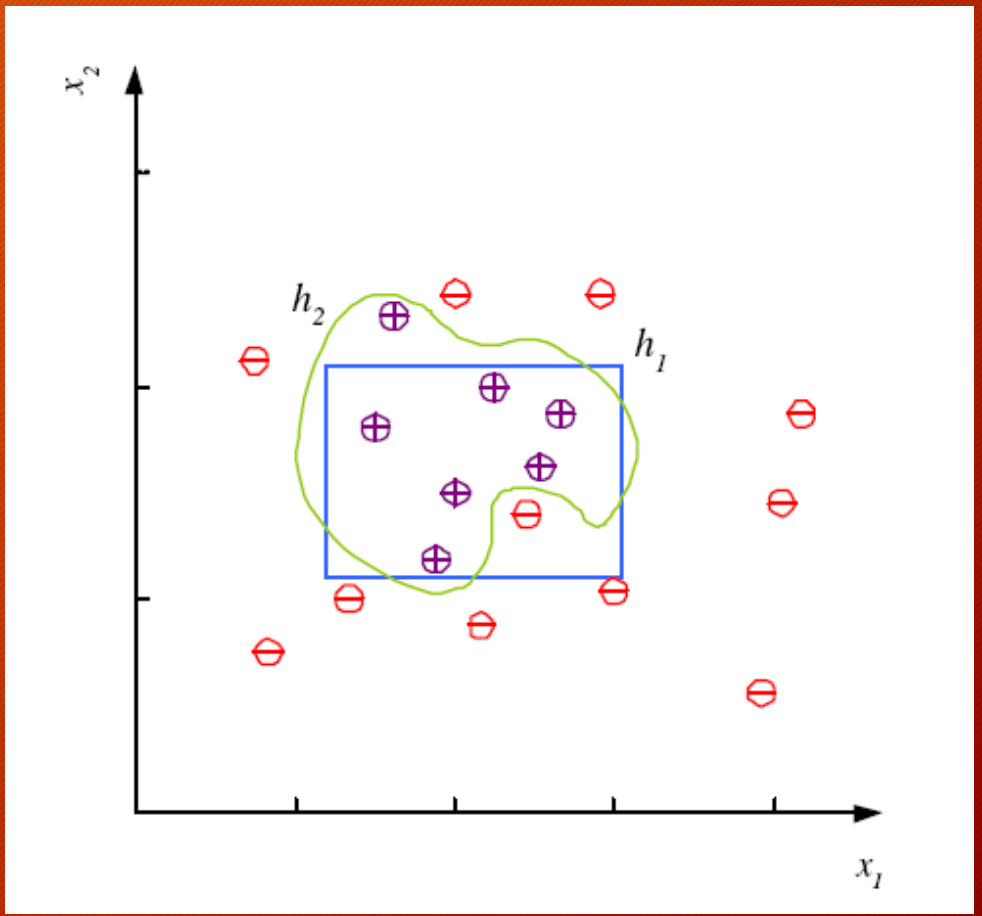
- Each strip is at most $\varepsilon/4$
- Pr that we miss a strip $1 - \varepsilon/4$
- Pr that N instances miss a strip $(1 - \varepsilon/4)^N$
- Pr that N instances miss 4 strips $4(1 - \varepsilon/4)^N$
- $4(1 - \varepsilon/4)^N \leq \delta$ and $(1 - x) \leq \exp(-x)$
- $4\exp(-\varepsilon N/4) \leq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



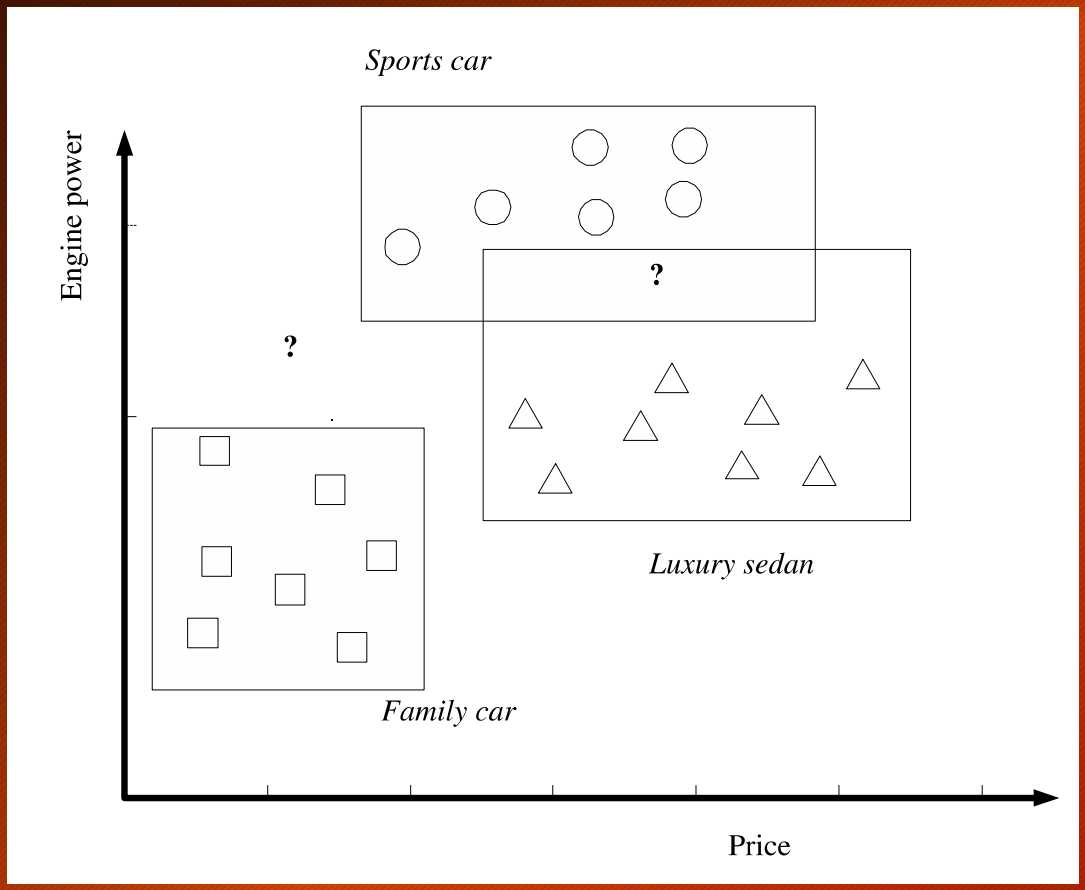
Noise and Model Complexity

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



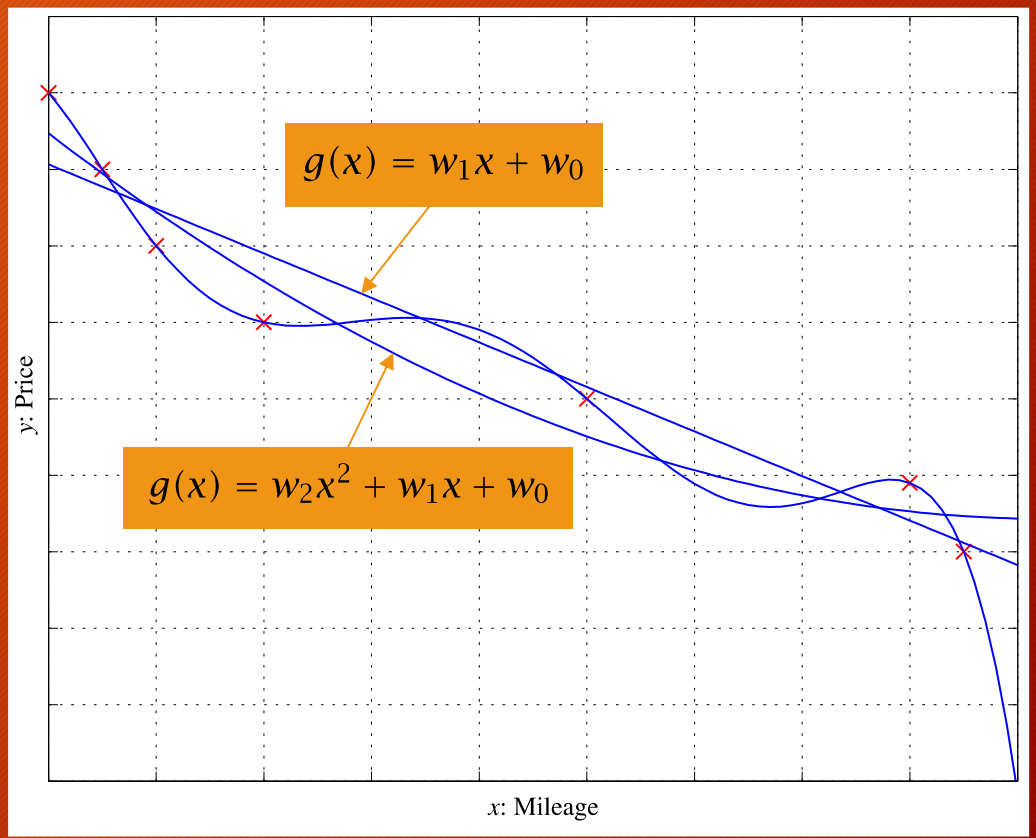
Train hypotheses $h_i(\boldsymbol{x})$, i =1,...,K, one to separate each class from the union of all others

?: reject regions where no or more than one $h_i(\boldsymbol{x})>0$.

Regression

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

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



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 $\mathcal{H}$
- 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

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 *increases*, E *decreases*
- As $c(\mathcal{H})$ *increases*, first E *decreases* and then E increases.

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

Dimensions of a Supervised Learner

1. Model: $g(x|\theta)$
2. Loss function: $E(\theta|\mathcal{X}) = \sum_t L(r^t, g(x^t|\theta))$
3. Optimization procedure:

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