

Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

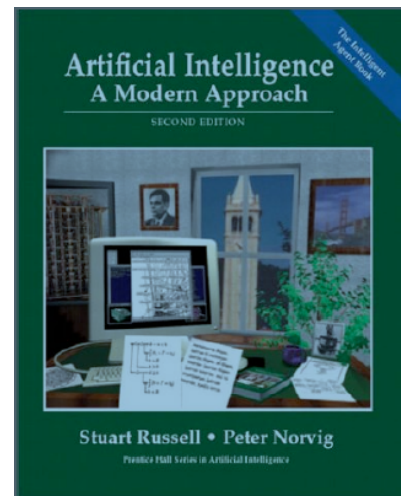
© The MIT Press, 2004

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml>

Lab Class and literature

- n Friday, 10.00 – 11.00, Schwarzenbergstrasse 95, H0.09
- n First Lab Class 11.04.08
- n Mainly used book





CHAPTER 1:

Introduction



Why “Learn” ?

- n Machine learning is programming computers to optimize a *performance criterion* using example data or past experience.
- n There is no need to “learn” to calculate payroll
- n Learning is used when:
 - Human expertise does not exist (navigating on Mars),
 - Humans are unable to explain their expertise (speech recognition)
 - Solution changes in time (routing on a computer network)
 - Solution needs to be adapted to particular cases (user biometrics)



What We Talk About When We Talk About “Learning”

- n Learning general models from a data of particular examples
- n Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- n Example in retail: Customer transactions to consumer behavior:
 - People who bought “Da Vinci Code” also bought “The Five People You Meet in Heaven” (www.amazon.com)*
- n Build a model that is *a good and useful approximation* to the data.



Data Mining

- n **Retail:** Market basket analysis, Customer relationship management (CRM)
- n **Finance:** Credit scoring, fraud detection
- n **Manufacturing:** Optimization, troubleshooting
- n **Medicine:** Medical diagnosis
- n **Telecommunications:** Quality of service optimization
- n **Bioinformatics:** Motifs, alignment
- n **Web mining:** Search engines
- n ...



What is Machine Learning?

- n Optimize a performance criterion using example data or past experience.
- n Role of Statistics: Building mathematical models, core task is inference from a sample
- n Role of Computer science: Efficient algorithms to
 - “ Solve the optimization problem
 - “ Representing and evaluating the model for inference



Sample Applications

- n Learning Associations
- n Supervised Learning
 - Classification
 - Regression
- n Unsupervised Learning
- n Reinforcement Learning

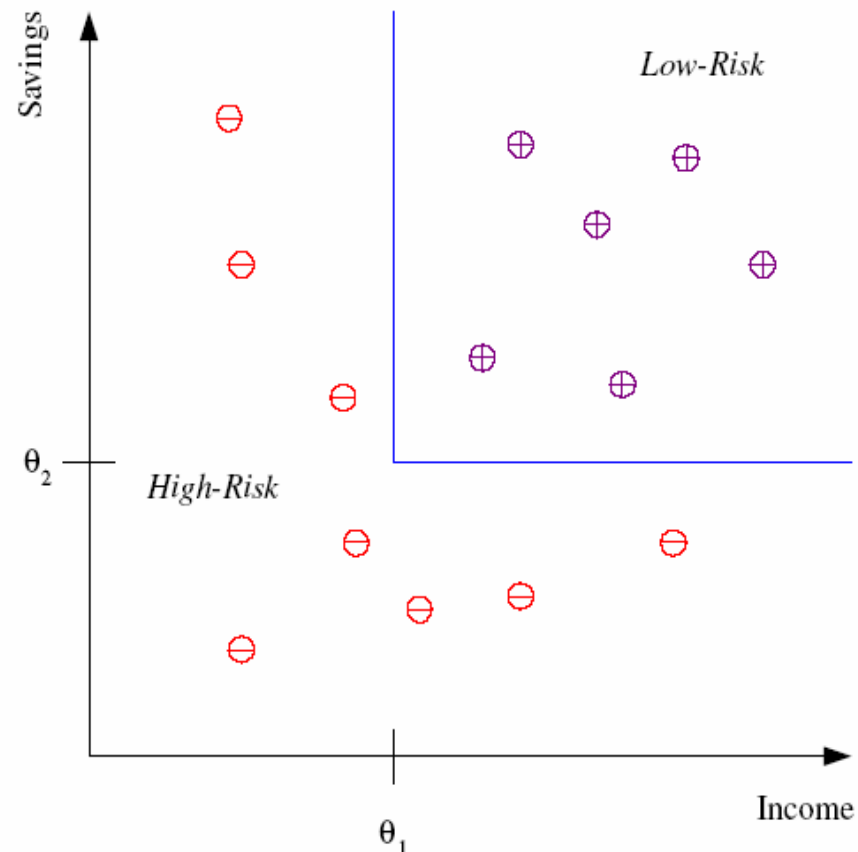


Learning Associations

- n Basket analysis:
 $P(Y | X)$ probability that somebody who buys X also buys Y where X and Y are products/services.
Example: $P(\text{chips} | \text{beer}) = 0.7$
- n If we know more about customers or make a distinction among them:
 - $P(X | Y, D)$
where D is the customer profile (age, gender, marital status, ...)
 - In case of a Web portal, items correspond to links to be shown/prepared/downloaded in advance

Classification

- n Example: Credit scoring
- n Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF $income > \theta_1$ AND $savings > \theta_2$
THEN low-risk ELSE high-risk



Classification: Applications

- n Aka Pattern recognition
- n **Character recognition:** Different handwriting styles.
- n **Face recognition:** Pose, lighting, occlusion (glasses, beard), make-up, hair style
- n **Speech recognition:** Temporal dependency.
 - Use of a dictionary or the syntax of the language.
 - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- n **Medical diagnosis:** From symptoms to illnesses
- n ...



Face Recognition

Training examples of a person



Test images



AT&T Laboratories, Cambridge UK
<http://www.uk.research.att.com/facedatabase.html>

Regression

n Example: Price of a used car

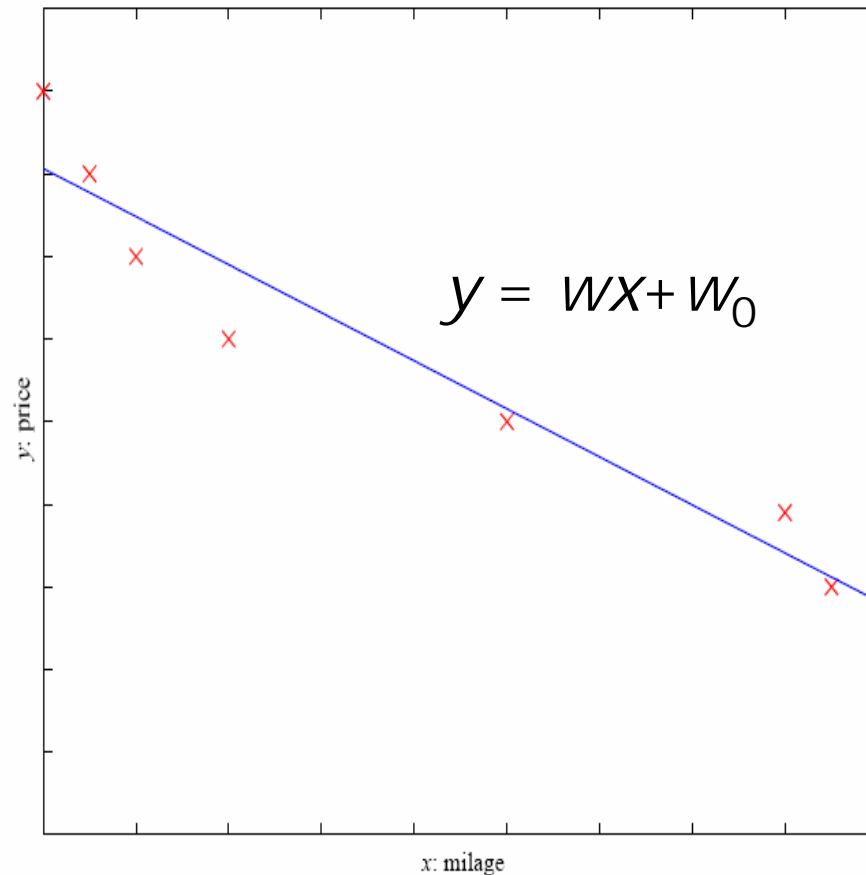
n x : car attributes

y : price

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

$g()$ model,

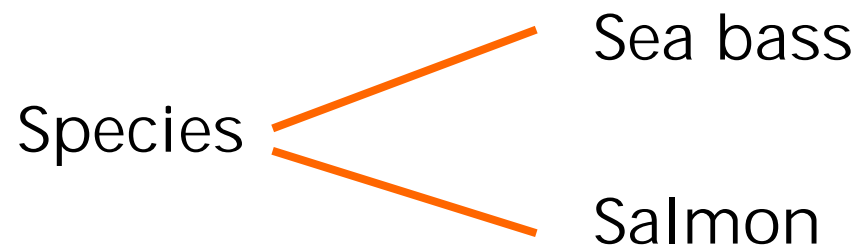
θ parameters





An Extended Example

- n “Sorting incoming Fish on a conveyor according to species using optical sensing”





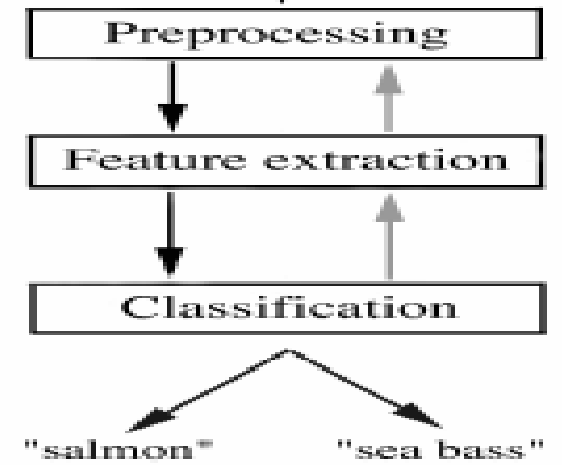
n Problem Analysis

- .. Set up a camera and take some sample images to extract features
 - n Length
 - n Lightness
 - n Width
 - n Number and shape of fins
 - n Position of the mouth, etc...
- n This is the set of all suggested features to explore for use in our classifier!



n Preprocessing

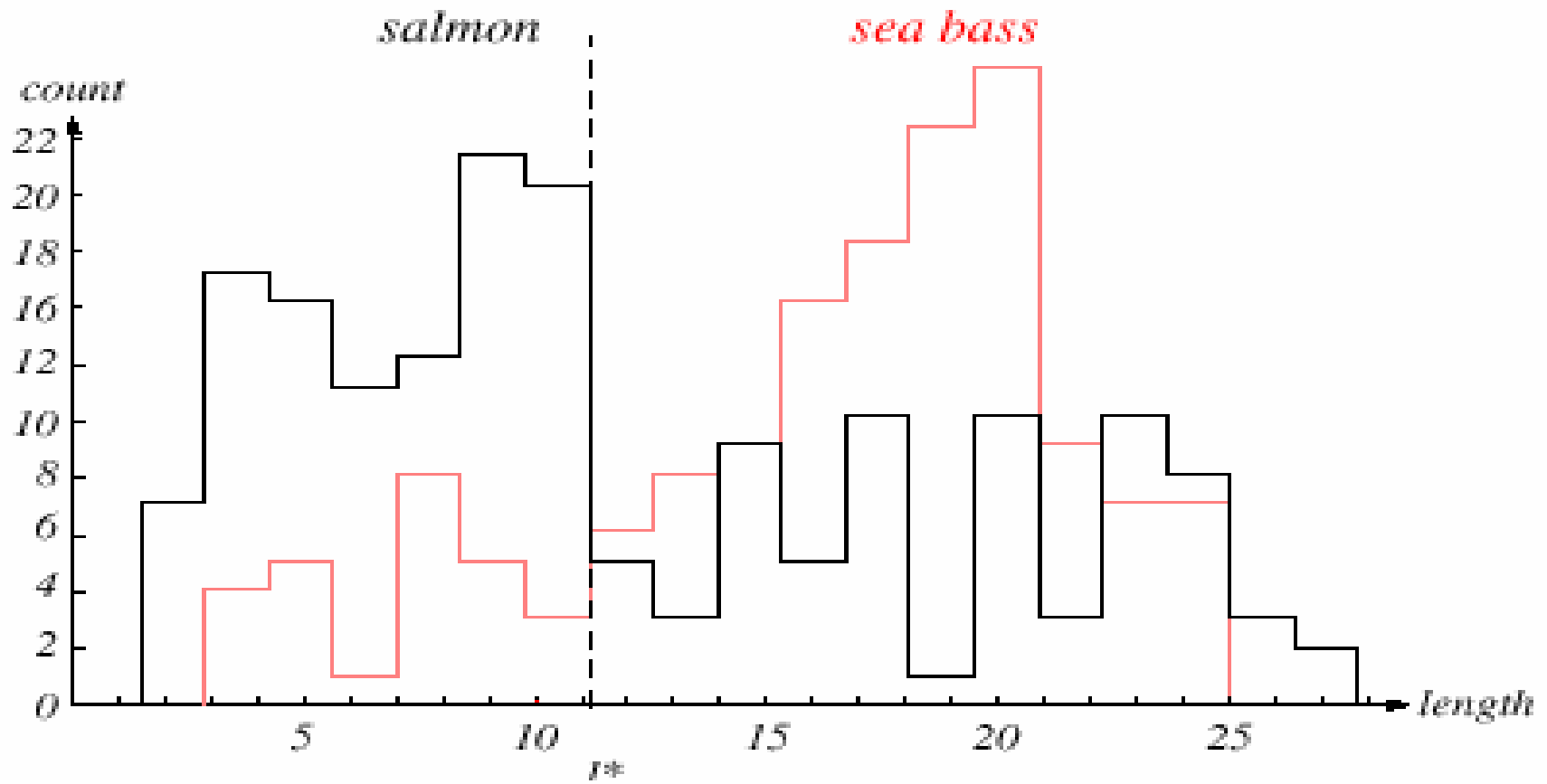
- Use a segmentation operation to isolate fishes from one another and from the background
- n Information from a single fish is sent to a feature extractor whose purpose is to reduce the data by measuring certain features
- n The features are passed to a classifier





n Classification

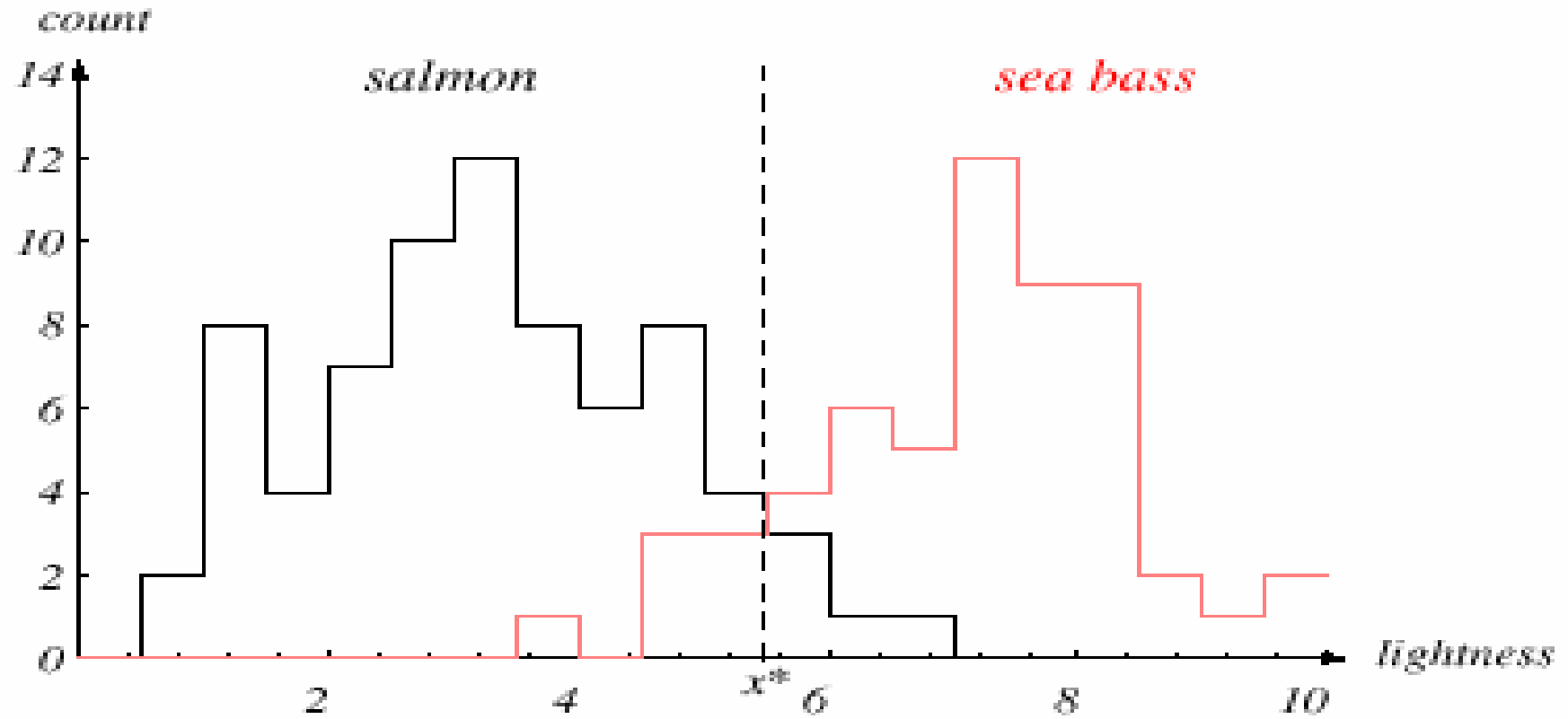
- Select the length of the fish as a possible feature for discrimination





The **length** is a poor feature alone!

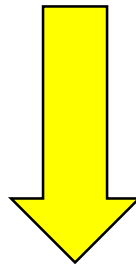
Select the **lightness** as a possible feature.





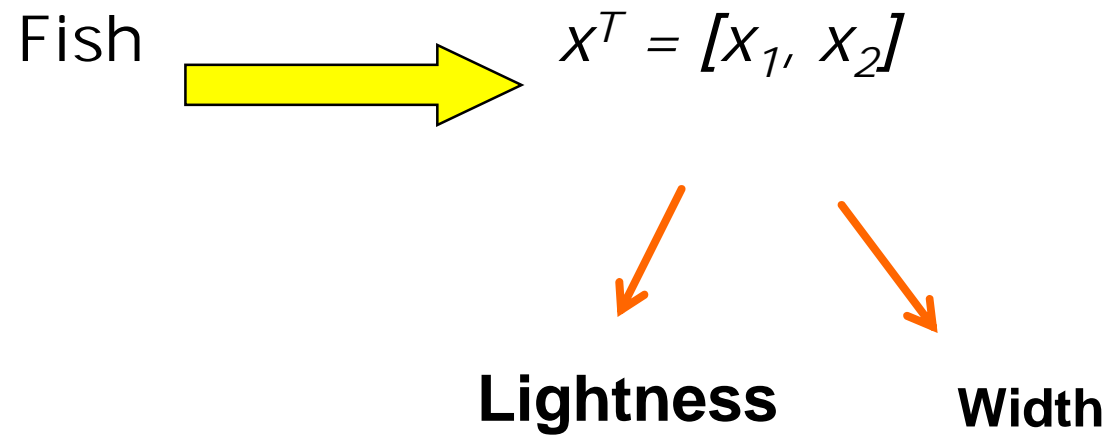
n Threshold decision boundary and cost relationship

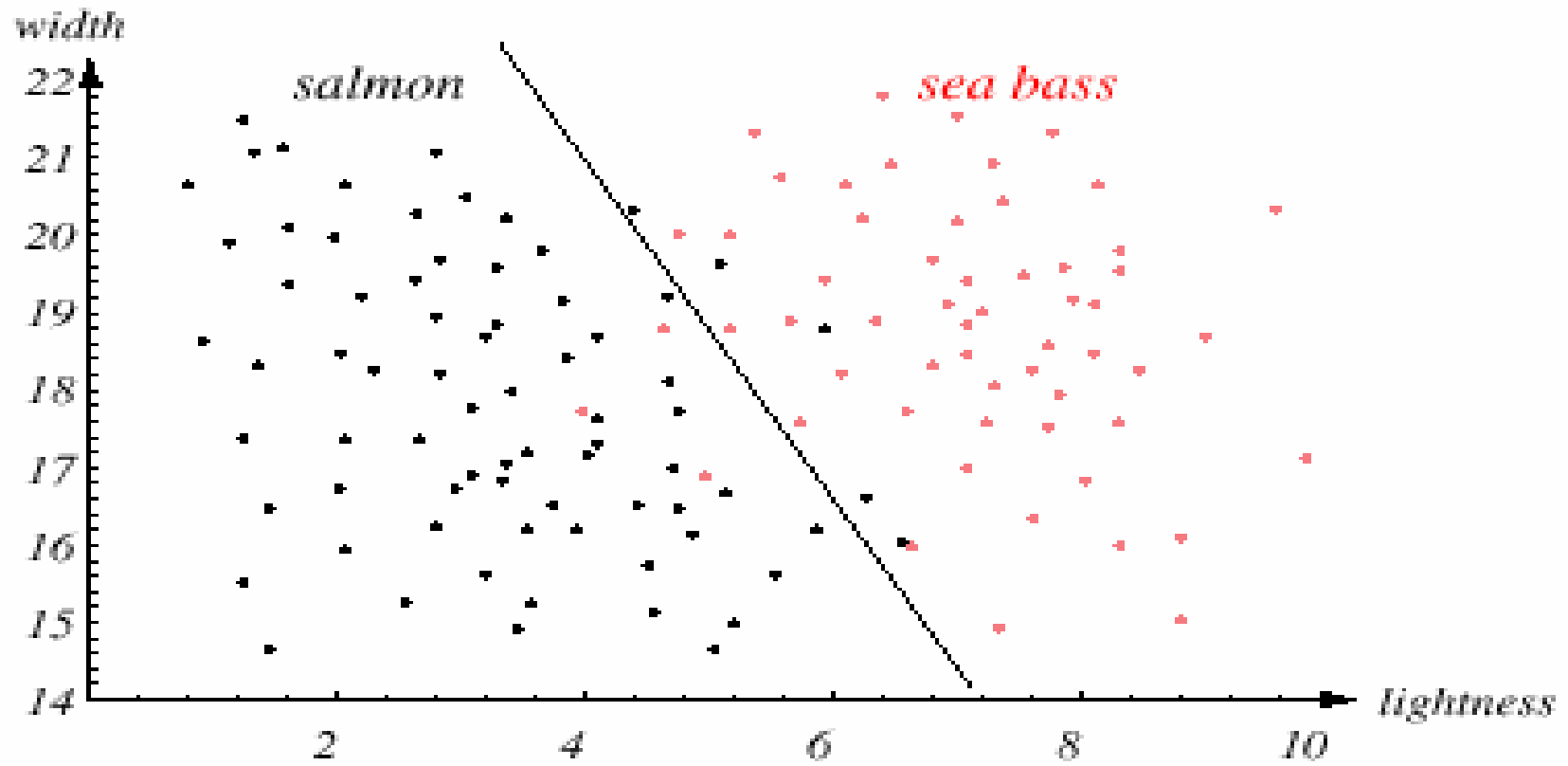
- Move our decision boundary toward smaller values of lightness in order to minimize the cost (reduce the number of sea bass that are classified salmon!)



Task of decision theory

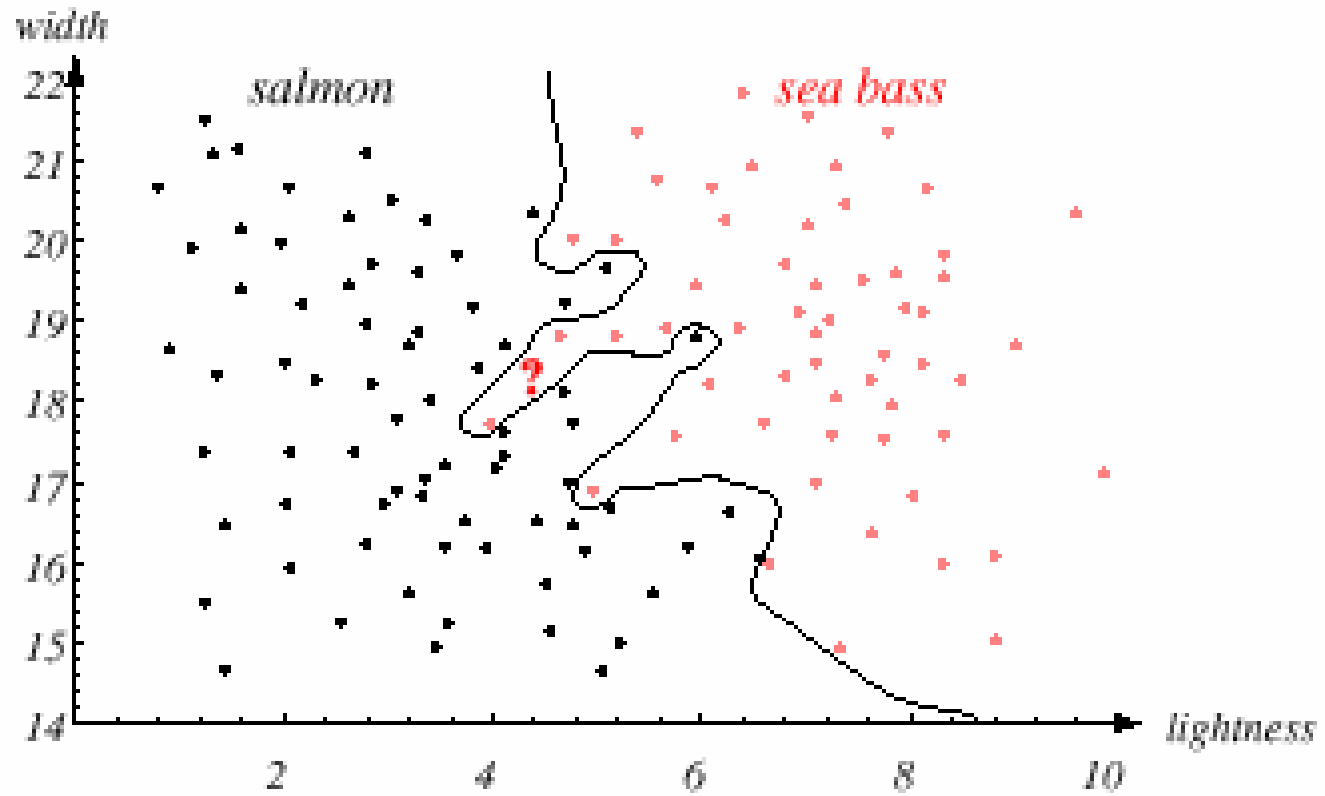
- n Adopt the lightness and add the width of the fish








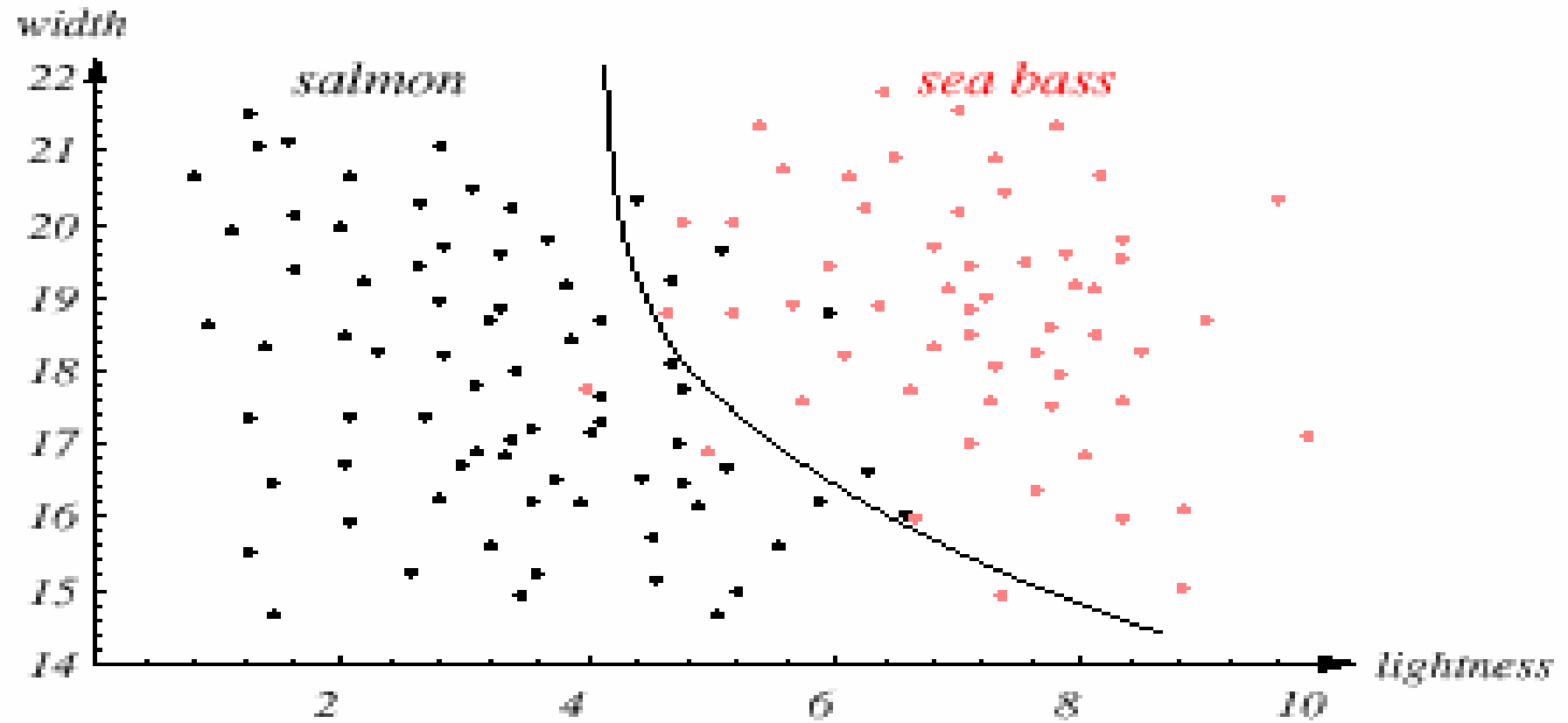
- n We might add other features that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding such “noisy features”
- n Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:



- 
- n However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input



Issue of generalization!





Supervised Learning: Uses

- n **Prediction of future cases:** Use the rule to predict the output for future inputs
- n **Knowledge extraction:** The rule is easy to understand
- n **Compression:** The rule is simpler than the data it explains
- n **Outlier detection:** Exceptions that are not covered by the rule, e.g., fraud



Unsupervised Learning

- n Learning “what normally happens”
- n No output
- n Clustering: Grouping similar instances
- n Example applications
 - “ Customer segmentation in CRM
 - “ Image compression: Color quantization
 - “ Bioinformatics: Learning motifs
 - “ Document Classification in unknown Domains.



Reinforcement Learning

- n Learning a policy: A **sequence** of actions/outputs
- n No supervised output but delayed reward
- n Credit assignment problem
- n Game playing
- n Robot in a maze
- n Multiple agents, partial observability, ...



CHAPTER 2:

Supervised Learning



Learning a Class from Examples

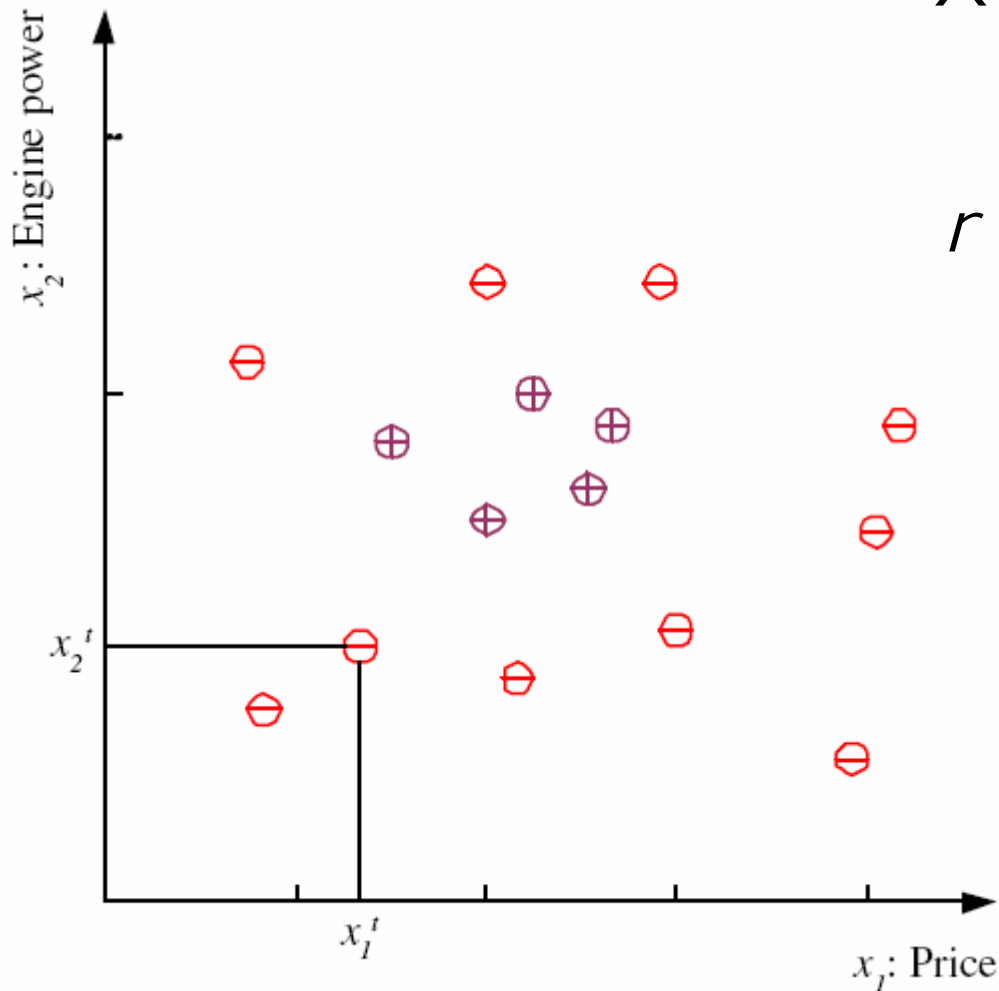
- n Class C of a “family car”
 - “ Prediction: Is car x a family car?
 - “ Knowledge extraction: What do people expect from a family car?
- n Output:
 - Positive (+) and negative (–) examples
- n Input representation:
 - x_1 : price, x_2 : engine power

Training set X

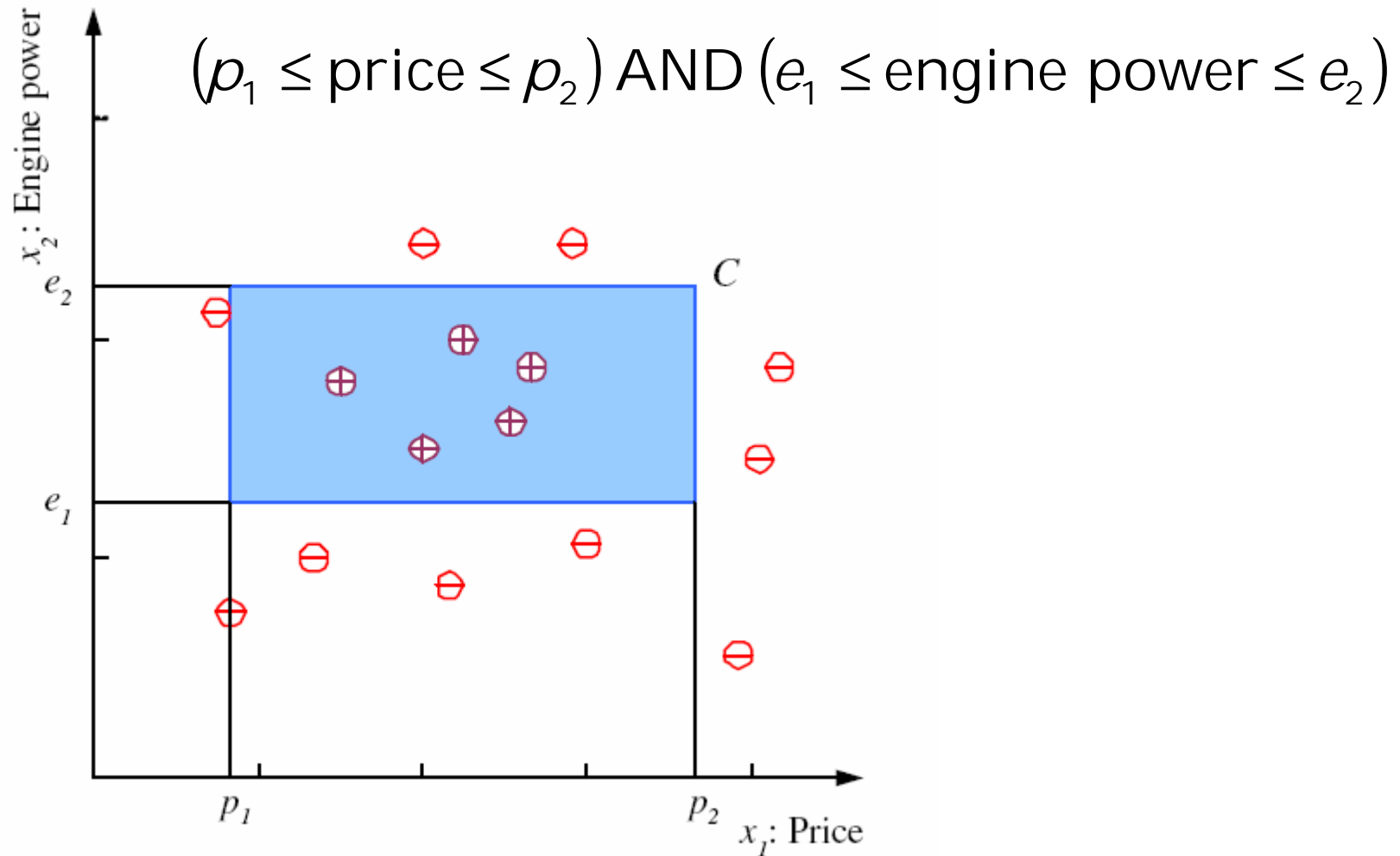
$$X = \{\mathbf{x}^t, r^t\}_{t=1}^N$$

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

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

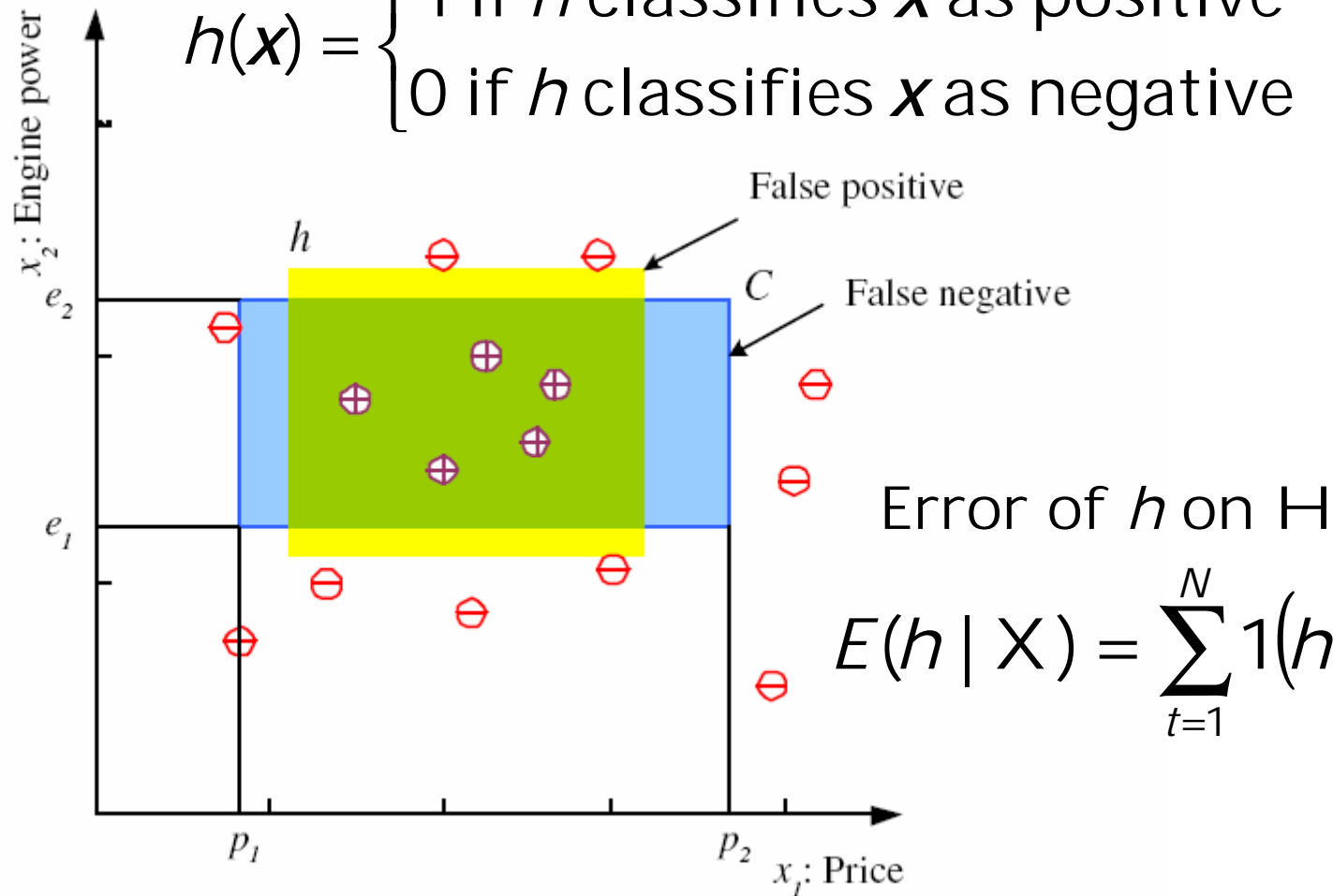


Class C



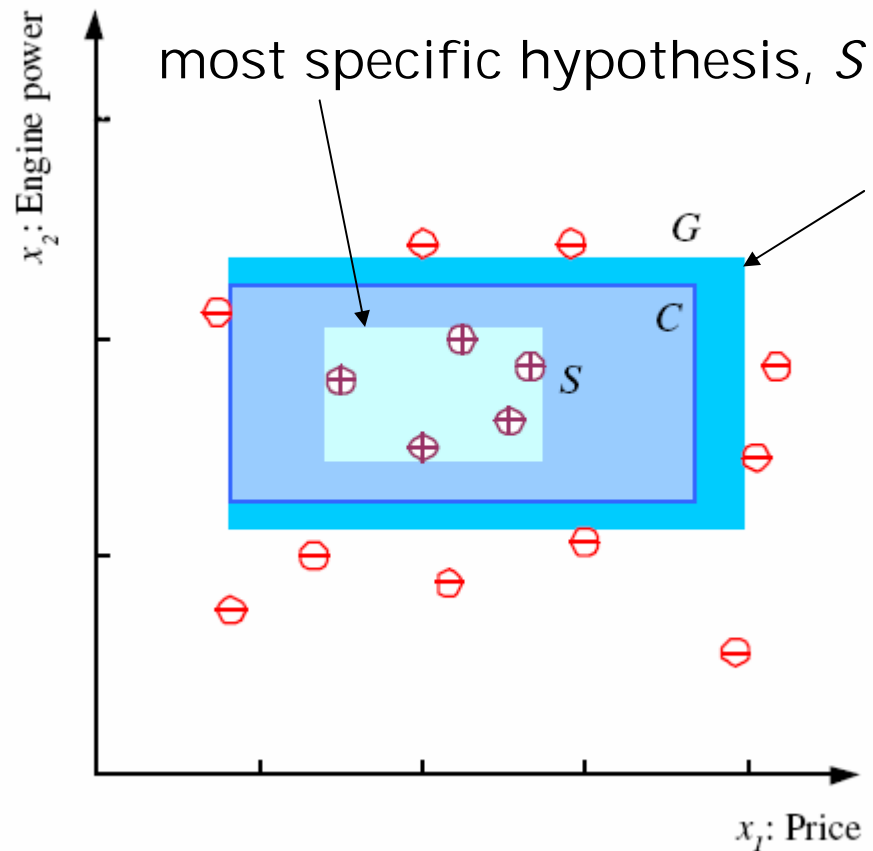
Hypothesis class H

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



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

S, G, and the Version Space



$h \in H$, between S and G is
consistent

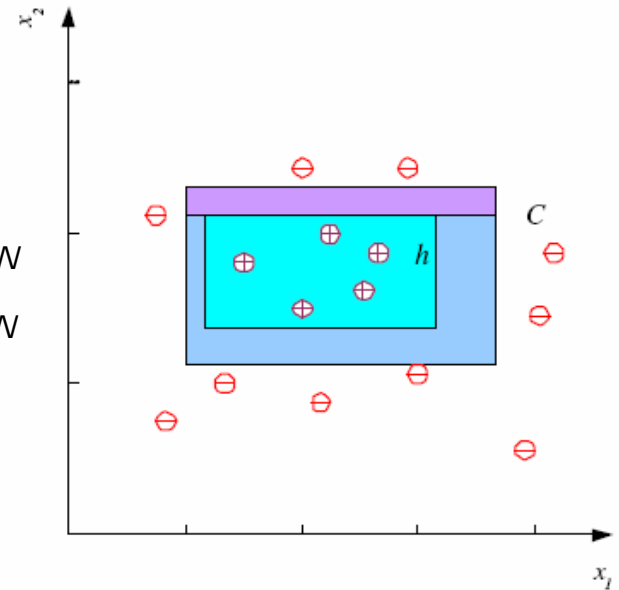
and make up the
version space

(Mitchell, 1997)

Probably Approximately Correct (PAC) Learning

n 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)

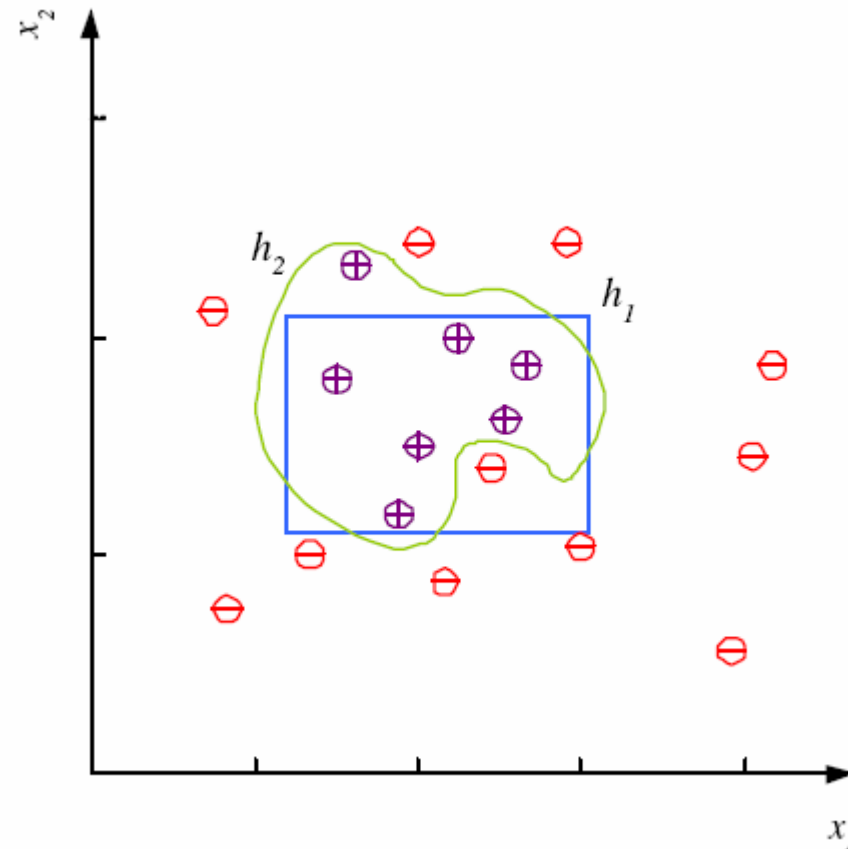
- n Upper bound for each strip should be $\epsilon/4$
è $4(\epsilon/4) = \epsilon$
- n Pr that we miss one strip $1 - \epsilon/4$
- n Pr that N instances miss one strip $(1 - \epsilon/4)^N$
- n Pr that N instances miss 4 strips $4(1 - \epsilon/4)^N$
- n $4(1 - \epsilon/4)^N \leq \delta$ use $(1 - x) \leq \exp(-x)$
- n $4\exp(-\epsilon N/4) \leq \delta$ and $N \geq (4/\epsilon)\log(4/\delta)$



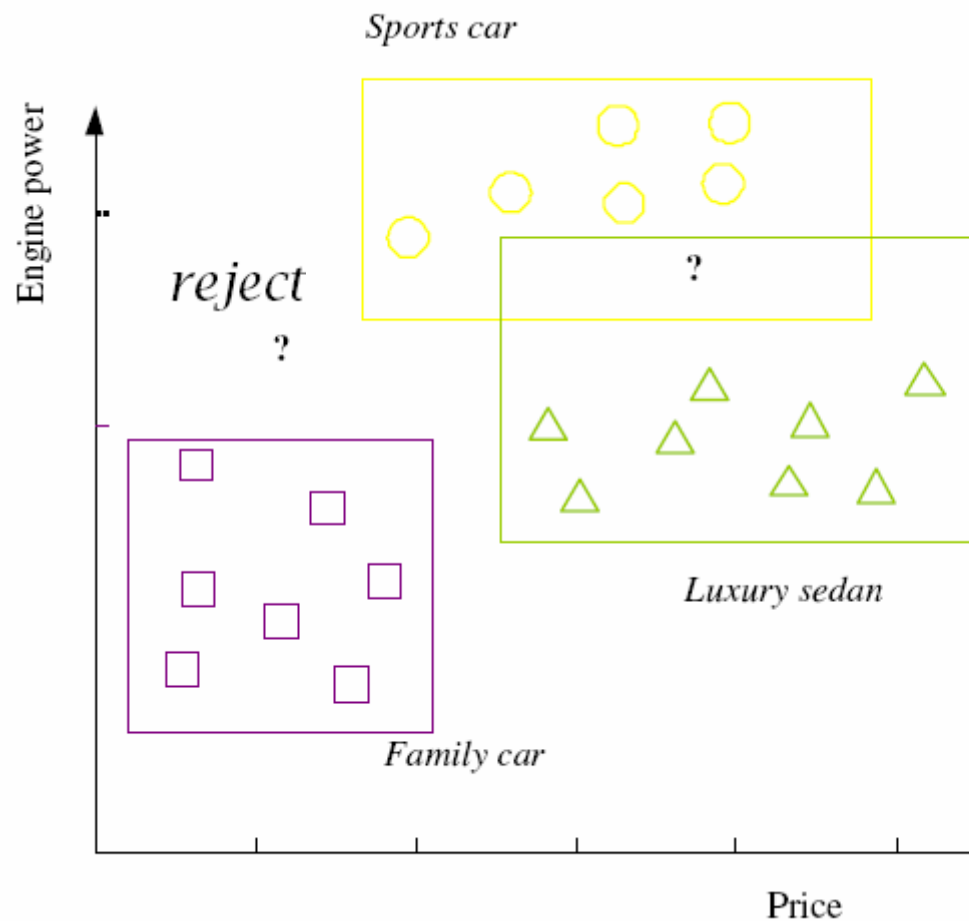
Noise and Model Complexity

Use the simpler one because

- n Simpler to use (lower computational complexity)
- n Easier to train (lower space complexity)
- n Easier to explain (more interpretable)
- n Generalizes better (lower variance - Occam's razor)



Multiple Classes, $C_i, i=1, \dots, K$



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

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

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

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

Regression

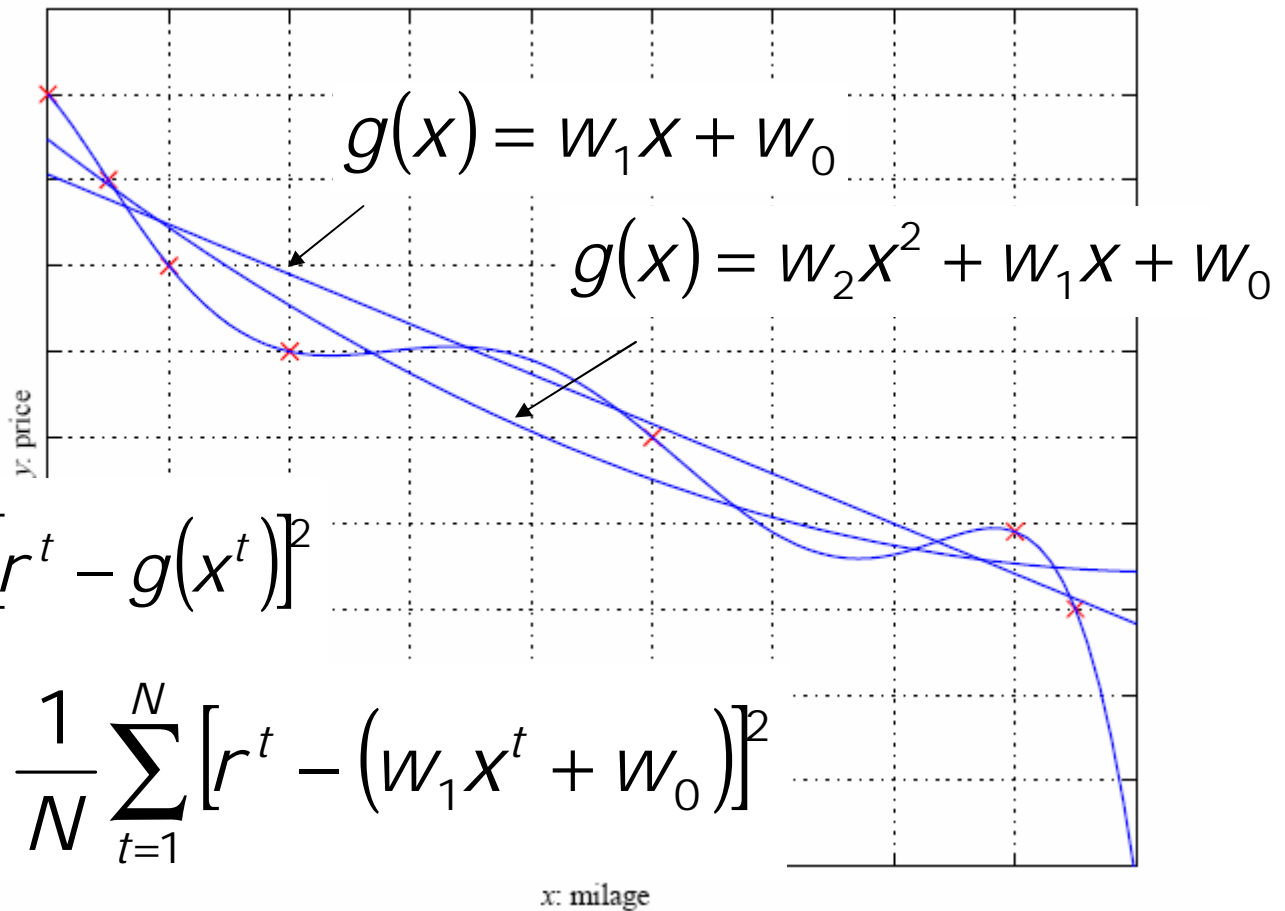
$$X = \{x^t, r^t\}_{t=1}^N$$

$$r^t \in \mathfrak{R}$$

$$r^t = f(x^t)$$

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

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





Model Selection & Generalization

- n Learning is an **ill-posed problem**; data is not sufficient to find a unique solution
- n The need for **inductive bias**, assumptions about H
- n **Generalization**: How well a model performs on new data
- n Overfitting: H more complex than C or f
- n Underfitting: H less complex than C or f



Triple Trade-Off

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



Cross-Validation

- n 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%)
- n Resampling when there is few data