UNIVERSITY of HOUSTON
DEPARTMENT OF COMPUTER SCIENCE

Machine Learning

Rakesh Verma

Basic Technical Concepts in Machine Learning

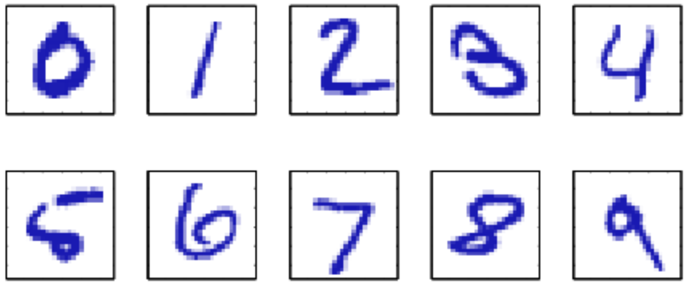
- Introduction
- Supervised learning
- Problems in supervised learning
- Bayesian decision theory

Ch1. Pattern Recognition & Machine Learning

- Automatically discover regularities in data.

- Based on regularities classify data into categories

Example:



Pattern Recognition & Machine Learning

- Supervised Learning:
  - ✓ Classification
  - ✓ Regression
- Unsupervised Learning:
  - ✓ Clustering
  - ✓ Density Estimation
  - ✓ Dimensionality reduction

Basic Technical Concepts in Machine Learning

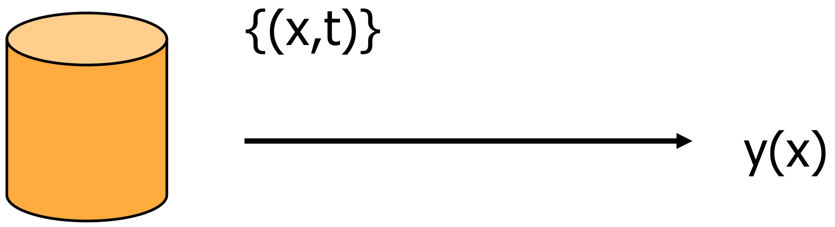
- Introduction
- Supervised learning
- Problems in supervised learning
- Bayesian decision theory

Supervised Learning

Training set x = {x1, x2, …, xN}

Class or target vector t = {t1, t2, …, tN}

Find a function y(x) that takes a vector x and outputs a class t.



Supervised Learning

Medical example:

X = {patient1, patient2, …. patientN}

patient1 = (high pressure, normal temp., high glucose,…)

t1 = cancer

patient2 = (low pressure, normal temp., normal glucose,…)

t1 = not cancer

new patient = (low pressure, low temp., normal glucose,…)

t = ?

Basic Technical Concepts in Machine Learning

- Introduction
- Supervised learning
- Problems in supervised learning
- Bayesian decision theory

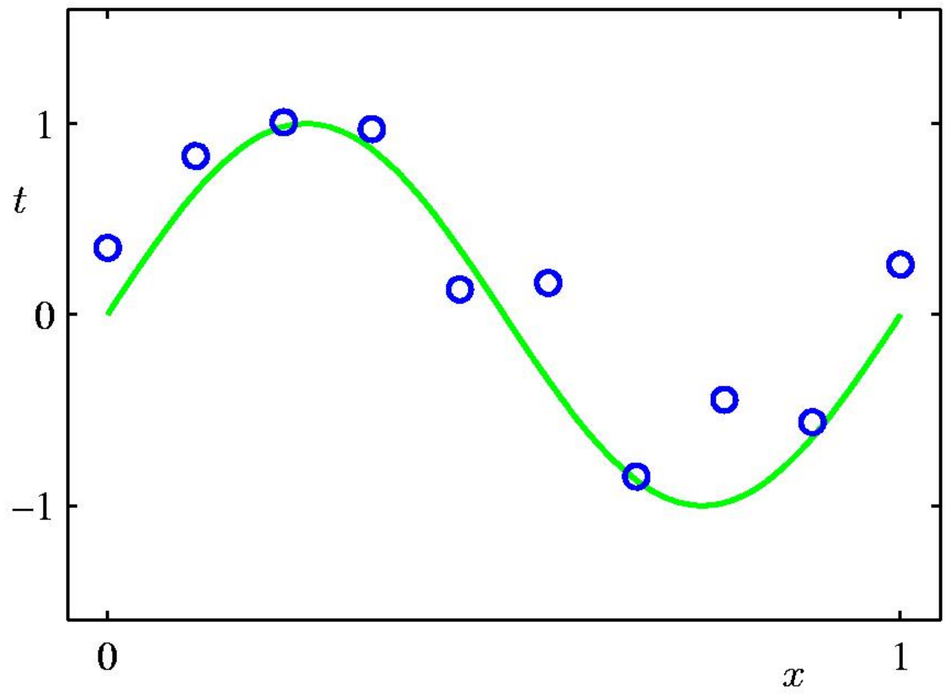
Problems in supervised learning

- Two problems in supervised learning:

1. Overfitting – Underfitting
2. Curse of Dimensionality

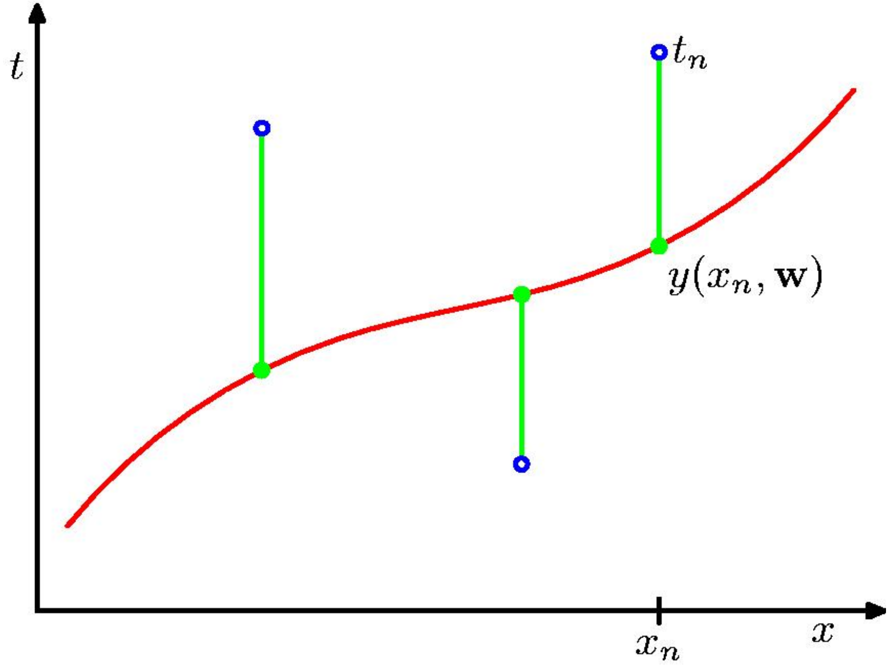
Problems in supervised learning

Example Polynomial Fitting (Regression):



Problems in supervised learning

Minimize Error:



Problems in supervised learning

- If our function is linear:

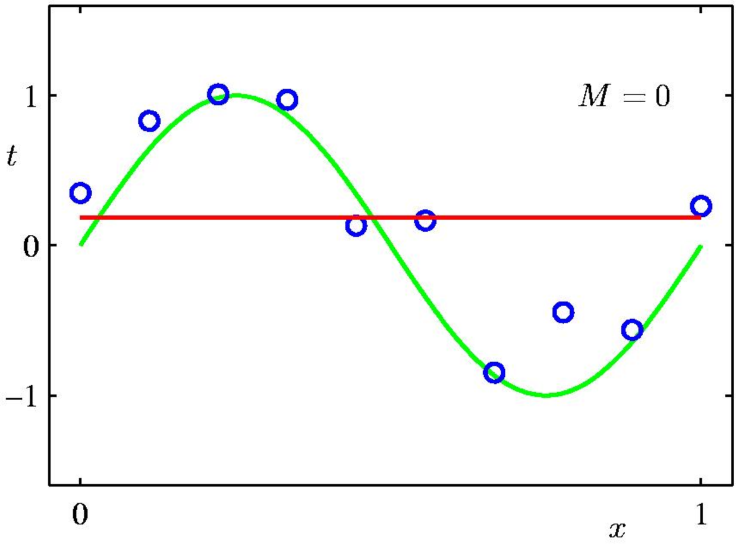
$y(x,w) = w0 + w1x + w2x^2 + \ldots + wMX^M$

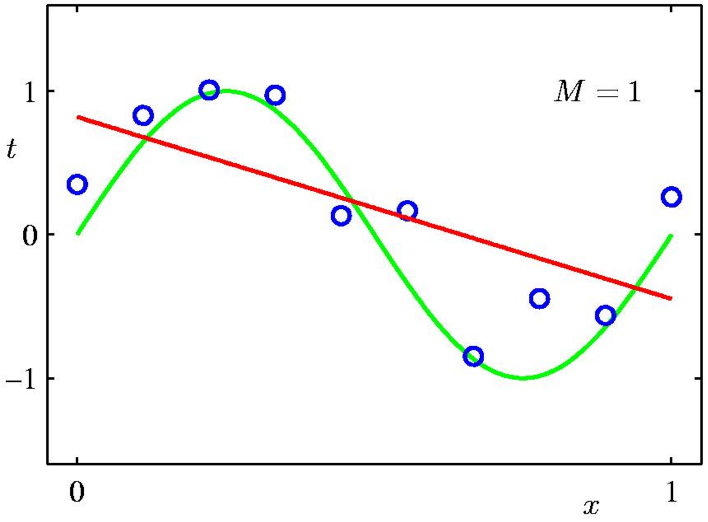
- Minimize error:

$E(w) = ½ \Sigma n \{y(xn,w) – tn)2$

- What happens as we vary M?

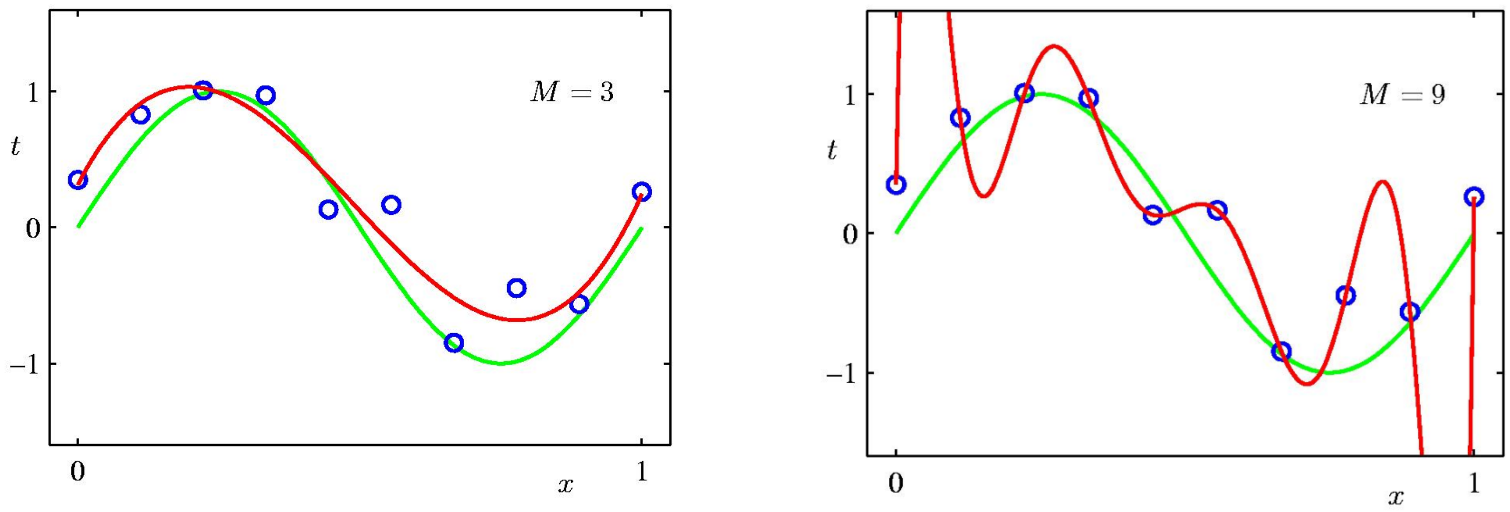
Underfitting





Underfitting

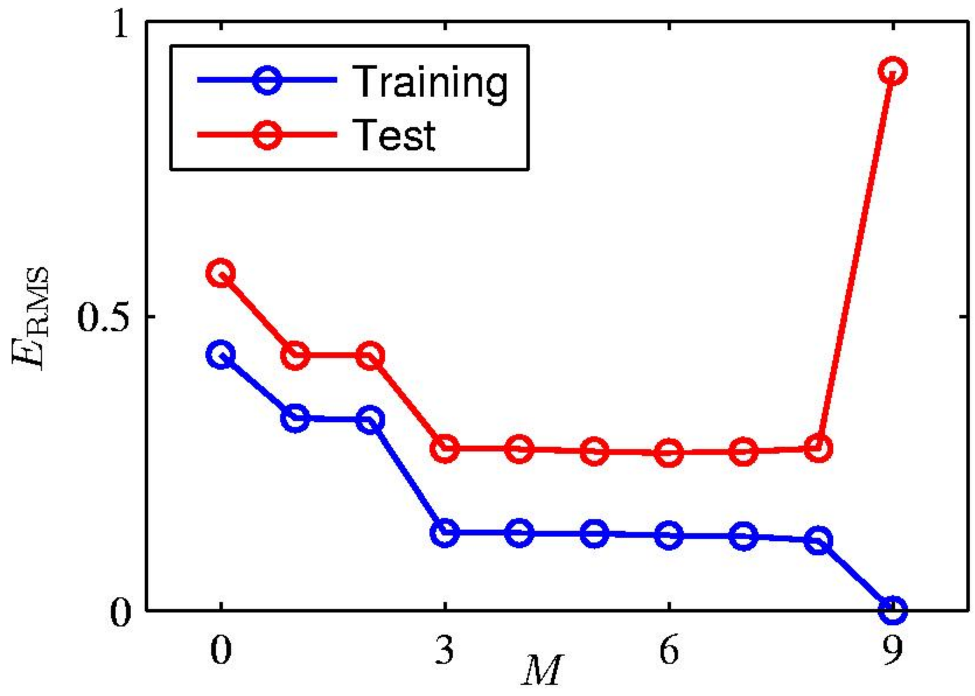
Overfitting



Overfitting

Overfitting

Root Mean Square Error (RSM)



Regularization

- One solution: Regularization
- Penalize for models that are too complex:
- Minimize error:

$E(w) = ½ \Sigma n \ \{y(xn,w) – tn)2 \ + \ \lambda/2 \ ||w||2$

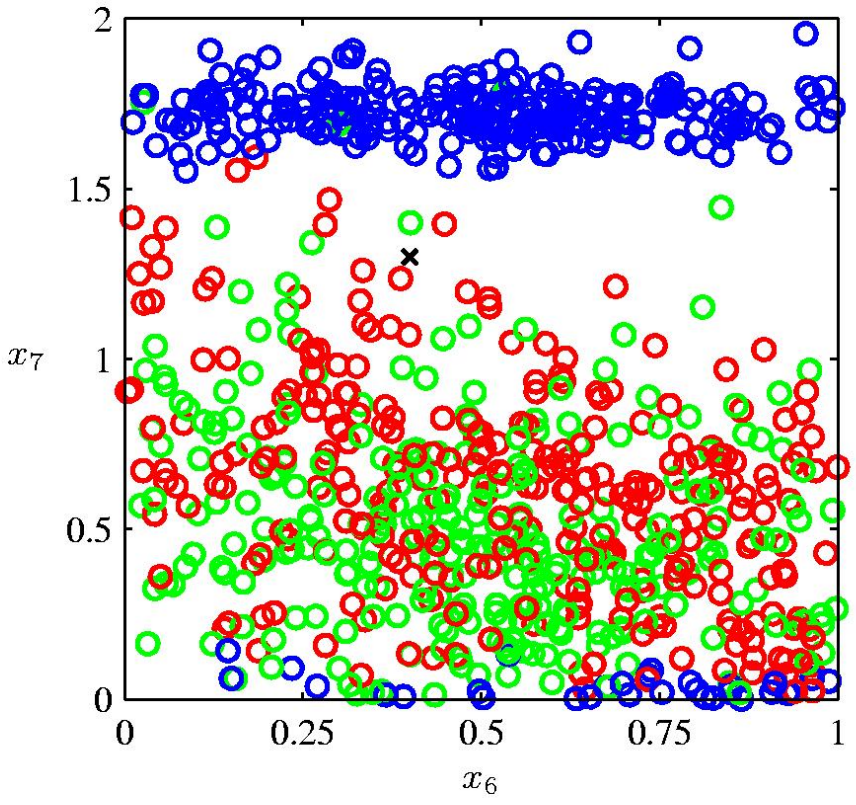
Problems in supervised learning

- Two problems in supervised learning:

1. Overfitting – Underfitting
2. Curse of Dimensionality

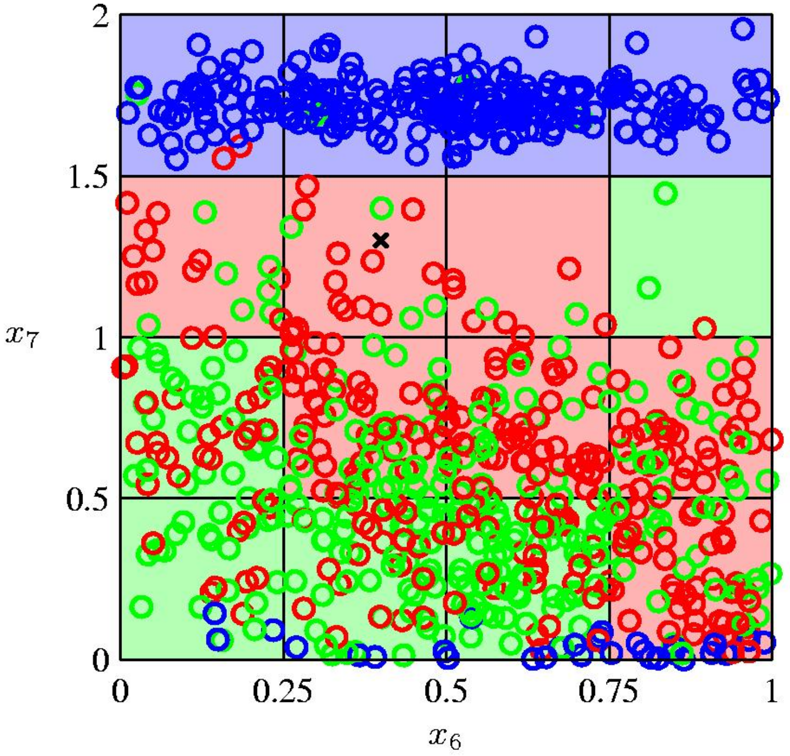
Curse of Dimensionality

Example: Classify vector x in one of 3 classes:



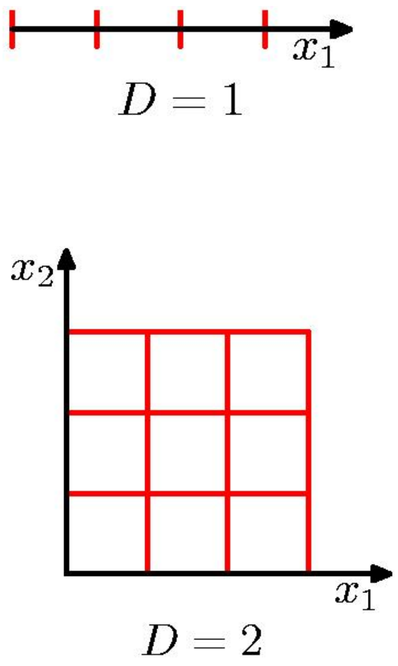
Curse of Dimensionality (cont.)

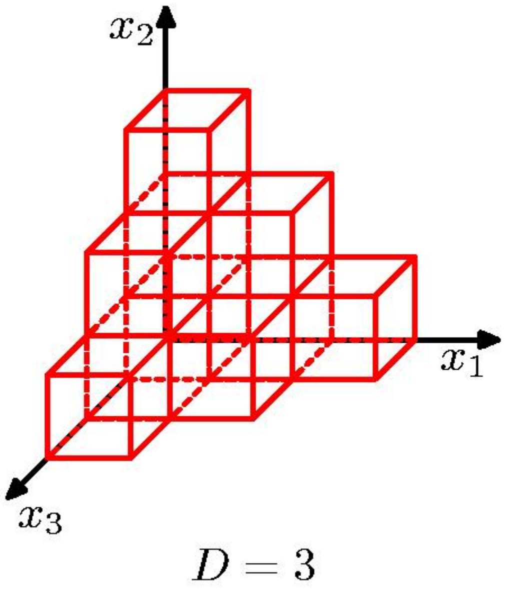
Solution: Divide space into cells:



Curse of Dimensionality (Cont.)

But if the no. of dimensions is high we need to take a huge amount of space to look at the ”neighborhood” of x.





Basic Technical Concepts in Machine Learning

- Introduction
- Supervised learning
- Problems in supervised learning
- Bayesian decision theory

Decision Theory

- **State of nature.**
  - ✓ Let C denote the state of nature. C is a random variable. (e.g., C = C1 for sea bass or C = C2 for salmon)
- **A priori probabilities.**
  - ✓ Let P(C1) and P(C2) denote the a priori probability of C1 and C2 respectively.
  - ✓ We know P(C1) + P(C2) = 1.
- **Decision rule.**
  - ✓ Decide C1 if P(C1) > P(C2); otherwise choose C2.

Basic Concepts

- **Class-conditional probability density function.**
  - ✓ Let x be a continuous random variable.
  - ✓ p(x|C) is the probability density for x given the state of nature C.
  - ✓ For example, what is the probability of lightness given that the class is salmon? p(lightness | salmon)?
  - ✓ Or what is the probability of lightness given sea bass? P(lightness | sea bass) ?

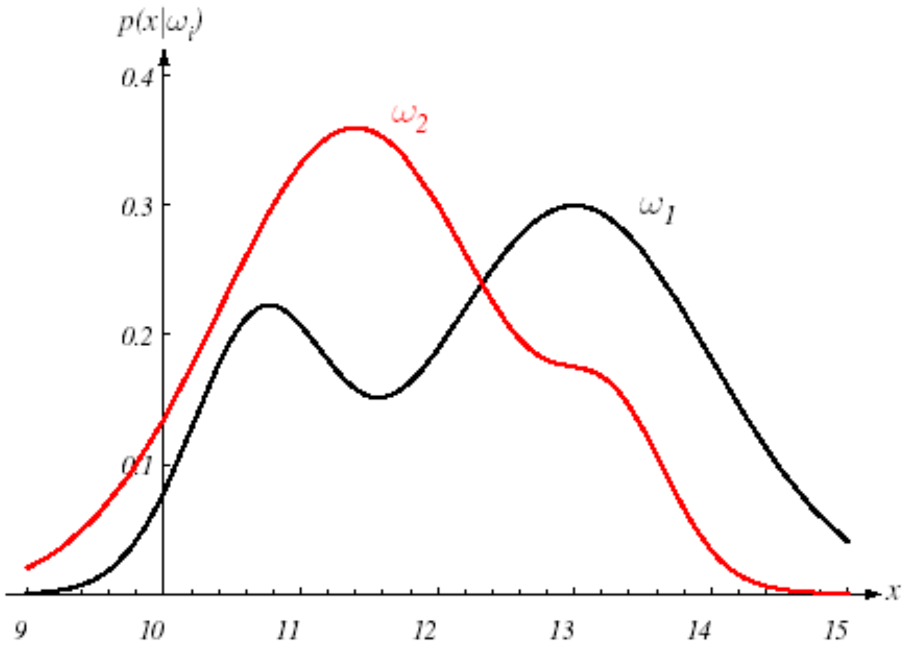


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Formula

How do we combine a priori and class-conditional Probabilities to know the probability of a state of nature?

Bayes Formula.

prior probability

$$\mathbf{P(Cj \mid x) = p(x|Cj)\, P(Cj) \;/\; p(x)}$$

posterior probability

likelihood

evidence

Bayes Decision:
Choose C1 if P(C1|x) > P(C2|x); otherwise choose C2.

Figure 2.2

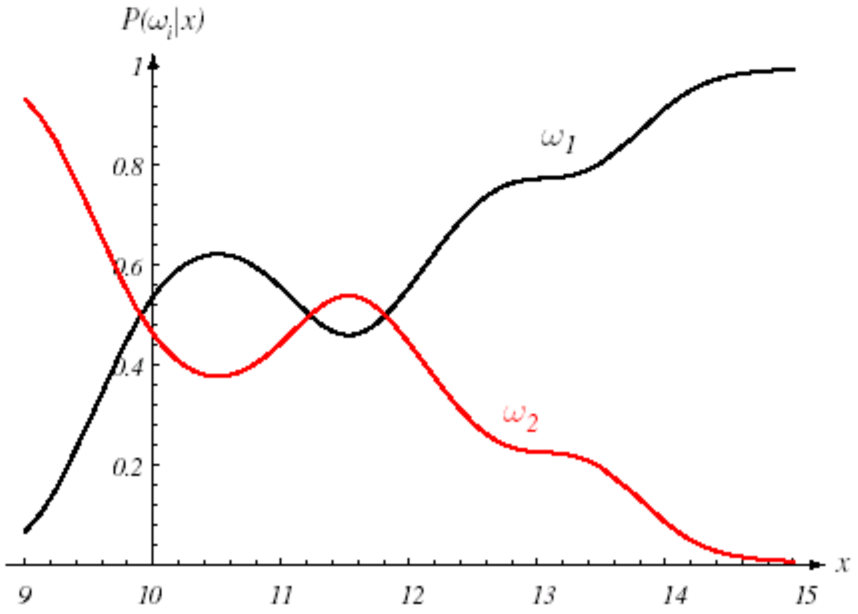


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Minimizing Error

What is the probability of error?

$P(\text{error} \mid x) =$ $P(C1|x)$ if we decide C2

$P(C2|x)$ if we decide C1

Does Bayes rule minimize the probability of error?

$P(\text{error}) = \int P(\text{error},x)\,dx = \int P(\text{error}|x)\,p(x)\,dx$

and if for every x we minimize the error then $P(\text{error}|x)$ is as small as it can be.

Answer is “yes”.

Simplifying Bayes Rule

Bayes Formula: $P(C_j \mid x) = p(x|C_j)\, P(C_j) \;/\; p(x)$

The evidence $p(x)$ is the same for all states or classes so

we can dispense with it to make a decision.

Rule:

Choose C1 if $p(x|C1)P(C1) > p(x|C2)P(C2)$;

otherwise decide C2

If $p(x|C1) = p(x|C2)$ then decision depends on the priors

If $P(C1) = P(C2)$ then decision depends on the likelihoods.

Loss Function

Let {C1,C2, …, Cc} be the possible states of nature.

Let {a1, a2, …, ak} be the possible actions.

Loss function:

$\lambda(ai|Cj)$ is the loss incurred for taking action ai when the state of nature is Cj.

Expected loss:

$$R(ai|x) = \Sigma_j \; \lambda(ai|Cj) \, P(Cj|x)$$

Decision: Select the action that minimizes the conditional risk

(** best possible performance **)

Zero-One Loss

We will normally be concerned with the symmetrical or zero-one loss function:

$$\lambda\,(a_i|C_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

In this case the conditional risk is:

$$R(a_i|x) = \Sigma_j\ \lambda(a_i|C_j)\ P(C_j|x)$$

$$= 1 - P(C_i|x)$$