Introduction to Statistical Pattern Recognition

Outline

- Human and Machine Perception
- What is pattern recognition?
- Sample pattern recognition problem
- Pattern recognition systems
- Bayes decision theory
- Matlab illustrations and sample programs
- Conclusion

Human Perception

Humans have developed highly sophisticated skills for sensing their environment and taking actions according to what they observe, e.g.,

- Recognizing a face
- Understanding spoken words
- Reading handwriting
- Distinguishing fresh food from its smell

We would like to give similar capabilities to machines.

What is pattern recognition?

A pattern is an entity, vaguely defined, that could be given a name, e.g.,

- Fingerprint image
- Handwritten word
- Human face
- Speech signal
- DNA sequence, etc.

Pattern recognition is the study of how machines can

- Observe the environment
- Learn to distinguish patterns of interest
- Make sound and reasonable decisions about the categories of patterns

Human and Machine Perception

- We are often influenced by the knowledge of how patterns are modeled and recognized in nature when we develop pattern recognition algorithms
- Research on machine perception also helps us gain deeper understanding and appreciation for pattern recognition systems in nature
- Yet, we may also apply many techniques that are purely numerical and do not have any correspondence in natural systems

Pattern Recognition Applications

Problem Domain	Application	Input Pattern	Pattern Classes
Document image analysis	Optical character recognition	Document image	Characters, words
Document classification	Internet search	Text document	Semantic categories
Multimedia database retrieval	Internet search	Video clip	Video genres
Speech recognition	Telephone directory assistance	Speech waveform	Spoken words
Natural language processing	Information extraction	Sentences	Parts of speech
Biometric recognition	Personal identification	Face, iris, fingerprint	Authorized users for access control
Medical	Diagnosis	Microscopic image	Cancerous/healthy cell
Military	Automatic target recognition	Optical or infrared image	Target type
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective/non-defective product
Industrial automation	Fruit sorting	Images taken on a conveyor belt	Grade of quality
Remote sensing	Forecasting crop yield	Multispectral image	Land use categories
Bioinformatics	Sequence analysis	DNA sequence	Known types of genes
Data mining	Searching for meaningful patterns	Points in multidimensional space	Compact and well-separated clusters

Sample Problem[1]

Problem: Sorting incoming fish on a conveyor belt, according to species

Assume we have only 2 kinds of fish:

- Sea bass
- Salmon

Figure: The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed.

[1]R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2000

Sample Problem: Decision Process

What kind of information can distinguish one species from the other?

- Length, width, weight, number and shape of fins, tail shape, etc.

What can cause problems during sensing (i.e., capturing image)?

- Lighting conditions, position of fish on the conveyor belt, camera noise, etc.

What are the steps in the process?

- Capture image $\rightarrow$ isolate fish $\rightarrow$ take measurements $\rightarrow$ make decision

Sample Problem: Selecting Features

- Assume a fisherman told us that a sea bass is generally longer than a salmon
- We can use length as a feature and decide between sea bass and salmon according to a threshold on length.
- How can we choose this threshold?

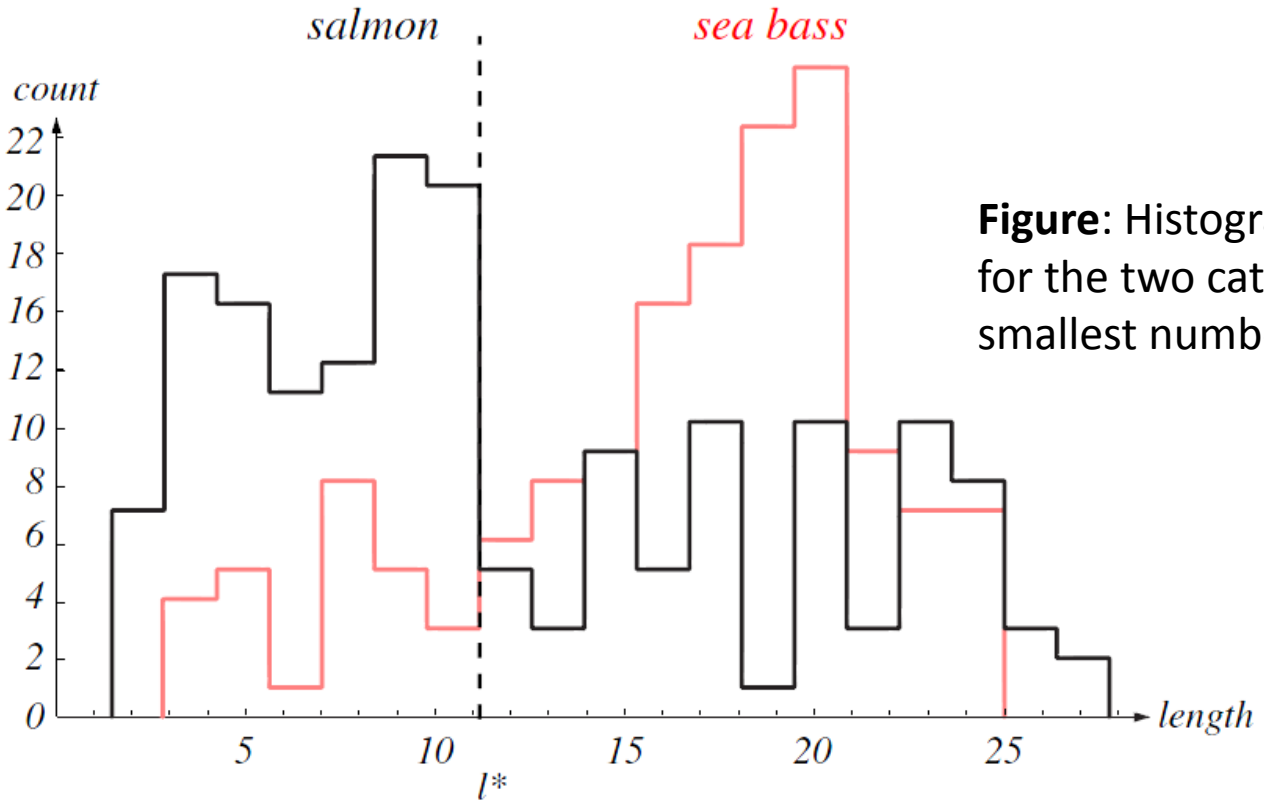


Figure: Histograms for the length feature for the two categories. l^* will lead to the smallest number of errors, on average.

Sample Problem: Selecting Features

- Even though sea bass is longer than salmon on average, there are many examples where this observation does not hold
- Try another feature: average lightness of the fish scales

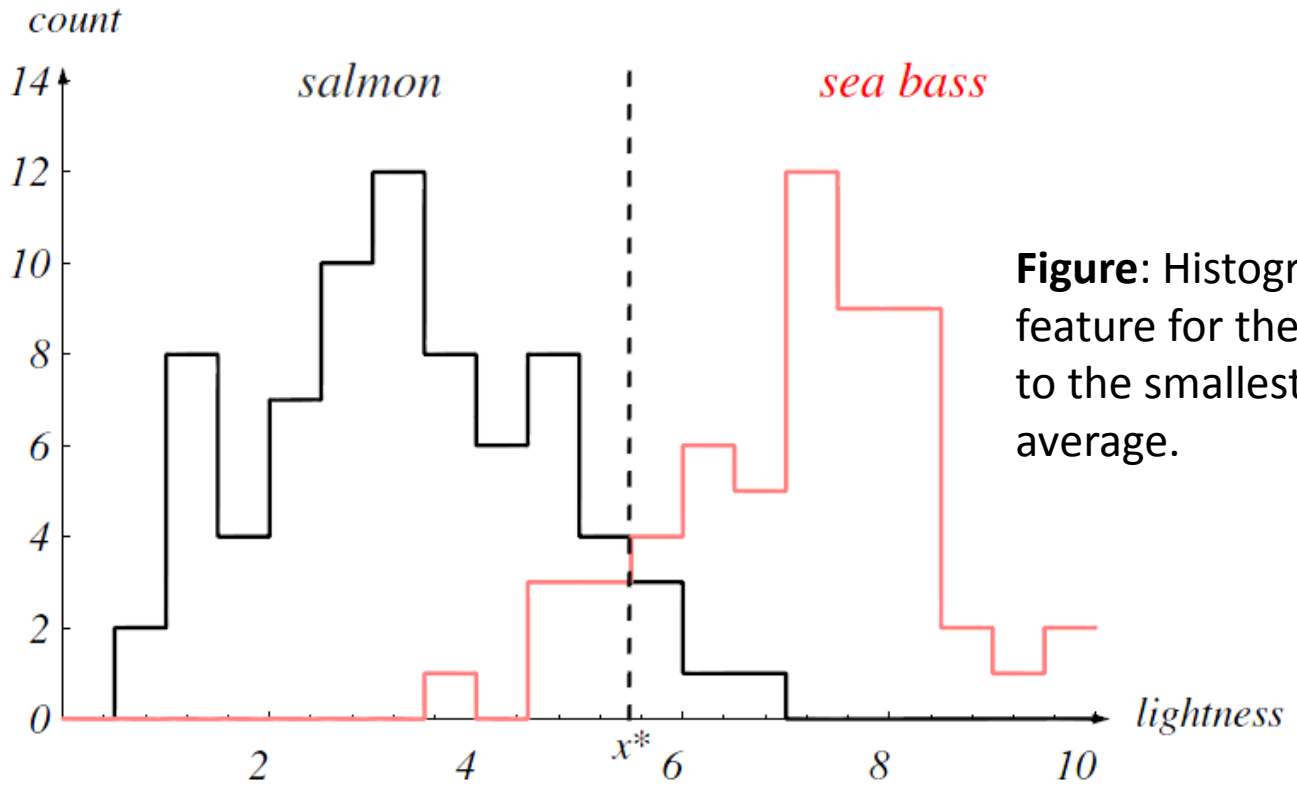


Figure: Histograms for the lightness feature for the two categories. x^* leads to the smallest number of errors, on average.

Sample Problem: Multiple Features

- Assume we also observed that sea bass are typically wider than salmon
- We can use two features in our decision: (a) lightness , x_1, and (b) width, x_2
- Each fish image is now represented as a point (feature vector) in a 2D space, i.e.,

$$\mathbf{x} = (x_1, x_2)^T$$

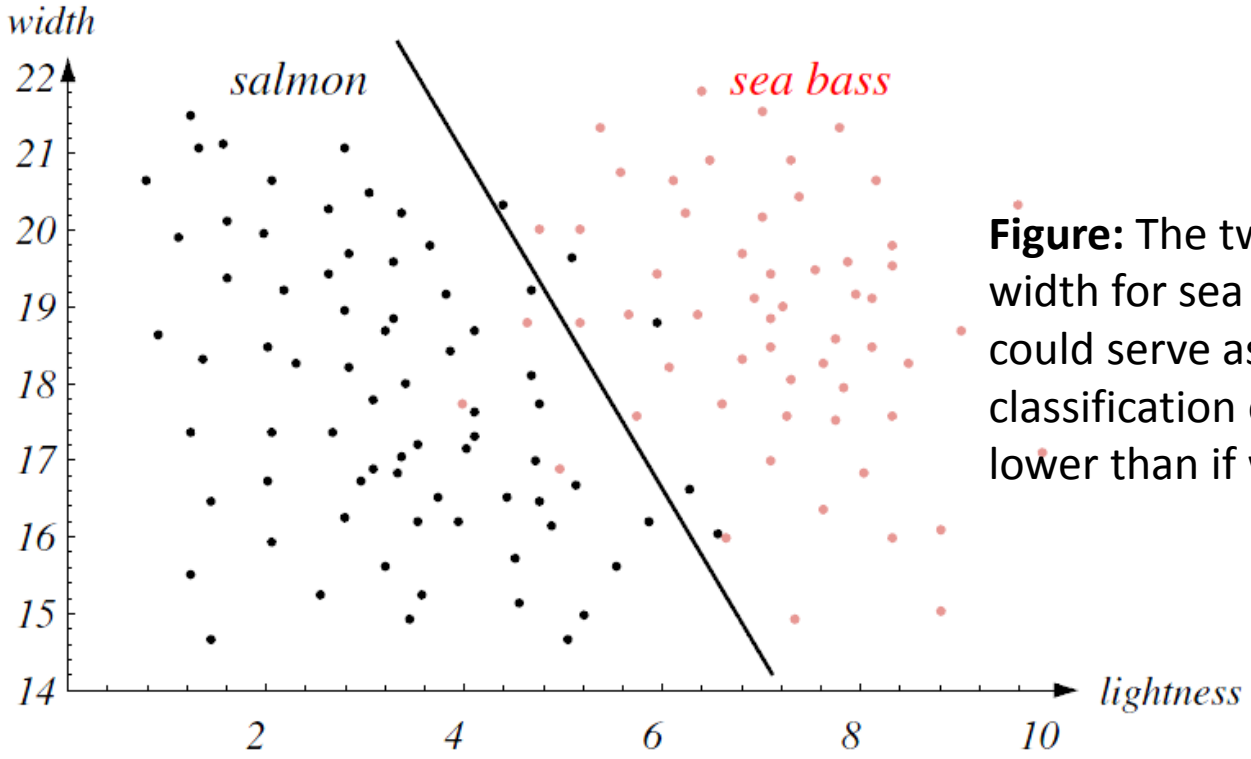


Figure: The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary. Overall classification error on the data shown is lower than if we use only one feature.

Pattern Recognition Systems

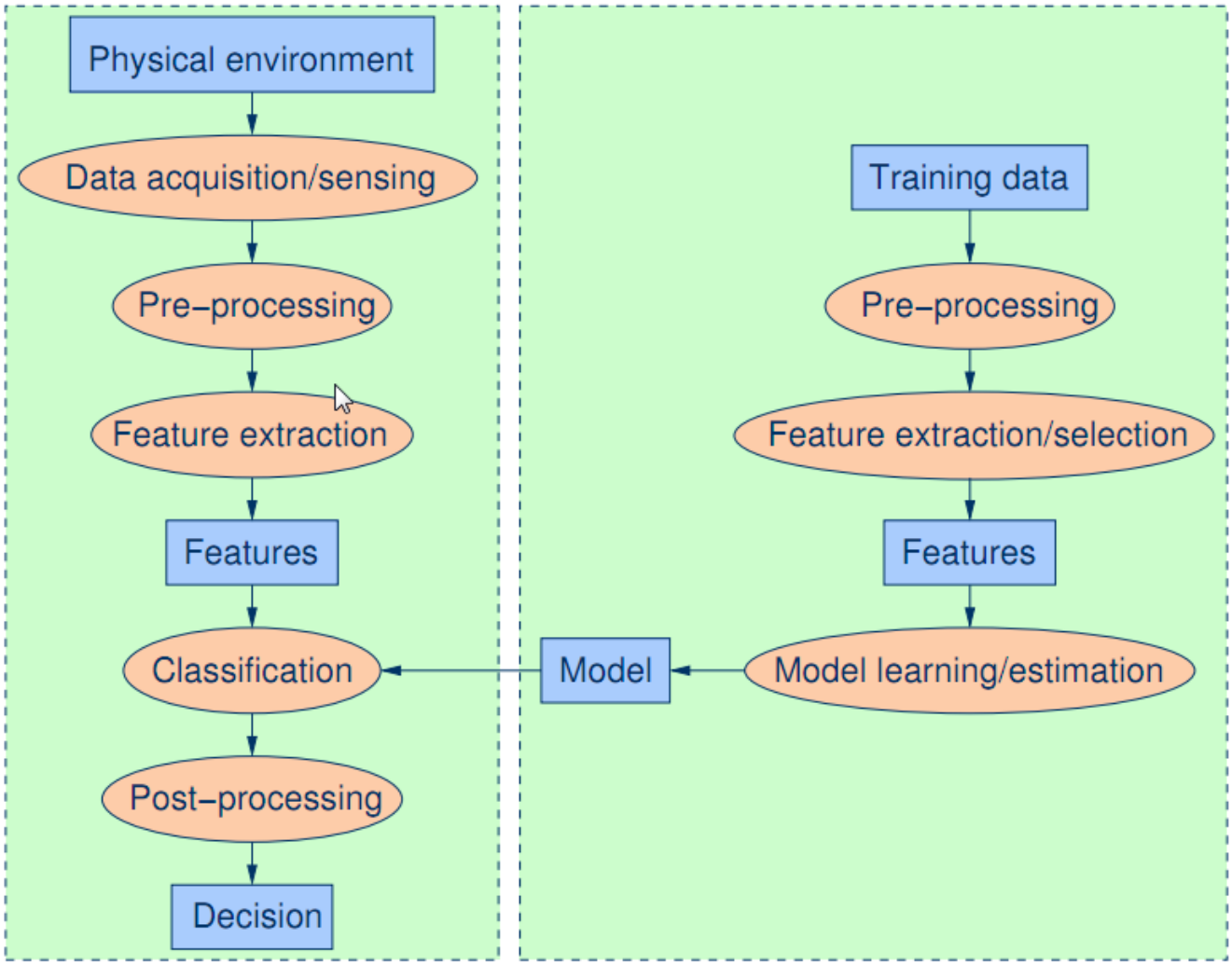


Figure: Object/process diagram of a pattern recognition system.

Pattern Recognition Systems

Data Acquisition and Sensing

- Measurements of physical variables
- Important issues: bandwidth, resolution, distortion, SNR, latency, etc.

Preprocessing

- Removal of noise in data
- Isolation of patterns of interest from the background

Feature Extraction

- Finding a new representation in terms of new features

Model Learning and Estimation

- Learning a mapping between features and pattern groups and categories

Pattern Recognition Systems

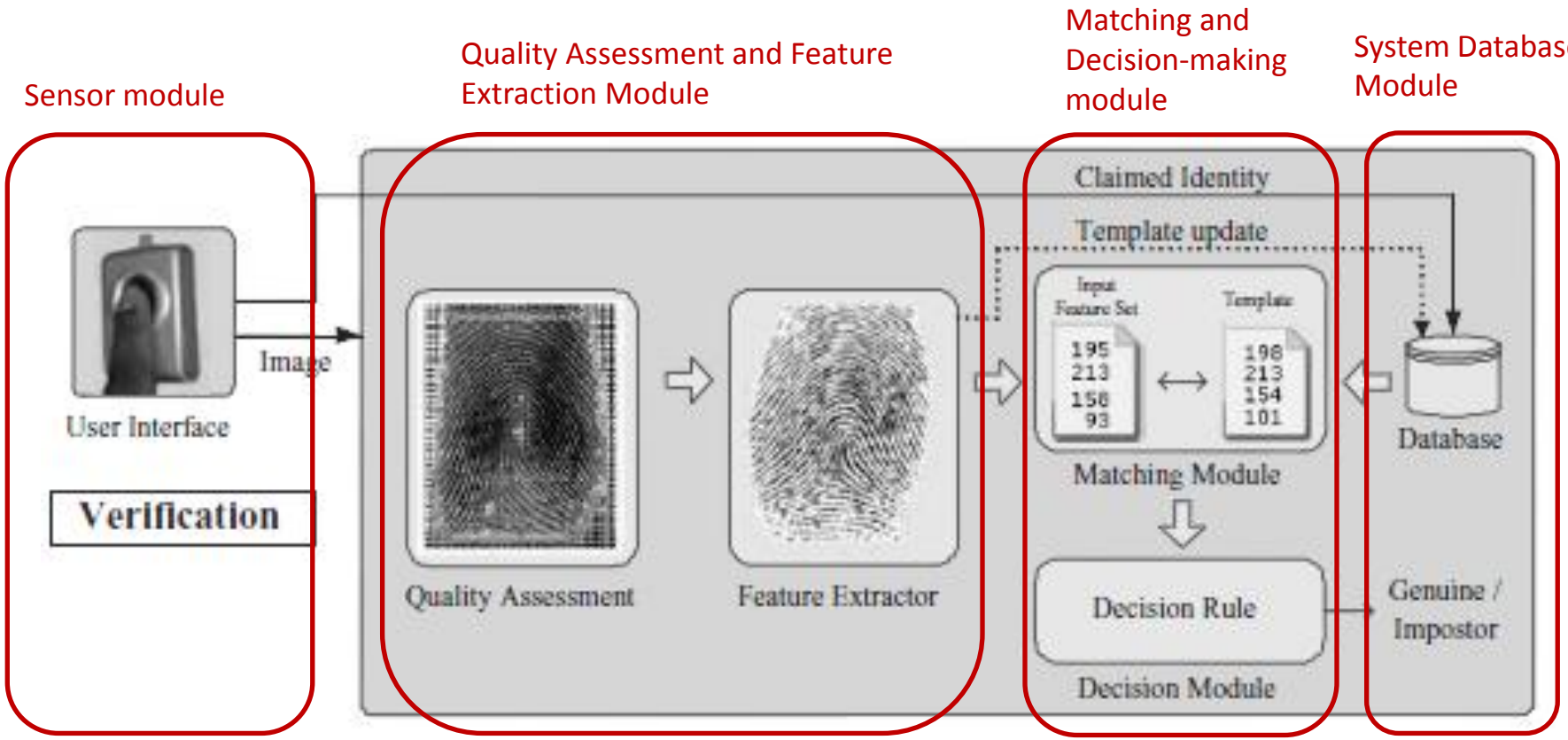
Classification

- Using features and learned models to assign a pattern to a category

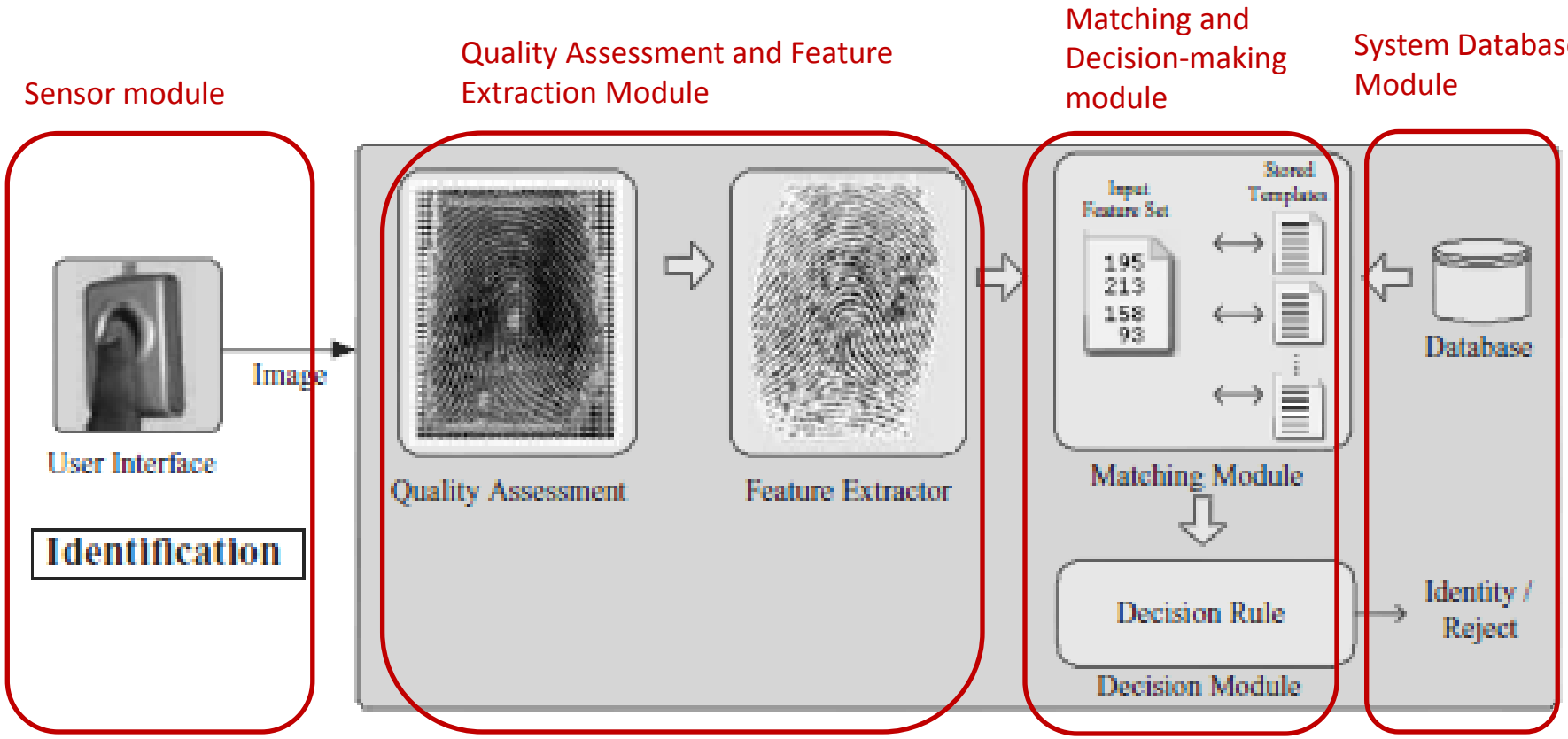
Post-processing

- Evaluation of confidence in decisions
- Exploitation of context to improve performance
- Combination of experts

Operation of a Biometric System: Verification



Operation of a Biometric System: Identification



Bayesian Decision Theory

- Bayesian Decision Theory – a fundamental statistical approach to the problem of pattern classification.
- Decision problem is posed in probabilistic terms
- **Face detection in color images using skin models**
- **Bayesian face recognition**

Bayes' Theorem

To derive Bayes' theorem, start from the definition of *conditional probability*. The probability of the event A given the event B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Equivalently, the probability of the event B given the event A is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Rearranging and combining these two equations, we find

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

Discarding the middle term and dividing both sides by P(B), provided that neither P(B) nor P(A) is 0, we obtain Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

Bayes' Theorem

Bayes' theorem is often completed by noting that, according to the *Law of Total Probability*

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(B|A)P(A) + P(B|A^c)P(A^c) \tag{2}$$

where A^c is the complementary event of A.

Substituting (2) into (1)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

More generally, the *Law of Total Probability* states that given a partition ,$\{A_i\}$, of the event space

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

Therefore, for any partition A_i

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)} \tag{3}$$

Bayes' Theorem: Example

1% of women at age forty who participate in routine screening have breast cancer.
80% of women with breast cancer will get positive mammograms.
9.6% of women without breast cancer will also get positive mammograms.

A woman in this age group had a positive mammography in a routine screening.
What is the probability that she actually has breast cancer?

- Solve for P(B|M) – probability that women at age forty actually having breast cancer given a positive mammogram
- P(B) – probability that women in the group have breast cancer
- P(M|B) – probability that women with breast cancer get a positive mammogram
- $P(B^c)$ – probability that women in the group do not have breast cancer
- $P(M|B^c)$ – probability that women without breast cancer will also get positive mammograms

$$P(B|M) = \frac{P(M|B)P(B)}{P(M|B)P(B) + P(M|B^c)P(B^c)} = \frac{0.8(0.01)}{0.8(0.01) + 0.096(0.99)} = 0.0776$$

Bayes' Theorem

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)} \tag{3}$$

Equation (3) can be extended to *random vectors* described by probability density functions, i.e.

$$P(A_i|\mathbf{x}) = \frac{P(\mathbf{x}|A_i)P(A_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|A_i)P(A_i)}{\sum_i P(\mathbf{x}|A_i)P(A_i)} \tag{4}$$

where we replaced B with $\mathbf{x}$.

Fish Sorting Example Revisited

Define w as the type of fish we observe (state of nature, class), where

- $w = w_1$ for sea bass
- $w = w_2$ for salmon

- $P(w_1)$ is the prior probability that the next fish is a sea bass
- $P(w_2)$ is the prior probability that the next fish is a salmon

Prior probabilities reflect our knowledge of how likely each type of fish will appear before we actually see it

How can we choose $P(w_1)$ and $P(w_2)$

- Set $P(w_1) = P(w_2)$ if they are equiprobable (uniform priors)
- Estimate from available training data, i.e., if N is the total number of available training patterns, and N_1, N_2 of them belong to w_1 and w_2, respectively, then

$$P(w_1) = \frac{N_1}{N}, P(w_2) = \frac{N_2}{N}$$

Making a Decision

Assume there are no other types of fish

$$P(w_1) + P(w_2) = 1$$

How can we make a decision with only prior information?

Decide $$w_i = \begin{cases} w_1 & \text{if } P(w_1) > P(w_2) \\ w_2 & \text{otherwise} \end{cases}$$

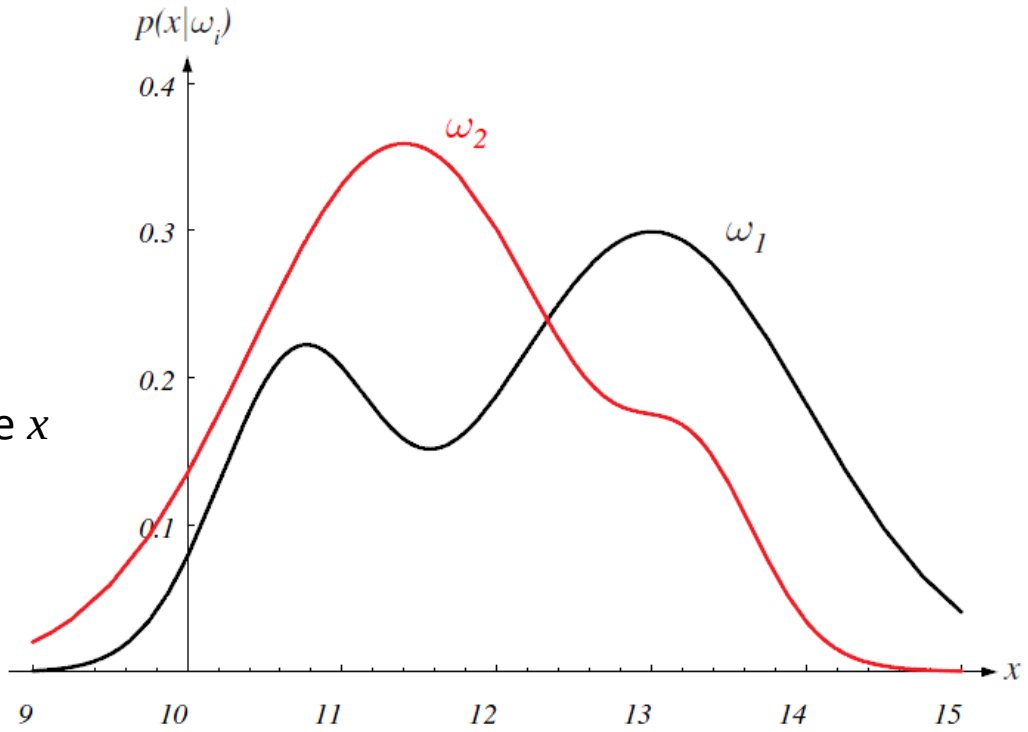
What is the probability of error for this decision?

$$P(error) = min\{P(w_1), P(w_2)\}$$

Class-Conditional Probabilities

- Let's try to improve the decision using the lightness measurement x
- Let x be a continuous random variable
- Define $p(x|w_j)$ as the class-conditional probability density (or probability of x given the state of nature is w_j
- $p(x|w_1)$ and $p(x|w_2)$ describe the difference in lightness between populations of sea bass and salmon

Hypothetical class-conditional pdf's show the probability density of measuring a particular feature value x given the pattern is in category ω_i.



Posterior Probabilities

- Suppose we know $P(w_j)$ and $p(x|w_j)$ and measure the lightness of a fish as the value x
- Define $P(w_j|x)$ as the a posteriori probability (probability of the state of nature being w_j given the measurement of feature value x)
- We can use the Bayes formula to convert the prior probability to posterior probability

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

where $$p(x) = \sum_{i=1}^{2} p(x|w_j)P(w_j)$$

- $p(x|w_j)$ is also called the likelihood and $p(x)$ is called the evidence

Making a Decision

- How can we make a decision after observing the value of x?

$$\text{Decide} \quad w_i = \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

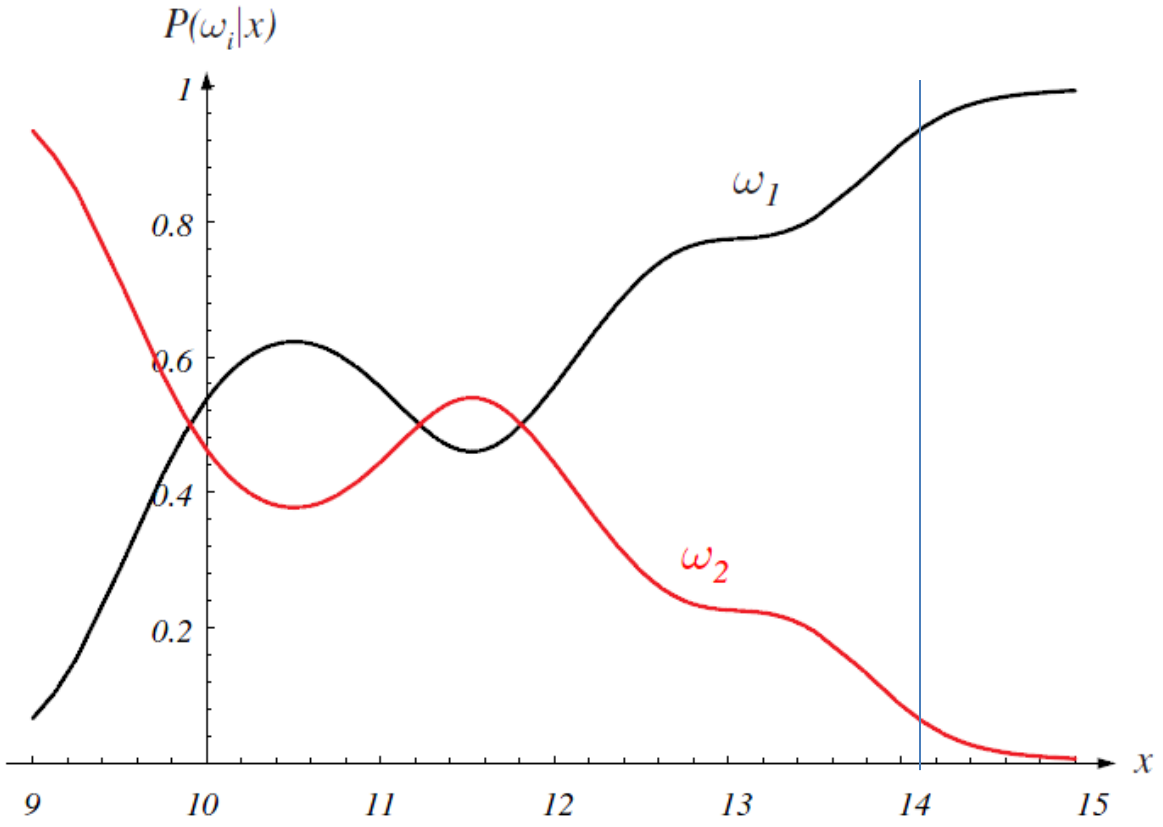
- Rewriting the rules gives

$$\text{Decide} \quad w_i = \begin{cases} w_1 & \text{if } \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} \\ w_2 & \text{otherwise} \end{cases}$$

- Note that at every x

$$P(w_1|x) + P(w_2|x) = 1$$

Making a Decision

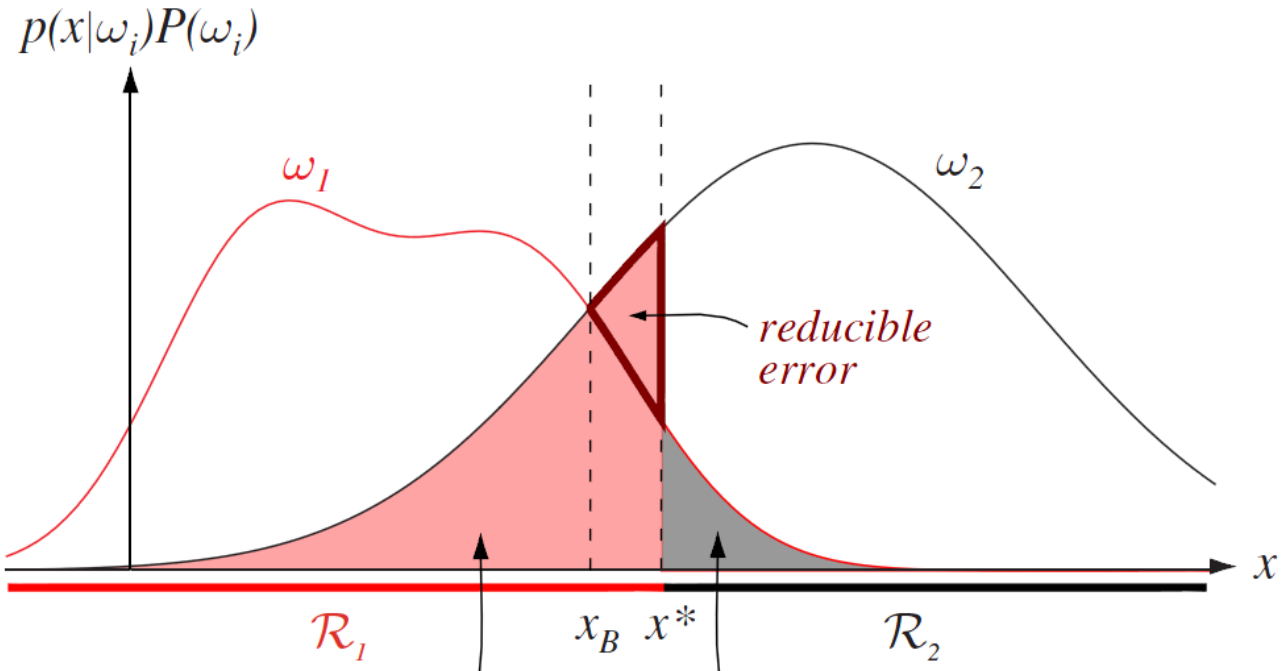


Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown previously. Thus in this case, given that a pattern is measured to have feature value x = 14, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92.

Probability of Error

- What is the probability of error for this decision?

$$P(error|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases}$$



P(error) = $\int p(x|\omega_2)P(\omega_2)\,dx$ + $\int p(x|\omega_1)P(\omega_1)\,dx$

Univariate Gaussian

- The structure of a Bayes classifier is determined by the conditional densities $p(\mathbf{x}|\omega_i)$ as well as by the prior probabilities.

- Of the various density functions that have been investigated, none has received more attention than the multivariate normal or Gaussian density.

For $x \in \mathbb{R}$

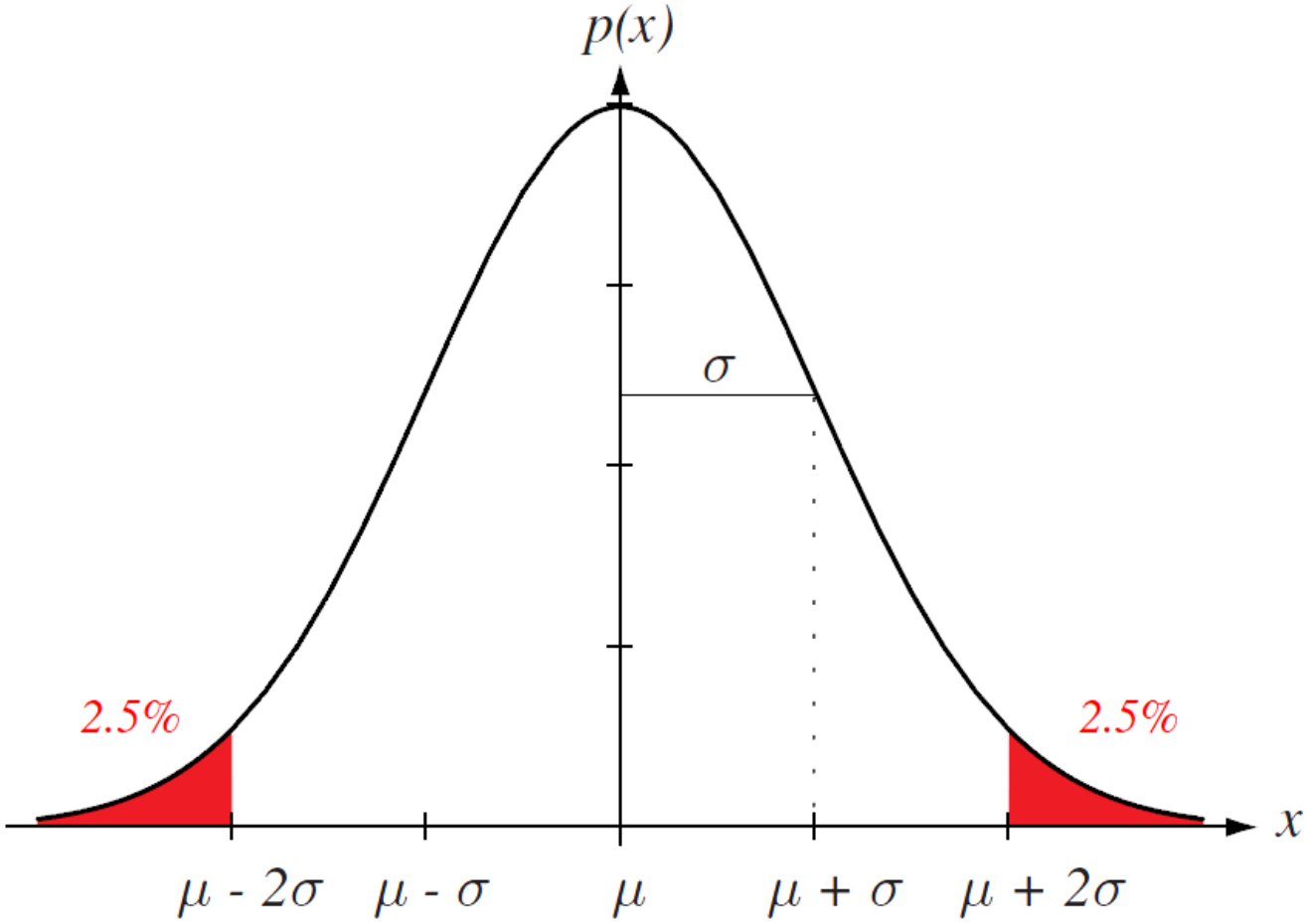
$$\begin{aligned} p(x) &= N(\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \end{aligned}$$

where

$$\mu = E[x] = \int_{-\infty}^{\infty} x p(x)\, dx$$

$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx$$

Univariate Gaussian



A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$

Multivariate Gaussian

For $\mathbf{x} \in \mathbb{R}^d$

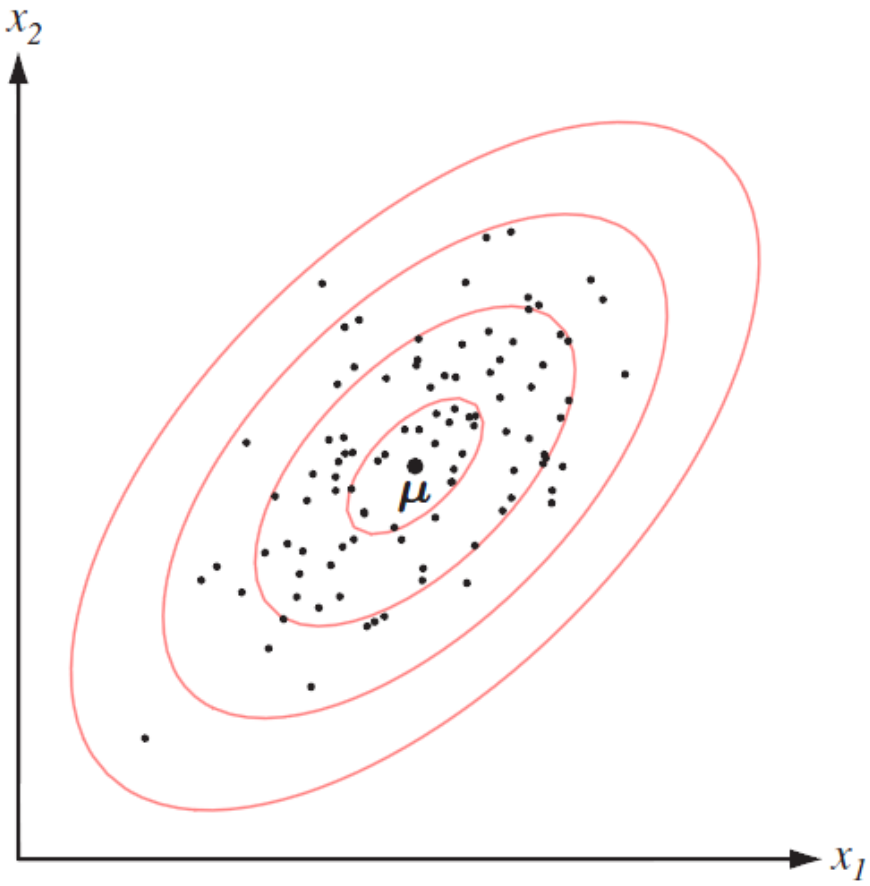
$$\begin{aligned} p(\mathbf{x}) &= N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \end{aligned}$$

where

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x}p(\mathbf{x})\,d\mathbf{x}$$

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T p(\mathbf{x})\,d\mathbf{x}$$

Multivariate Gaussian



Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean . The ellipses show lines of equal probability density of the Gaussian.

Matlab Illustrations

We will look at a univariate classification problem with equal priors and two classes. The class-conditionals are given by the normal distributions as follows:

$$p(x|w_1) = N(x; -1, 1)$$
$$p(x|w_2) = N(x; 1, 1)$$

The priors are

$$P(w_1) = 0.6$$
$$P(w_2) = 0.4$$

First step: Create a function that returns the value of the univariate/multivariate probability function

Matlab Illustrations

```
function prob = csevalnorm(x,mu,cov_mat)


[n,d]=size(x);
prob = zeros(n,1);
a=(2*pi)^(d/2)*sqrt(det(cov_mat));
covi = inv(cov_mat);
for i = 1:n
    xc = x(i,:)-mu;
    arg=xc*covi*xc';
    prob(i)=exp((-.5)*arg);
end
prob=prob/a;
```

$$p(\mathbf{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

[2]W. Martinez and A. Martinez, *Computational Statistics Handbook with MATLAB*, 2nd edition, Chapman and Hall/CRC, Inc., 2007

Matlab Illustrations

```
% Get the domain for the densities.
dom = -6:.1:8;
dom = dom';

pxg1 = csevalnorm(dom,-1,1); % Class-conditional 1
pxg2 = csevalnorm(dom,1,1);   % Class-conditional 2

figure, plot(dom,pxg1,'r',dom,pxg2,'b')
xlabel('Feature-x')
legend('Class-conditional 1','Class-conditional 2')

% Posterior
ppxg1 = pxg1*0.6; % Multiply by priors
ppxg2 = pxg2*0.4;

figure, plot(dom,ppxg1,'r',dom,ppxg2,'b')
xlabel('Feature-x')
legend('Posterior 1','Posterior 2')
```

Matlab Illustrations

Let's see what happens when x = -0.75

```
x = -0.75;
% Evaluate each un-normalizd posterior.
po1 = csevalnorm(x,-1,1)*0.6
po2 = csevalnorm(x,1,1)*0.4
```

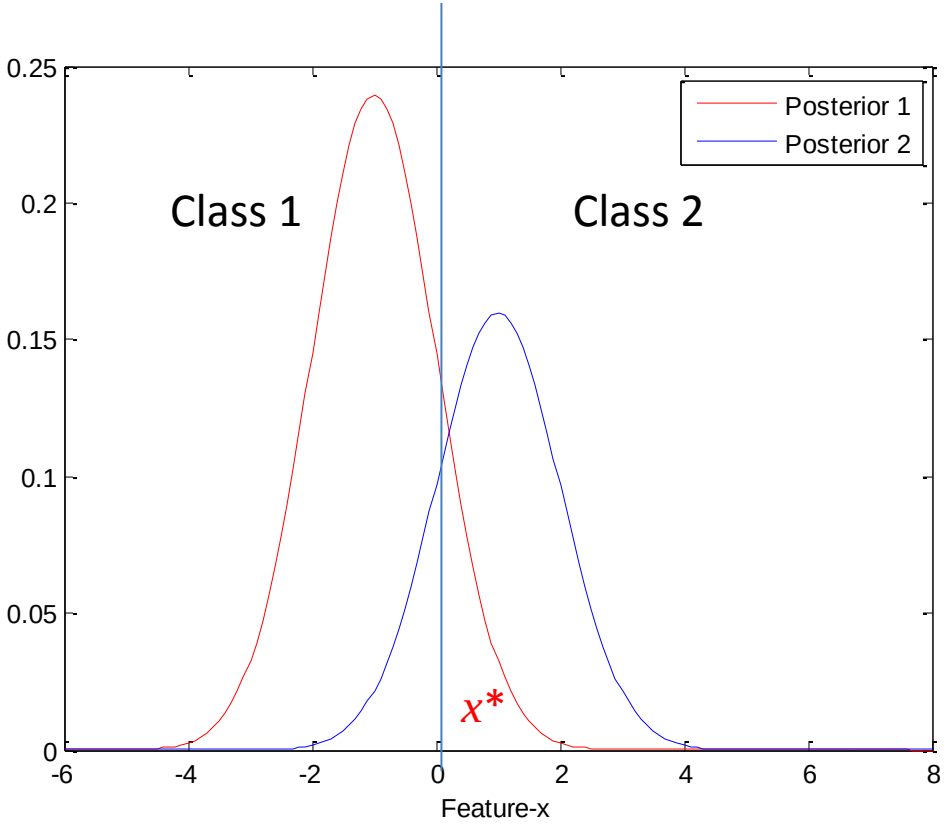
$$P(x = -0.75|w_1)P(w_1) = 0.23$$
$$P(x = -0.75|w_2)P(w_2) = 0.04$$

$$w_i = \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

Classify as **Class 1 (w_1).**

Matlab Illustrations: Bayesian Decision Rule, Prob. of Error

$$w_i = \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$



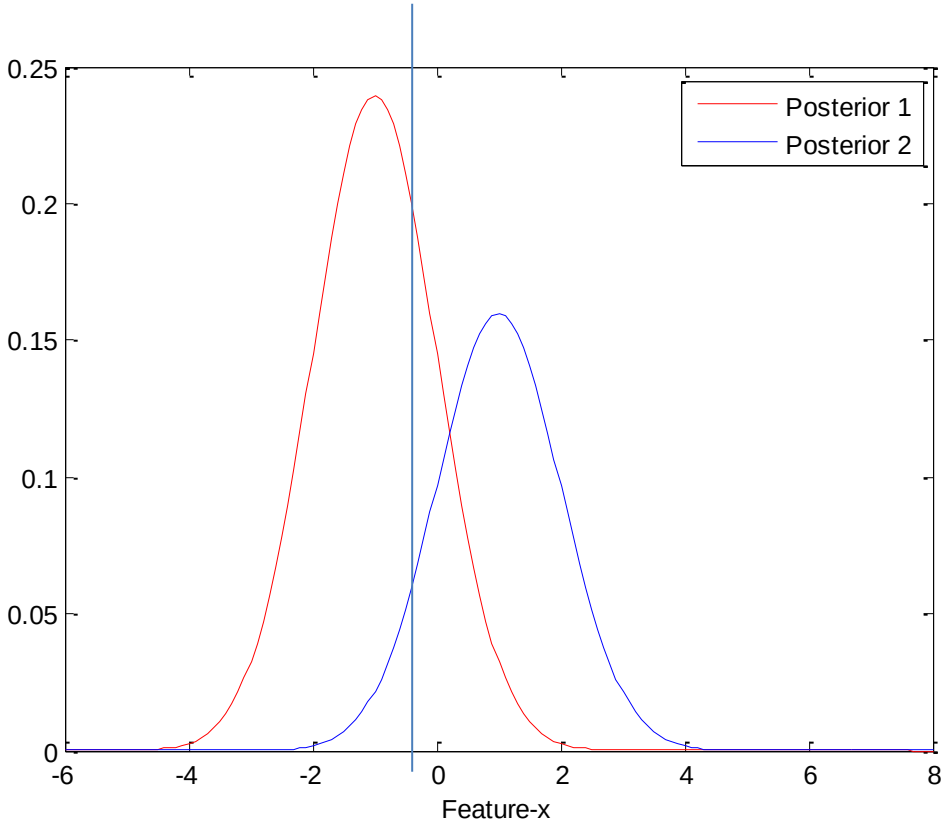

```
% Note that 0.1 is the step
size and we are
approximating the integral
using a % sum.

ind1 = find(ppxg1 >= ppxg2);
% Now find the other part.
ind2 = find(ppxg1<ppxg2);
pmis1 = sum(ppxg1(ind2))*.1;
pmis2 = sum(ppxg2(ind1))*.1;
errorhat = pmis1 + pmis2
```

$$P(error) = \int_{-\infty}^{x*} P(w_2|x)\,dx + \int_{x*}^{\infty} P(w_1|x)\,dx = 0.1539$$

Matlab Illustrations: Probability of Error

What happens to the error if the threshold is set at x = -0.5




```
% Change the decision boundary.
bound = -0.5;
ind1 = find(dom <= bound);
ind2 = find(dom > bound);
pmis1 = sum(ppxg1(ind2))*.1;
pmis2 = sum(ppxg2(ind1))*.1;
errorhat = pmis1 + pmis2
```

If we change the boundary to other than the threshold solved by the Bayes decision rule, the error will be greater.

$$P(error) = \int_{-\infty}^{-0.5} P(w_2|x)\,dx + \int_{0.5}^{\infty} P(w_1|x)\,dx = 0.2040$$

Summary

- Human and Machine Perception
- What is pattern recognition?
- Sample pattern recognition problem
- Pattern recognition systems
- Bayes decision theory
- Matlab illustrations and sample programs
- Conclusion

Next Topic

- Receiver Operating Characteristics (ROCs)
- More Matlab illustrations
- Evaluation of Biometric Systems, Definition of Terms
- Face Detection in Color Images using Skin Models

References

R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2000

Selim Aksoy, CS 551(Pattern Recognition) Course Website, http://www.cs.bilkent.edu.tr/~saksoy/courses/cs551-Spring2010/index.html

W. Martinez and A. Martinez, *Computational Statistics Handbook with MATLAB*, 2nd edition, Chapman and Hall/CRC, Inc., 2007