Introduction to Machine Learning

Inas A. Yassine, PhD

Assoc. Prof. , Systems and Biomedical Engineering Department, Cairo University

inas.yassine@eng.cu.edu.eg

Linear regression

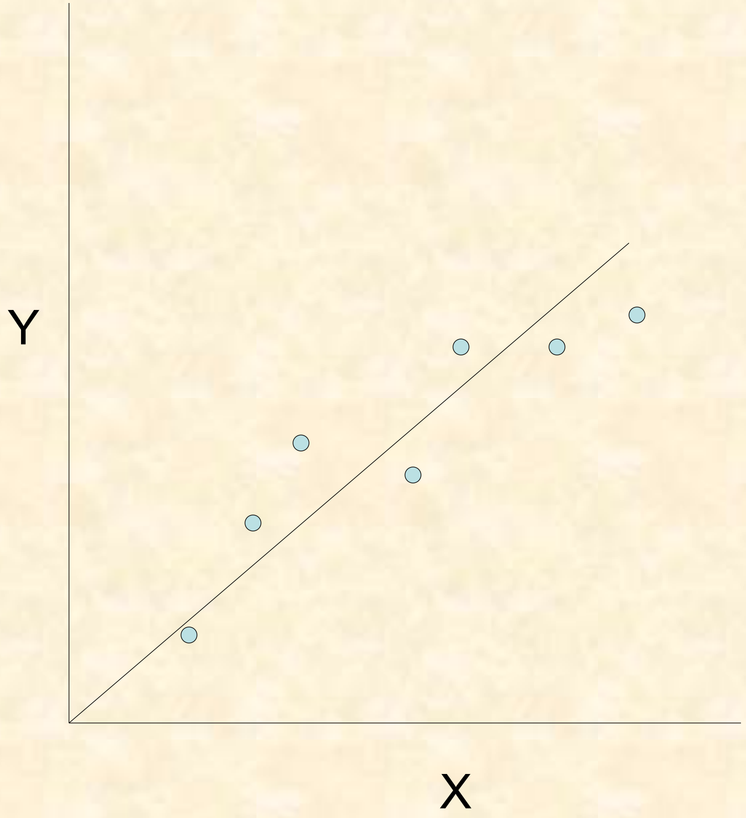
- Given an input x we would like to compute an output y
- In linear regression we assume that y and x are related with the following equation:

What we are trying to predict

Observed values

$$y = wx + \varepsilon$$

where w is a parameter and ε represents measurement or other noise



Linear regression

- Our goal is to estimate *w* from a training data of <x_i,y_i> pairs

- Optimization goal: minimize squared error (least squares):

$$\arg\min_w \sum_i (y_i - wx_i)^2$$

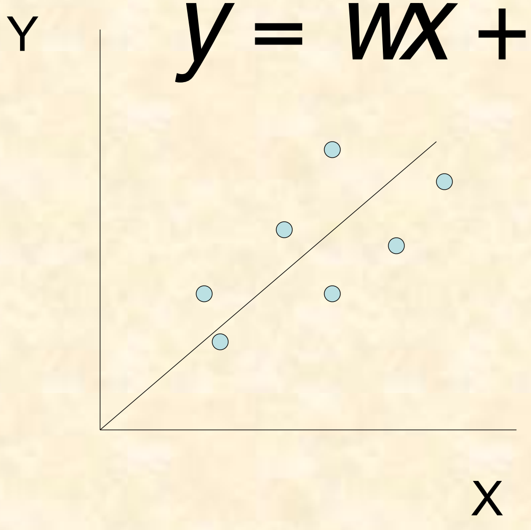
- Why least squares?

  - minimizes squared distance between measurements and predicted line

  - has a nice probabilistic interpretation

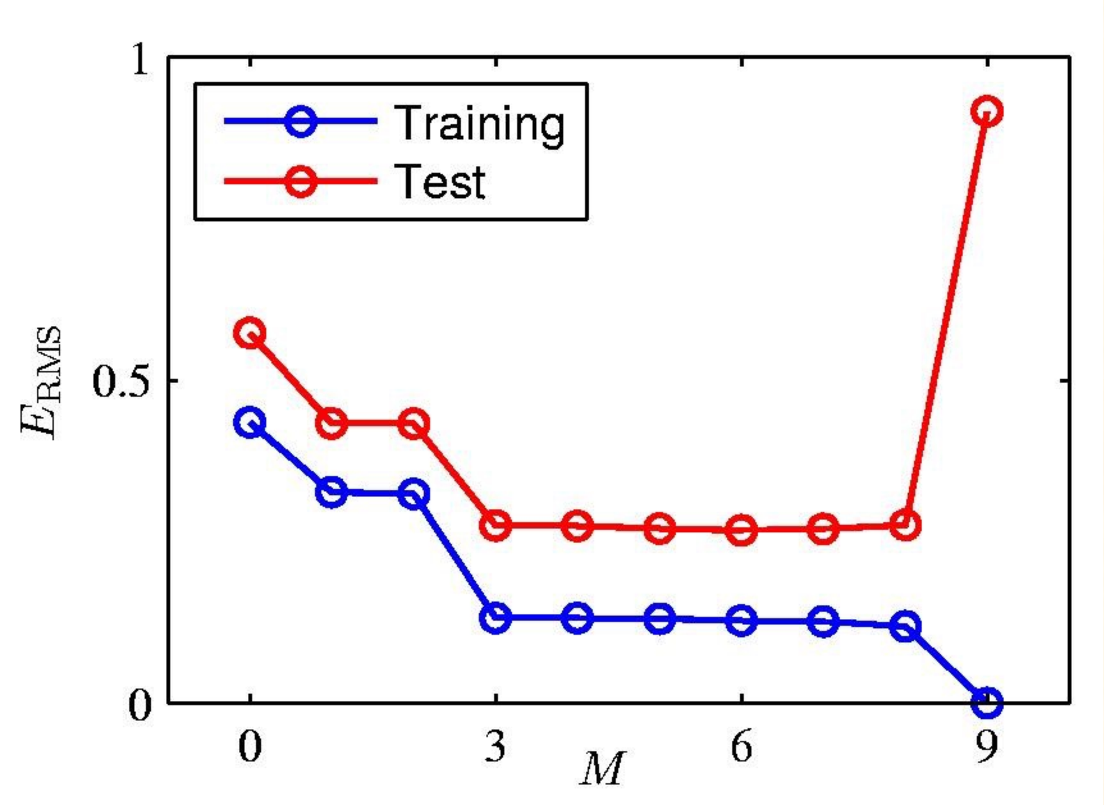
  - the math is pretty

$$y = wx + \varepsilon$$



Overfitting in Regression

Over-fitting



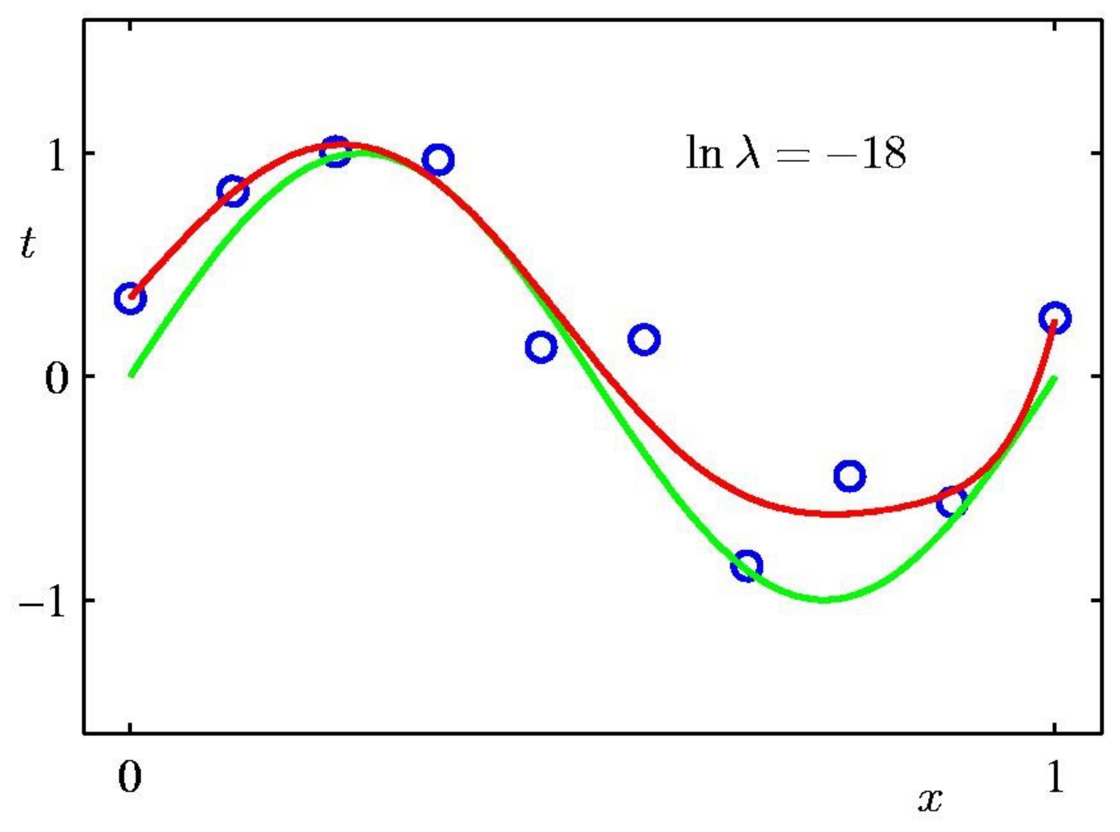
Root-Mean-Square (RMS) Error: $E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

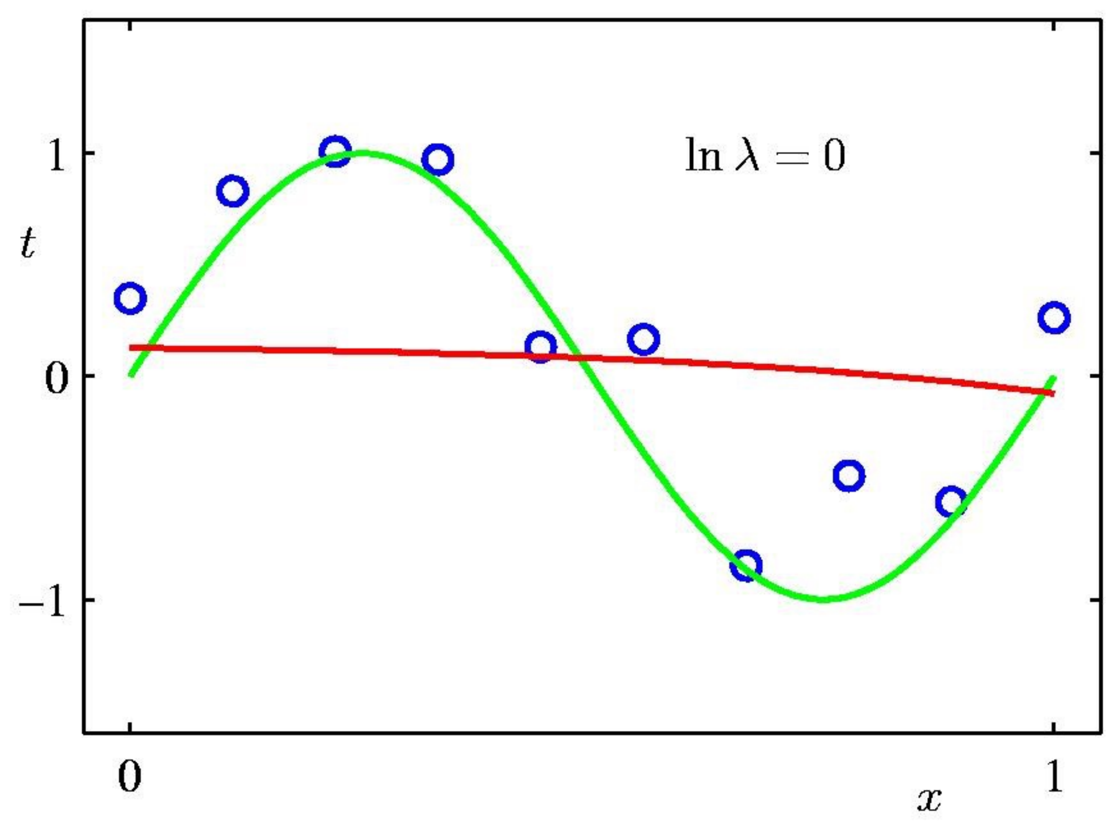
Regularization

- Penalize Large Coefficients

$$J_{\mathbf{x},\mathbf{y}}(\mathbf{w}) = \frac{1}{2}\sum_{i}\left(y^{i} - \sum_{j} w_{j}\phi_{j}(\mathbf{x}^{i})\right)^{2} - \frac{\lambda}{2}\left\|\mathbf{w}\right\|^{2}$$

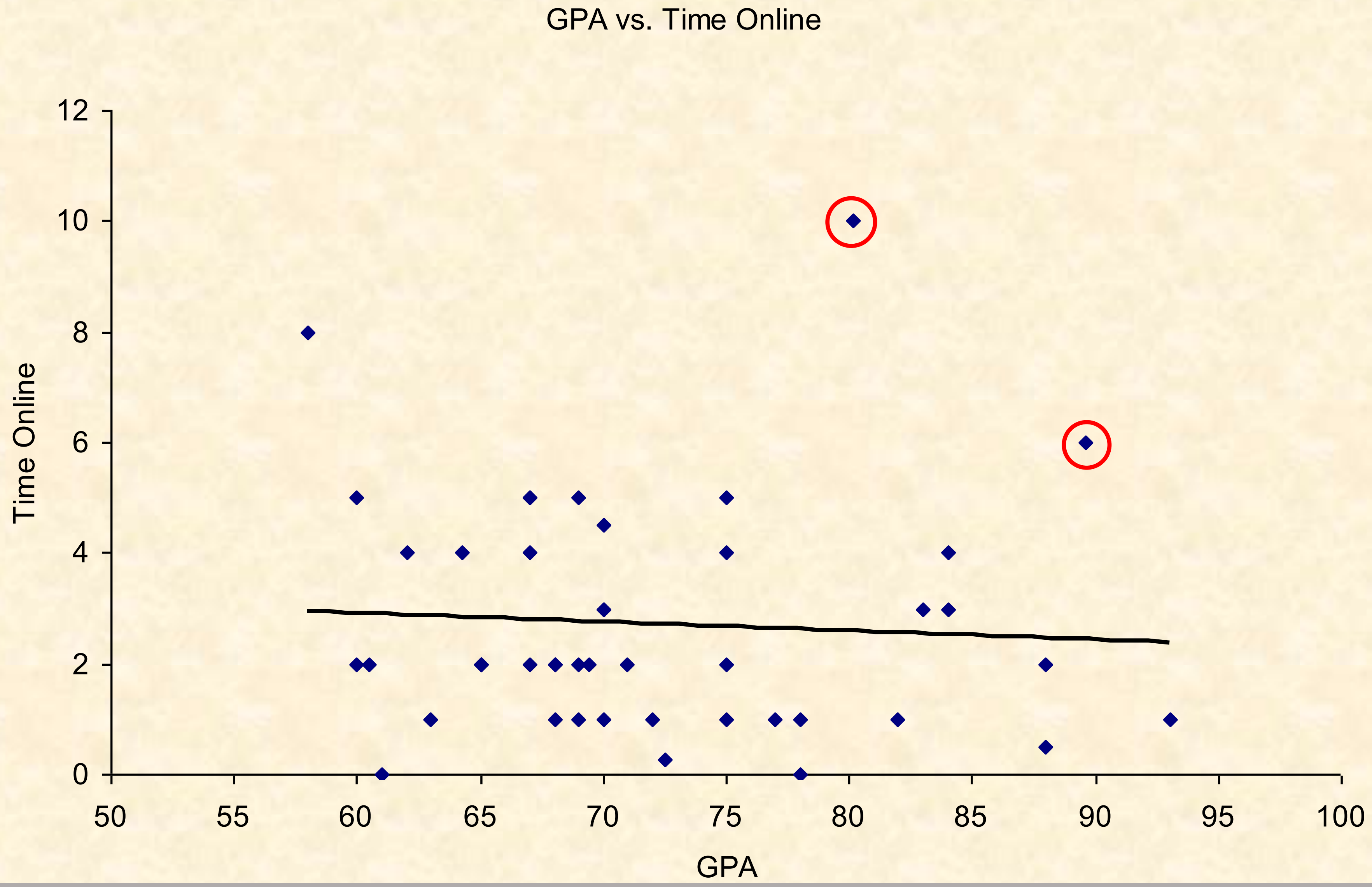
Regularization/ Over-Regularization





Outliers

- Rare/ Extreme values that may destroy the learning, which could be:
  - Error
  - Important observation
- outliers if detected if greater than 3 standard deviation from the mean


GPA vs. Time Online
Time Online
0
2
4
6
8
10
12
GPA
50
55
60
65
70
75
80
85
90
95
100

Generative Vs Discriminative classifier

- Generative classifier, e.g., Naïve Bayes:
  - –Assume some functional form for **P(XIY), P(Y)**
  - –Estimate parameters of P(XIY), P(Y) directly from training data
  - –Use Bayes rule to calculate P(YIX=x)
  - –This is ‘generative’ model
    - Indirect computation of P(YIX) through Bayes rule
    - But, can generate a sample of the data,

- Discriminative classifier, e.g., Logistic Regression:
  - –Assume some functional form for **P(YIX)**
  - –Estimate parameters of P(YIX) directly from training data
  - –This is the ‘discriminative’ model
  - Directly learn P(YIX)
  - But cannot sample data, because P(X) is not available

Bayesian Decision Theory

Outline

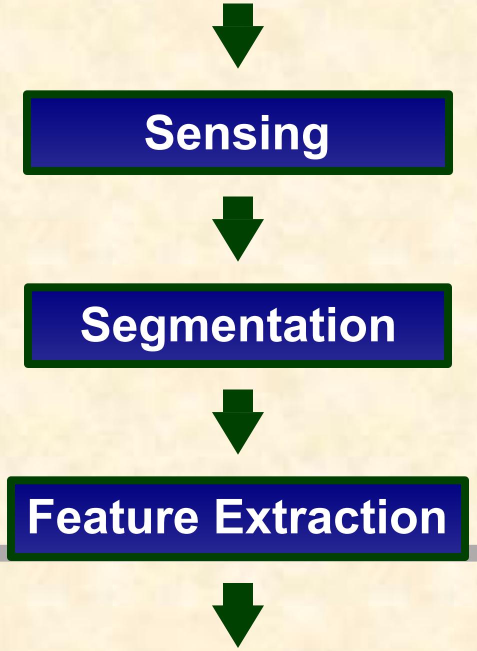
- What is classification?
- Classification by Bayesian Classification
- **Basic Concepts**
- **Bayes Rule**
- More General Forms of Bayes Rule
- Discriminated Functions
- Bayesian Belief Networks

TYPICAL APPLICATIONS OF PR

IMAGE PROCESSING EXAMPLE

- **Sorting Fish:** incoming fish are sorted according to species using optical sensing (sea bass or salmon?)

- **Problem Analysis:**
  - set up a camera and take some sample images to extract features
  - Consider features such as length, lightness, width, number and shape of fins, position of mouth, etc.



Sensing

Segmentation

Feature Extraction

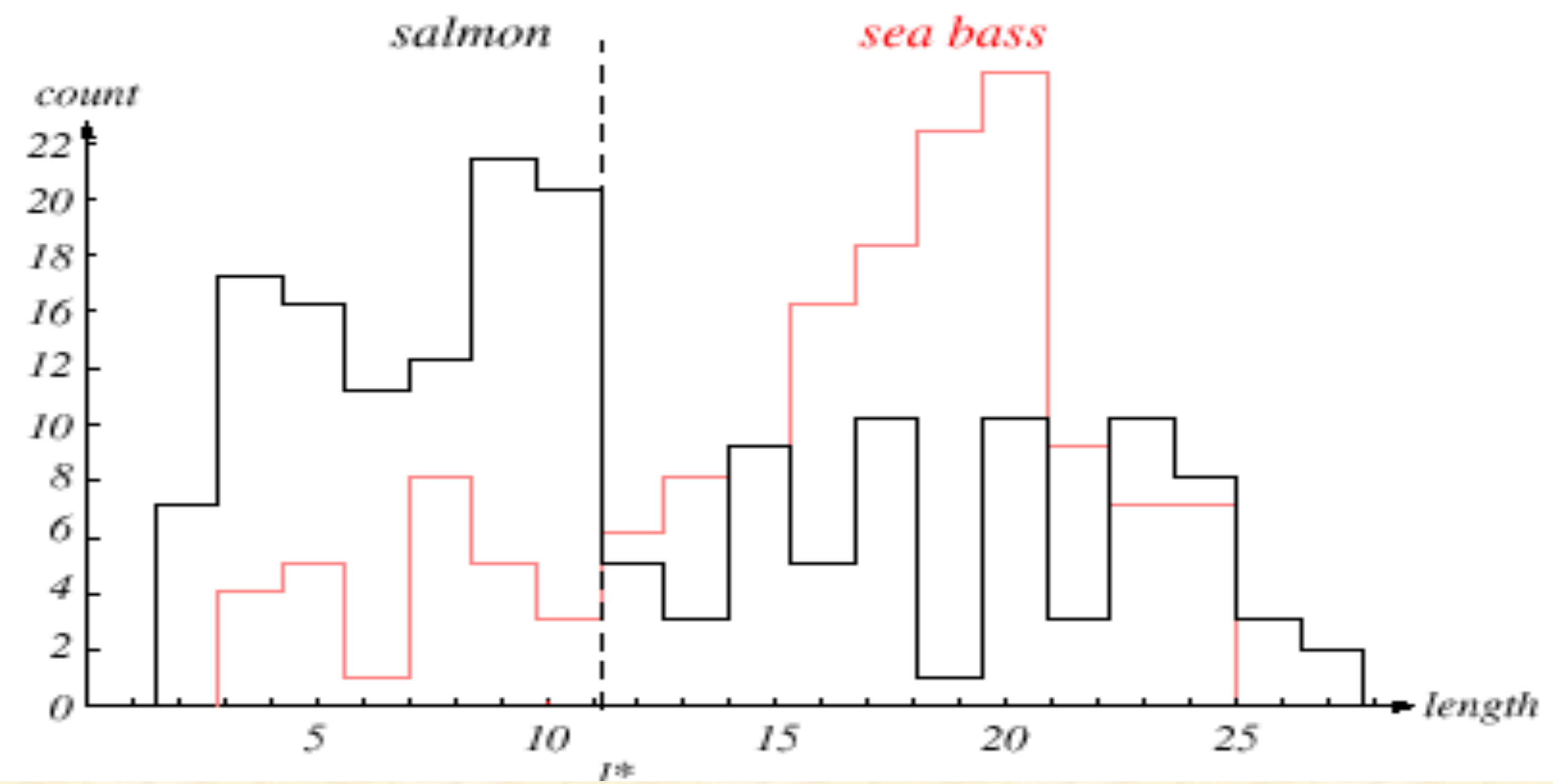
Pattern Classification System

- Preprocessing
  - Segment (isolate) fishes from one another and from the background
- Feature Extraction
  - Reduce the data by measuring certain features
- Classification
  - Divide the feature space into decision regions

Classification

- Initially use the length of the fish as a possible feature for discrimination

Length Discriminator



- **Length is a poor discriminator**

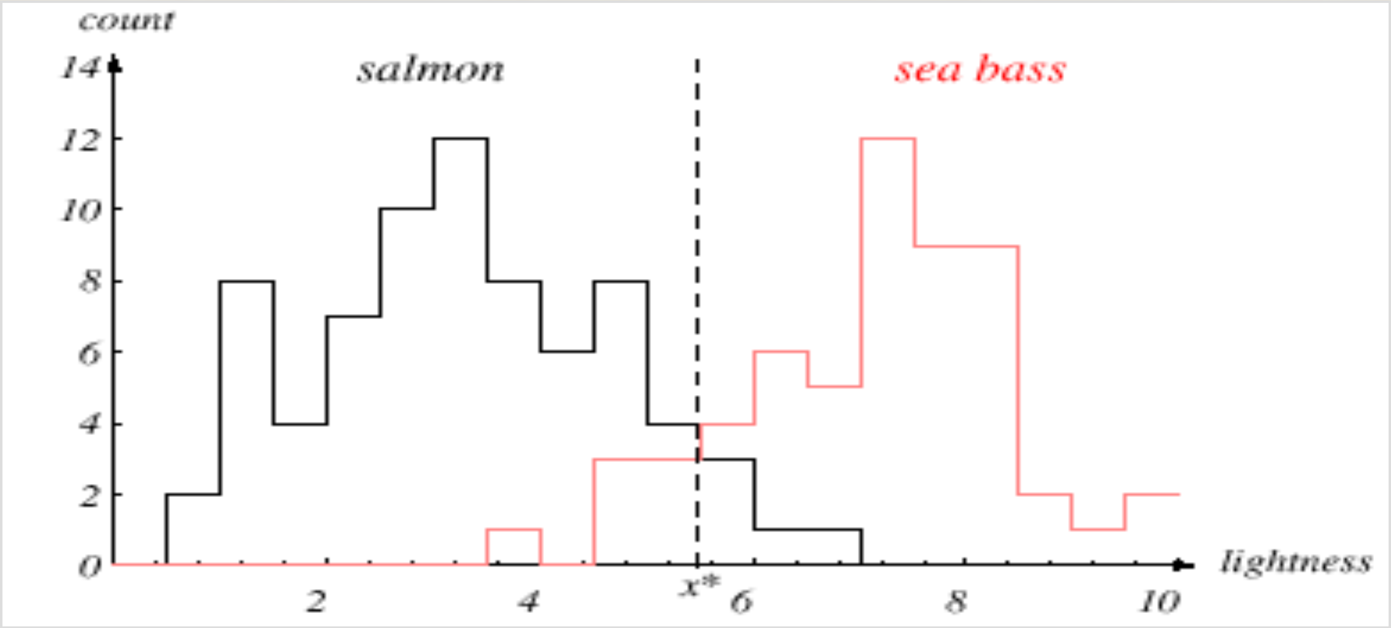
Feature Selection

The length is a poor feature alone!

Select the lightness as a possible feature

Another Feature

- **Lightness is a better feature than length because it reduces the misclassification error.**
- **Can we combine features in such a way that we improve performance? (Hint: correlation)**



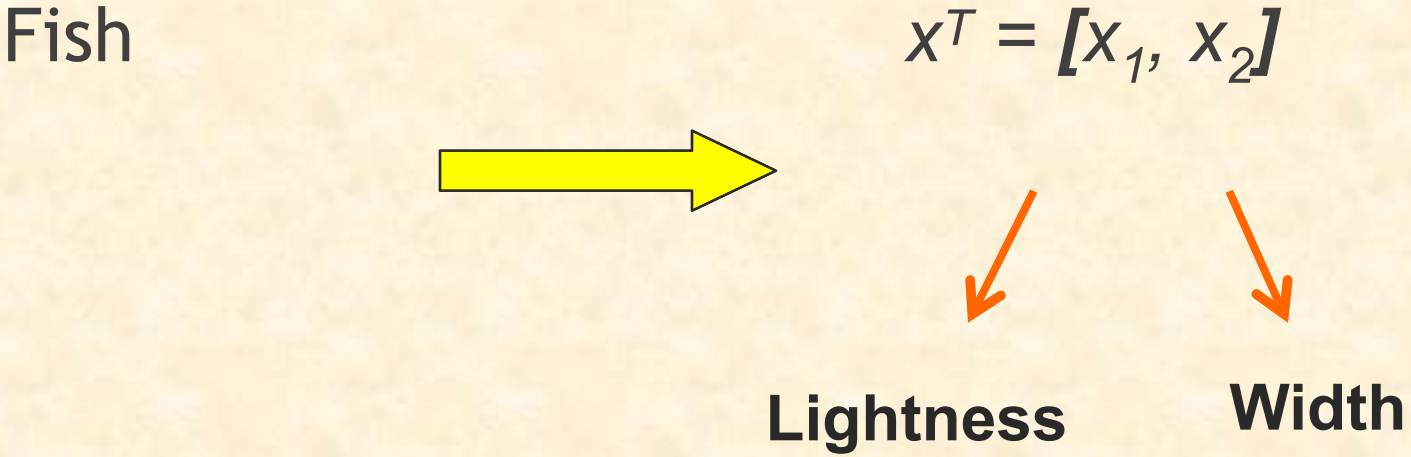
Threshold Decision Boundary and Cost Relationship

- Move 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

Feature Vector

- Adopt the lightness and add the width of the fish to the feature vector



Fish

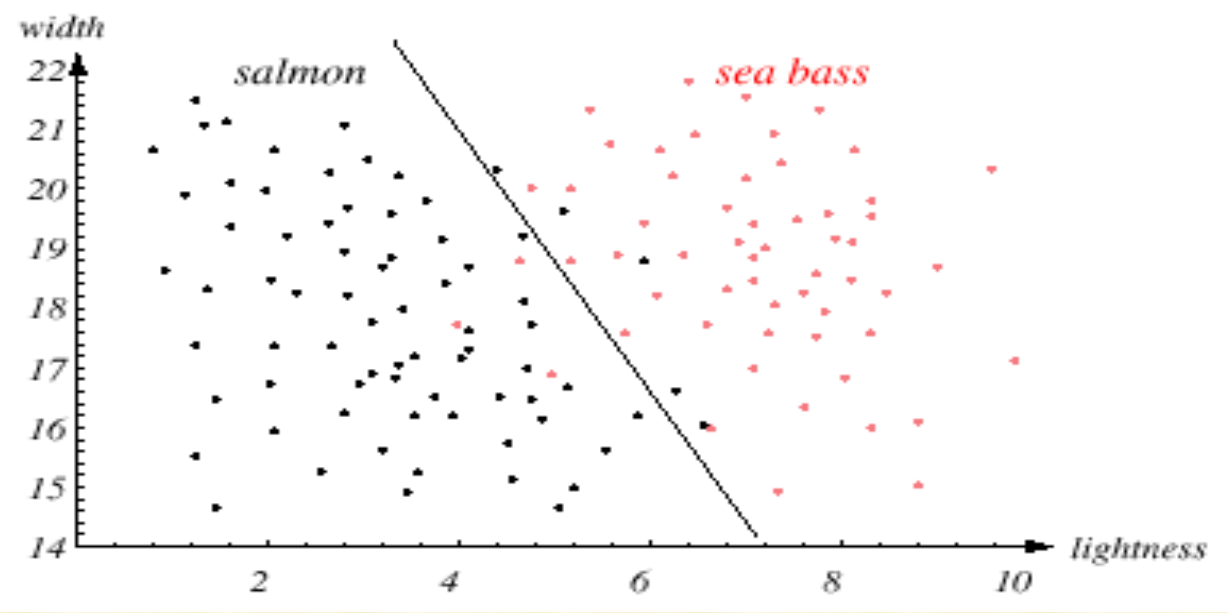
$x^T = [x_1, x_2]$

Lightness

Width

Width and Lightness Boundary

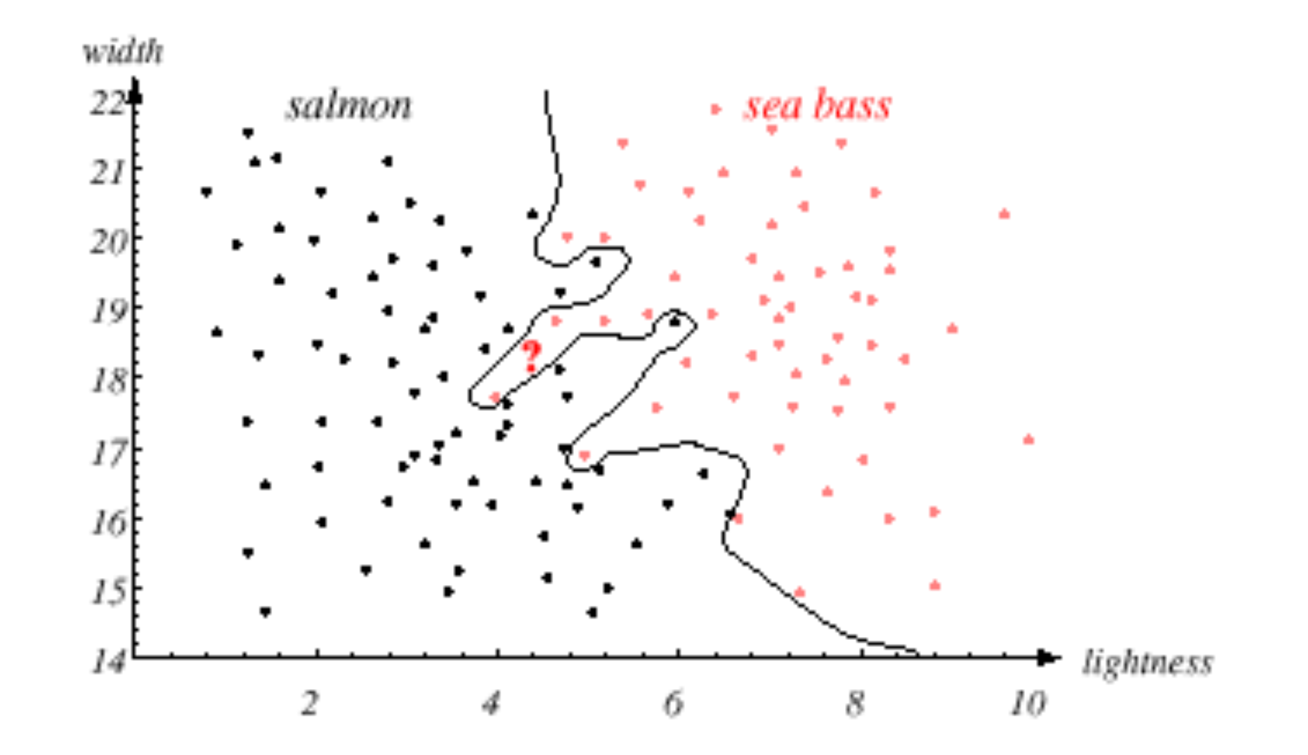
- Treat features as a N-tuple (two-dimensional vector)
- Create a scatter plot
- Draw a line (regression) separating the two classes



Features

- We might add other features that are not highly correlated with the ones we already have. Be sure not to reduce the performance by adding “noisy features”
- Ideally, you might think the best decision boundary is the one that provides optimal performance on the training data (see the following figure)

Generalization Problem



Is this a good decision boundary?

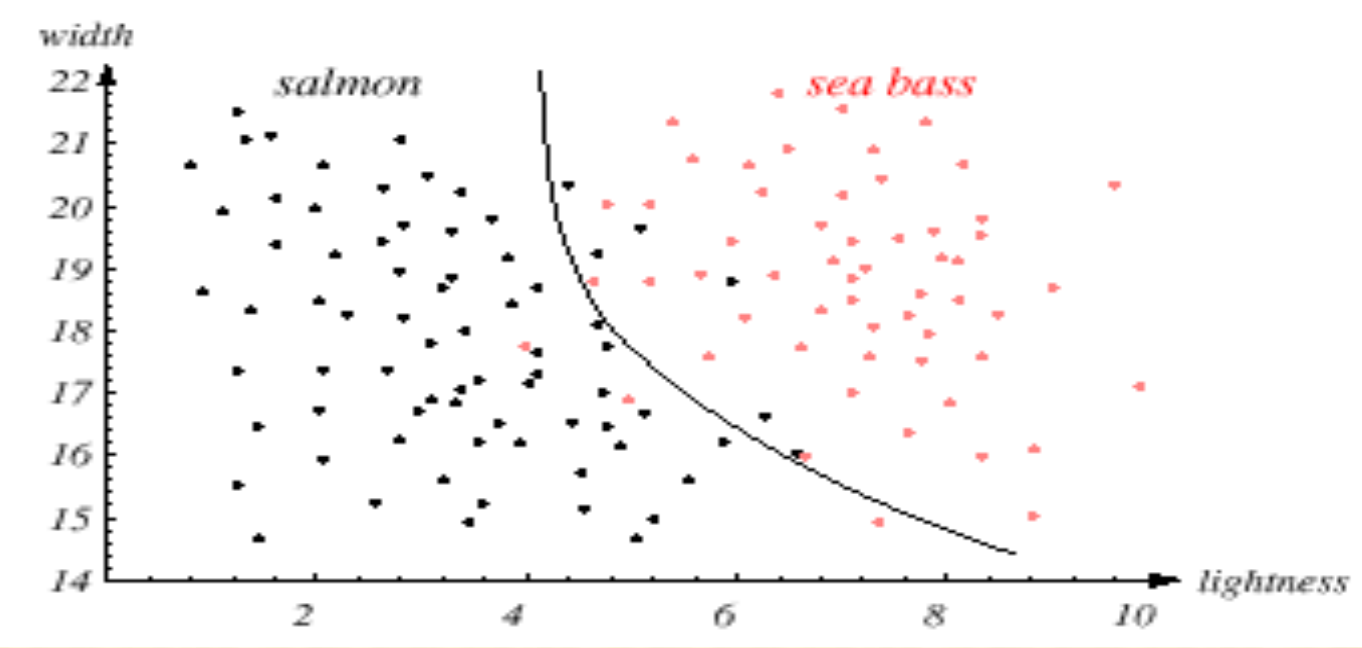
Decision Boundary Choice

- Our satisfaction is premature because the central aim of designing a classifier is to correctly classify new (test) input

Issue of generalization!

Generalization & Risk: Better Decision Boundary

- **Why might a smoother decision surface be a better choice? (hint: Occam's Razor).**
- **PR investigates how to find such "optimal" decision surfaces and how to provide system designers with the tools to make intelligent trade-offs.**



Need for Probabilistic Reasoning

- Most everyday reasoning is based on uncertain evidence and inferences.
- Classical logic, which only allows conclusions to be strictly true or strictly false, does not account for this uncertainty or the need to weigh and combine conflicting evidence.
- Todays expert systems employed fairly *ad hoc* methods for reasoning under uncertainty and for combining evidence.

Probabilistic Decision Theory

- **Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification.**
- **Using *probabilistic* approach to help making decision (e.g., classification) so as to *minimize the risk* (cost).**
- **Assume all relevant probability distributions are known (later we will learn how to estimate these from data).**

Prior Probability

- State of nature is *prior* information
  - ω denote the state of nature
- Model as a random variable, ω:
  - $\omega = \omega_1$: the event that the next fish is a sea bass
  - category 1: sea bass; category 2: salmon
- A priori probabilities:
  - $P(\omega_1)$ = probability of category 1
  - $P(\omega_2)$ = probabi
  - $P(\omega_1) + P(\omega_2)$ = 1 (eithe $_2$ must occur)
- Decision rule

  Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise, decide ω_2

But we know there will be many mistakes ….

http://www.stat.yale.edu/Courses/1997-98/101/ranvar.htm

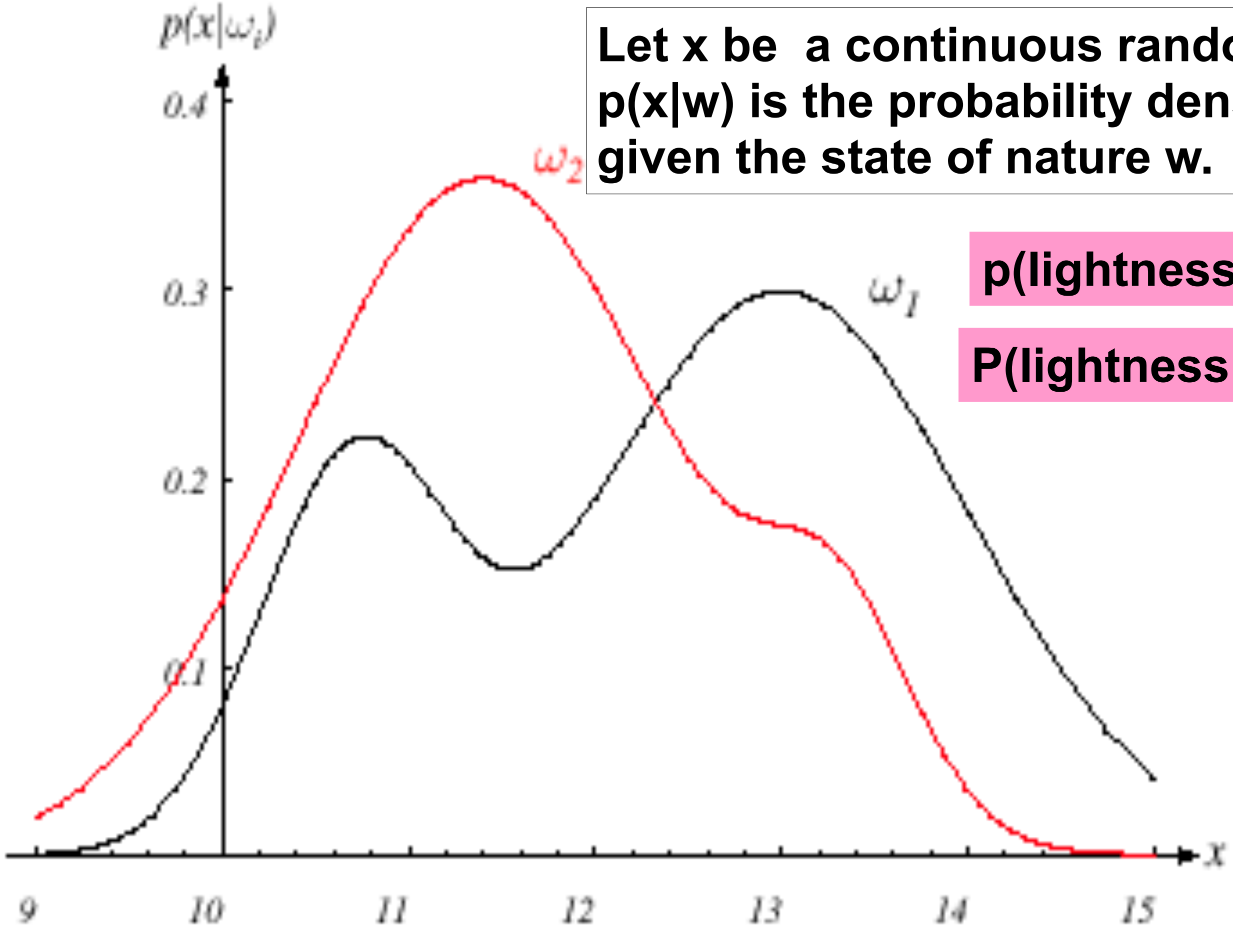
Class Conditional Probabilities

- A decision rule with only prior information always produces the same result and ignores measurements.
- If $P(\omega_1) >> P(\omega_2)$, we will be correct most of the time.

- Given a feature, x (lightness), which is a continuous random variable, $p(x|\omega_2)$ is the class-conditional probability density function:

- $p(x|\omega_1)$ and $p(x|\omega_2)$ describe the difference in lightness between populations of sea and salmon.

Conditional Probability



Let x be a continuous random variable. p(x|w) is the probability density for x given the state of nature w.

p(lightness | salmon) ?

P(lightness | sea bass) ?

Preliminaries and Notations

$\omega_i \in \{\omega_1, \omega_2, \boxed{?}, \omega_c\}$: a state of nature

$P(\omega_i)$: prior probability

$\mathbf{x}$: feature vector

$p(\mathbf{x} | \omega_i)$: class-conditional density

$P(\omega_i | \mathbf{x})$: posterior probability

Bayes Formula: Combining A prioiri and Conditional Probabilities

- Suppose we know both $P(\omega_j)$ and $p(x|\omega_j)$, and we can measure x. How does this influence our decision?
- The joint probability that of finding a pattern that is in category j and that this pattern has a feature value of x is:

$$p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$$

- Rearranging terms, we arrive at Bayes formula.

Casual Formulation

- The *prior* probability reflects knowledge of the relative frequency of instances of a class
- The *likelihood* is a measure of the probability that a measurement value occurs in a class
- The *evidence* is a scaling term

Posterior Probability

- Bayes formula:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

For two categories:

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j).$$

can be expressed in words as:

$$posterior = \frac{likelihood \times prior}{evidence}$$

- By measuring x, we can convert the prior probability, $P(\omega_j)$, into a posterior probability, $P(\omega_j|x)$.

Bayes Decision:
Choose w1 if P(w1|x) > P(w2|x); otherwise choose w2.

- Evidence can be viewed as a scale factor and is often ignored in optimization applications (e.g., speech recognition).

Two Categories

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2

Special cases:

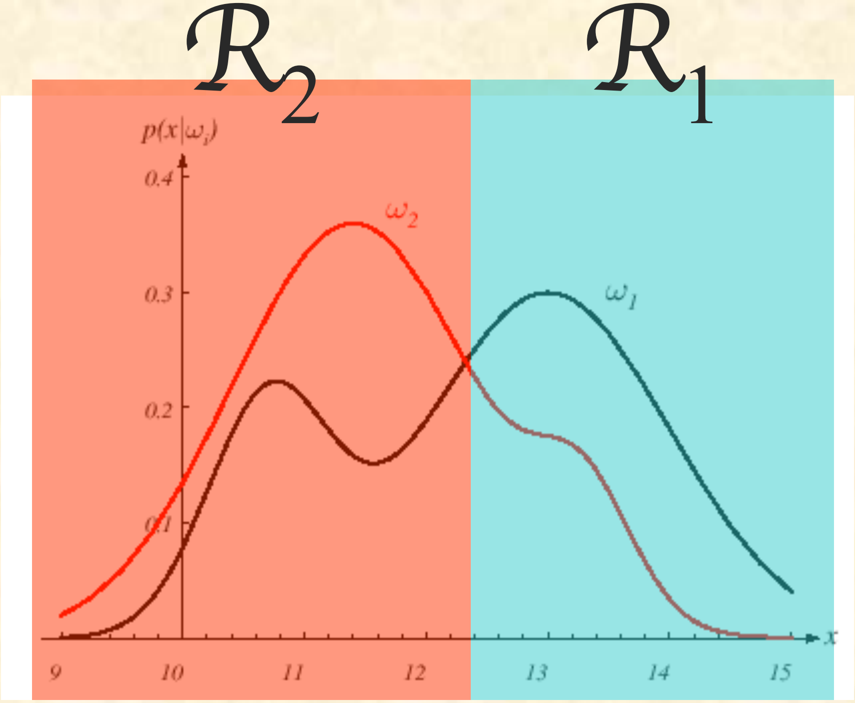
1. $P(\omega_1)=P(\omega_2)$

   Decide ω_1 if $p(x|\omega_1) > p(x|\omega_2)$; otherwise decide ω_2

2. $p(x|\omega_1)=p(x|\omega_2)$

   Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2

Example



Special cases:

1. $P(\omega_1)=P(\omega_2)$

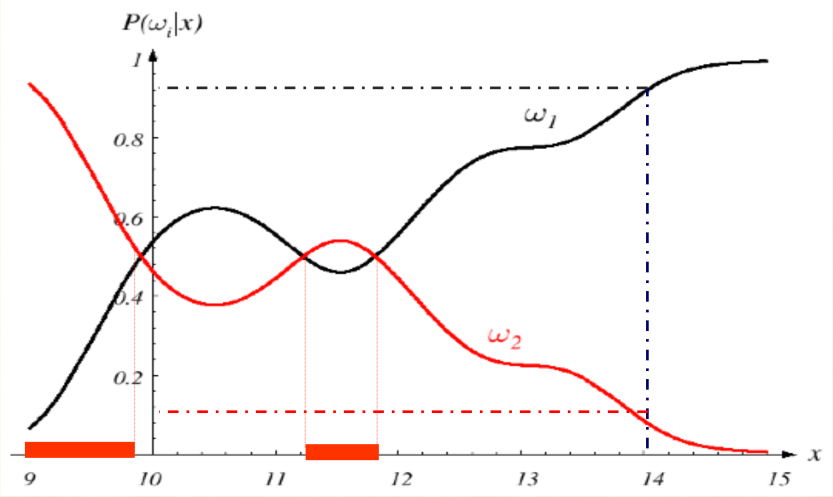
   Decide ω_1 if $p(\mathbf{x}|\omega > p(\mathbf{x}|\omega_2)$; otherwise decide ω_1

2. $p(\mathbf{x}|\omega_1)=p(\mathbf{x}|\omega_2)$

   Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2

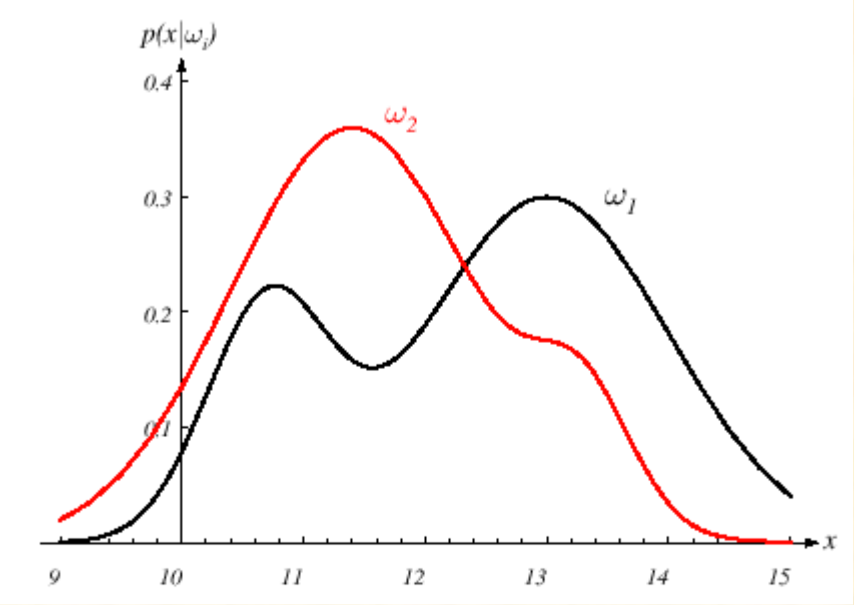
$$P(\omega_1)=P(\omega_2)$$

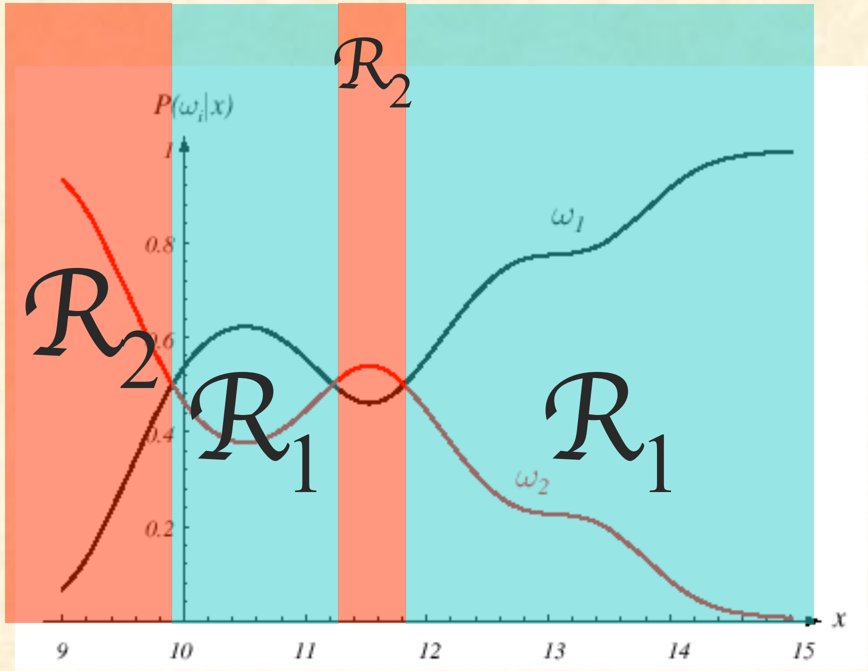
Example



$P(\omega_1)$=2/3

$P(\omega_2)$=1/3

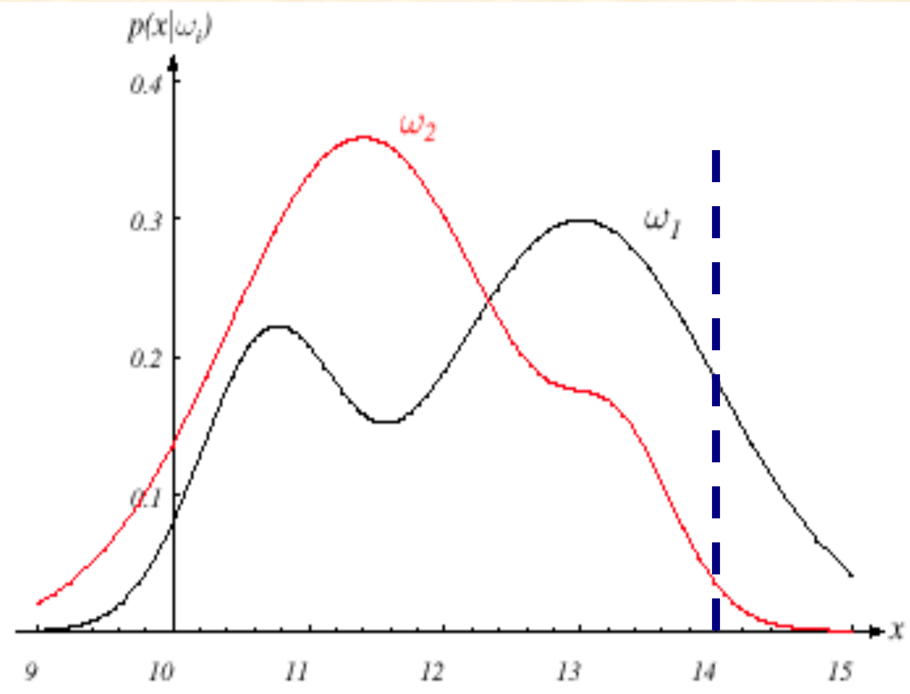


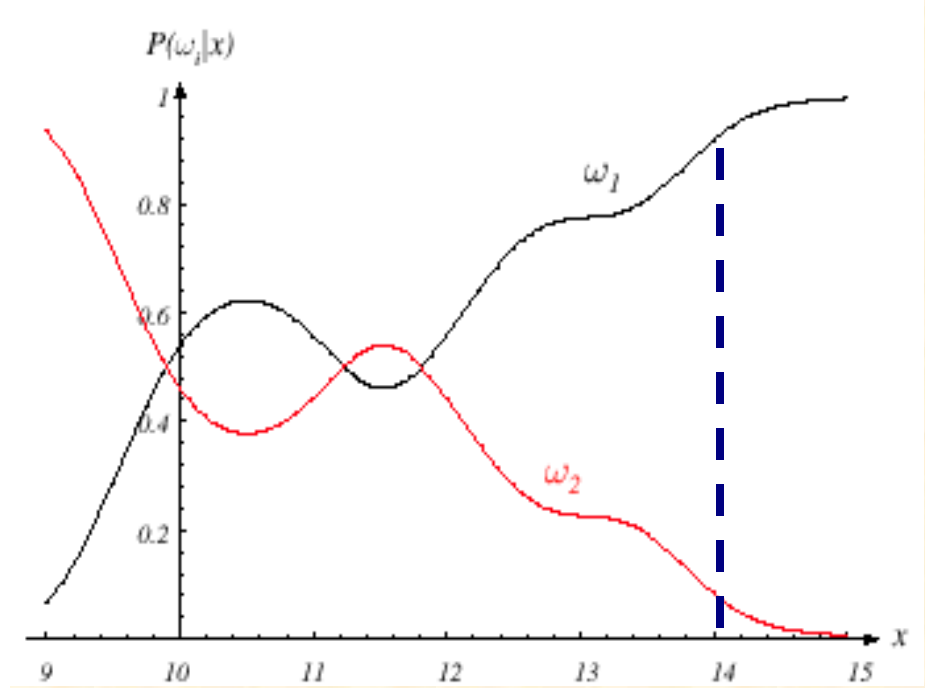


Bayes Decision Rule

Decide ω_1 if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$; otherwise decide ω_2

Posterior Probability





- **Two-class fish sorting problem (P(ω1) = 2/3, P(ω2) = 1/3):**

- **For every value of x, the posteriors sum to 1.0.**

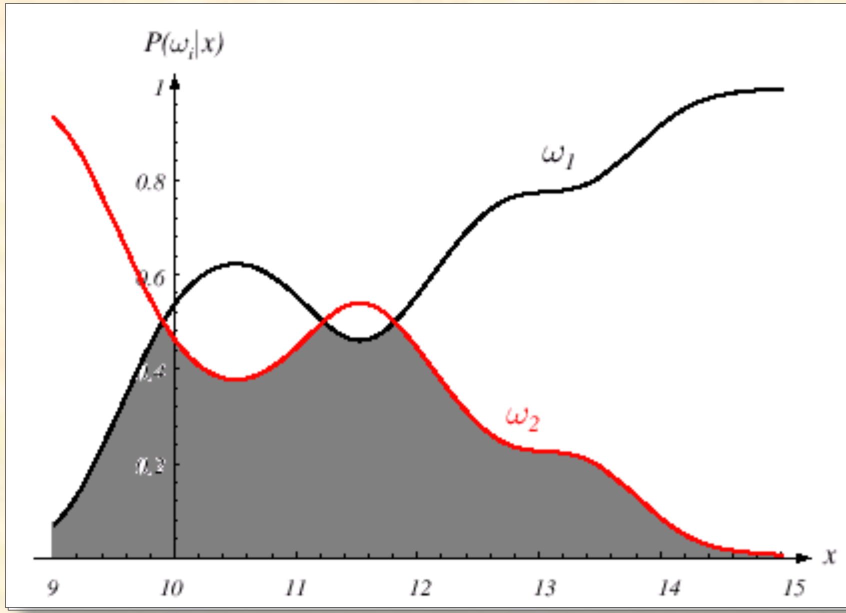
- **At x=14, the probability it is in category ω2 is 0.08, and for category ω1 is 0.92.**

Classification Error

- Decision rule:
  - For an observation x, decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise, decide ω_2
- Probability of error:

$$P(error\,|\,x) = \begin{cases} P(\omega_2|x) & x \in \omega_1 \\ P(\omega_1|x) & x \in \omega_2 \end{cases}$$

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error\,|\,x)p(x)dx$$



- The average probability of error is given by:

Consider two categories: $P(error\,|\,x) = min[\,P(\omega_1|x), P(\omega_2|x)]$

- If for every x we ensure that P(error|x) is as small as possible, then the integral is as small as possible. Thus, Bayes decision rule for minimizes P(error).

Generalization of Two-Class Problem

- **Generalization of the preceding ideas:**
  - **Use of more than one feature (e.g., length and lightness)**
  - **Use more than two states of nature (e.g., N-way classification)**
  - **Allowing actions other than a decision to decide on the state of nature (e.g., rejection: refusing to take an action when alternatives are close or confidence is low)**
  - **Introduce a loss of function which is more general than the probability of error (e.g., errors are not equally costly)**
  - **Let us replace the scalar** x **by the vector** $\mathbf{x}$ **in a** ***d*****-dimensional Euclidean space,** R^d**, called the *feature space*.**

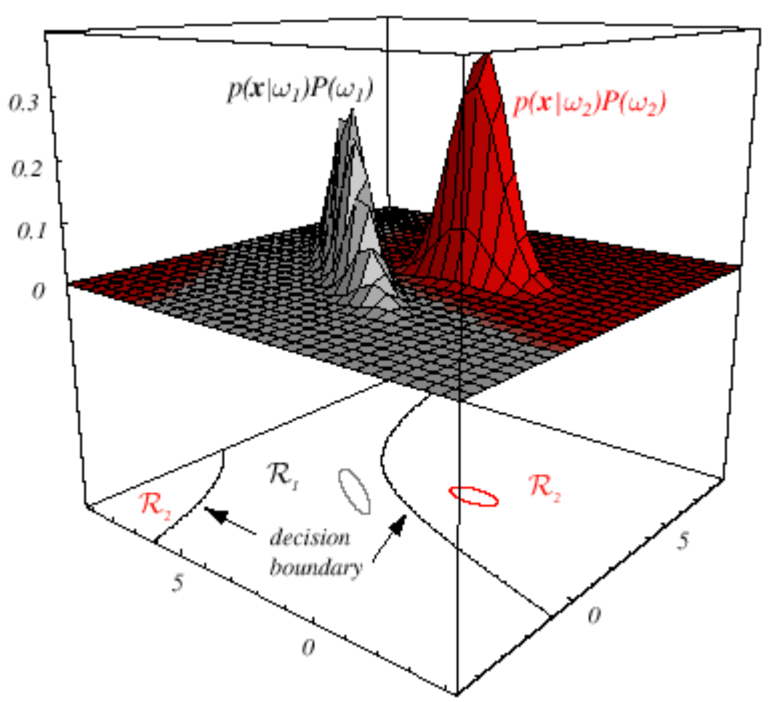
Decision Regions

The net effect is to divide the feature space into c regions (one for each class). We then have c *decision regions* separated by *decision boundaries*.

$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) > g_j(\mathbf{x}) \;\; \forall j \neq i\}$$

Two-category example

Decision regions are separated by *decision boundaries*.



Figur

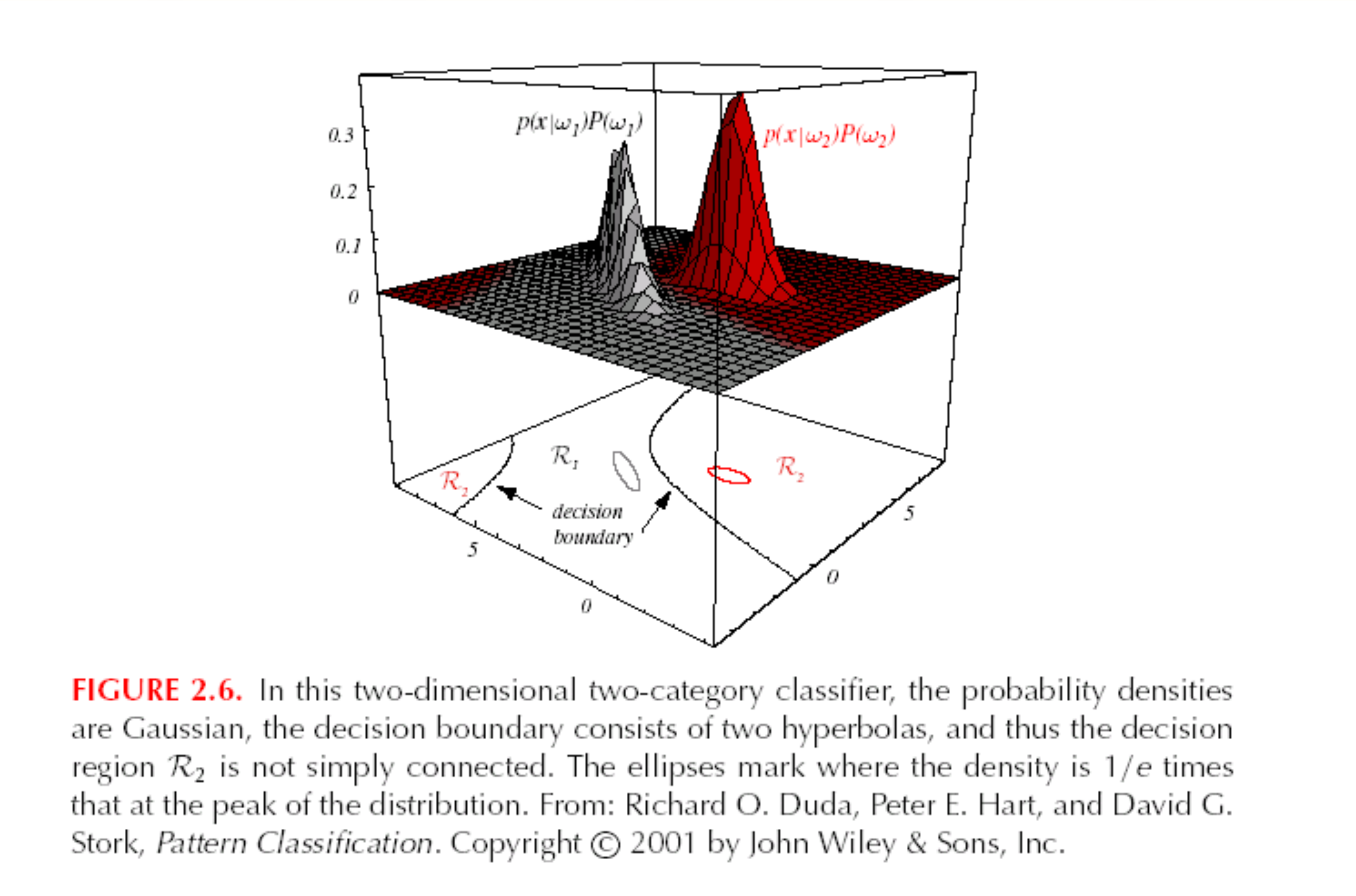


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayesian Decision Theory (Classification)

The Normal Distribution

Basics of Probability

Discrete random variable (*X*) - **Assume integer**

Probability mass function (pmf): $p(x) = P(X = x)$

Cumulative distribution function (cdf): $F(x) = P(X \le x) = \sum_{t=-\infty}^{x} p(t)$

Continuous random variable (*X*)

Probability density function (pdf): $p(x)$ or $f(x)$ not a probability

Cumulative distribution function (cdf): $F(x) = P(X \le x) = \int_{-\infty}^{x} p(t)dt$

Cairo University

Thank You ...