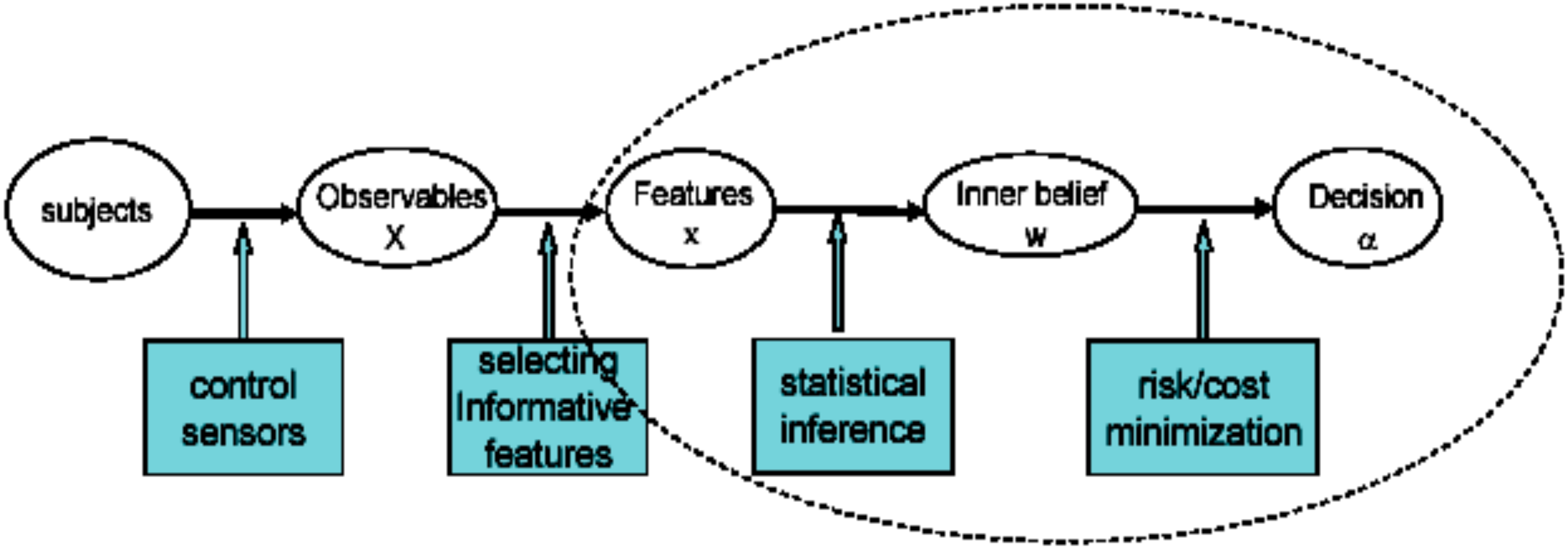
Bayesian Decision Theory

Introduction



In Bayesian decision theory, we are concerned with the last three steps in the big ellipse assuming that the observables are given and features are selected.

Introduction

- The sea bass/salmon example
  - State of nature, prior
    - State of nature is a random variable
    - EXAMPLE PRIOR: The catch of salmon and sea bass is equiprobable
      - $P(\omega_1) = P(\omega_2)$ (uniform priors)
      - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2

- Use of the class –conditional information

- $P(x \mid \omega_1)$ and $P(x \mid \omega_2)$ describe the difference in lightness between populations of sea and salmon

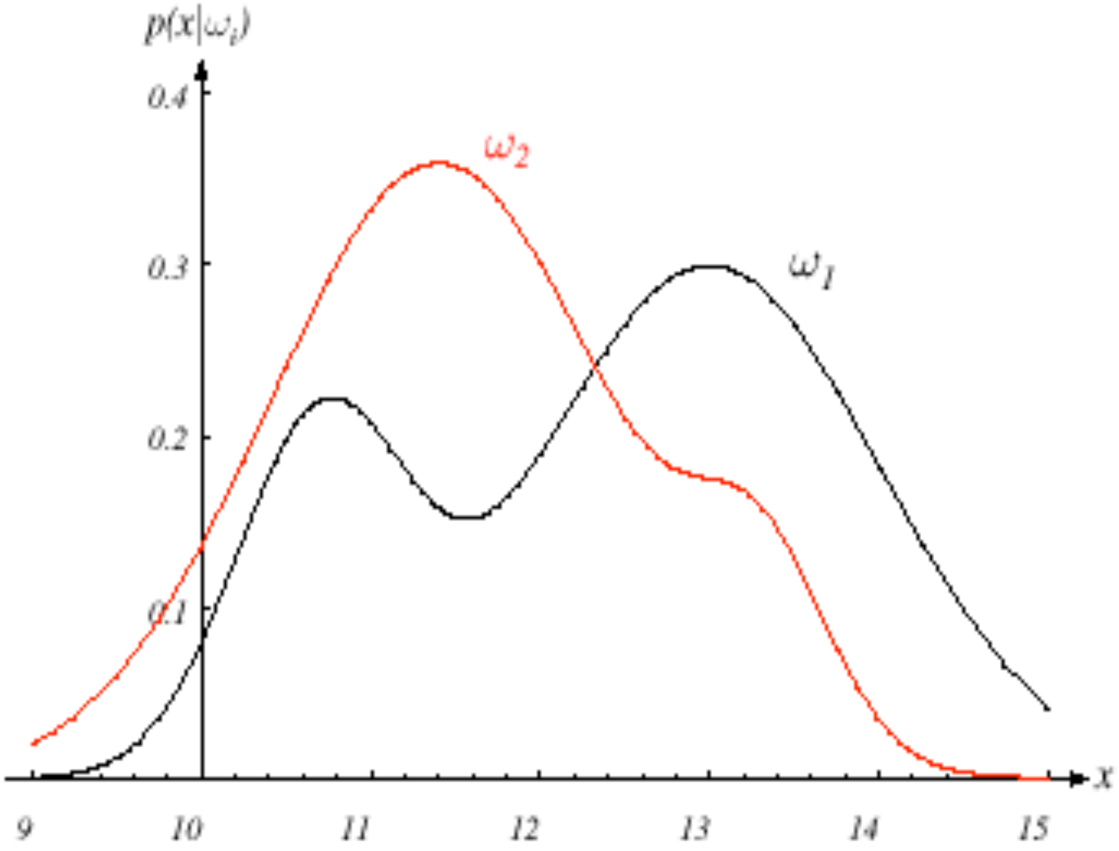


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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 theorem

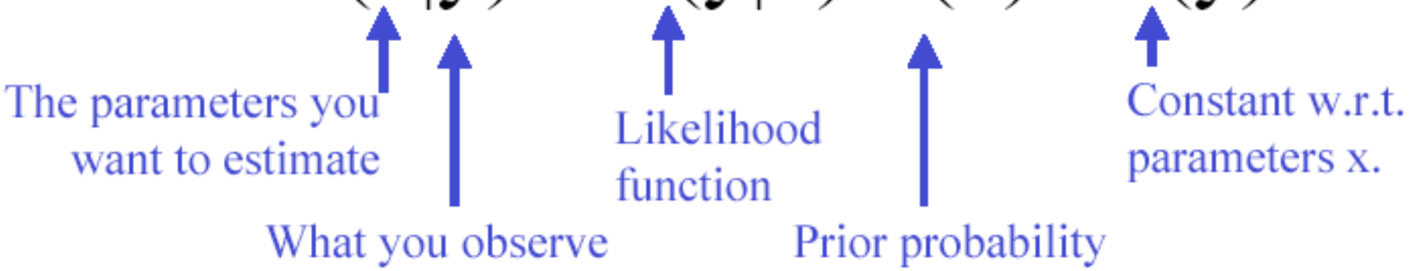
$$P(x, y) = P(x|y)\ P(y)$$

so

$$P(x|y)\ P(y) = P(y|x)\ P(x)$$

and

$$P(x|y) = P(y|x)\ P(x)\ /\ P(y)$$



The parameters you want to estimate

What you observe

Likelihood function

Prior probability

Constant w.r.t. parameters x.

- Posterior, likelihood, evidence

  - $P(\omega_j \mid x) = P(x \mid \omega_j) \,.\, P(\omega_j) / P(x)$

  - Where in case of two categories

$$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j) P(\omega_j)$$

  - Posterior = (Likelihood. Prior) / Evidence

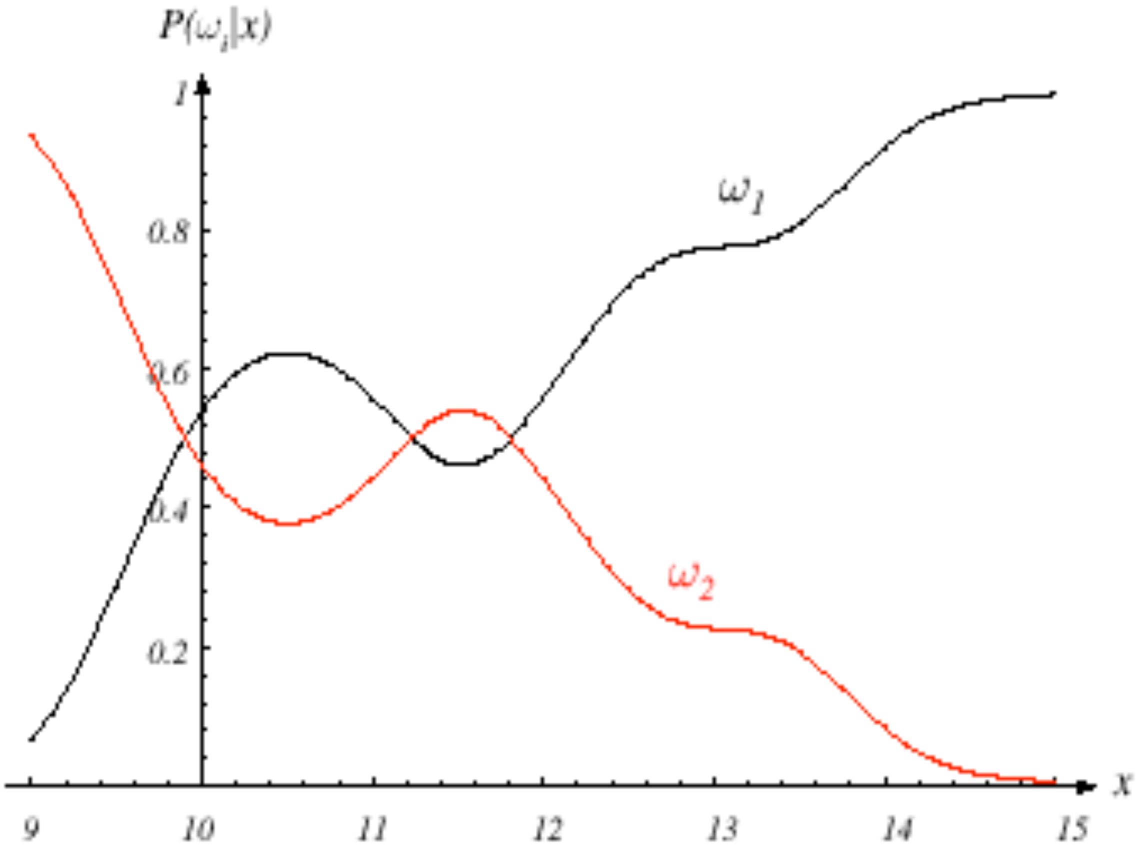


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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.

Example

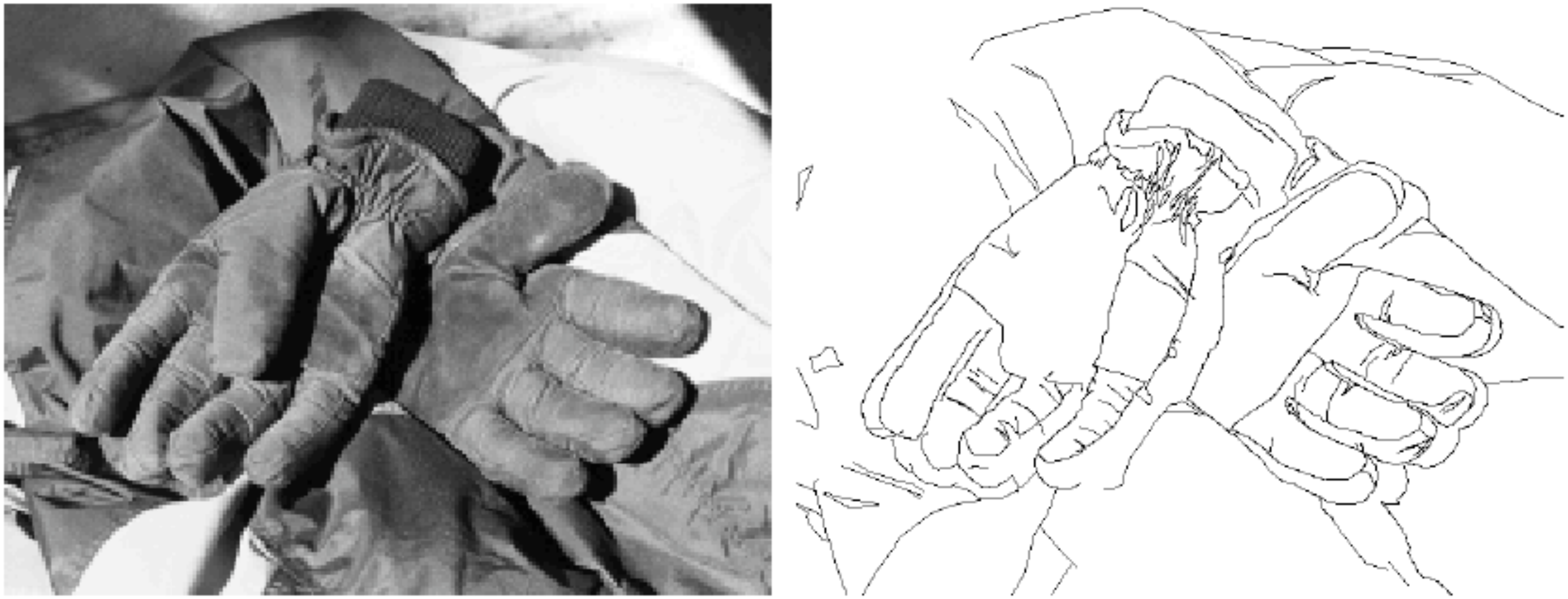


Figure 2.4 A typical image (left) and the ground truth segmentation (right). courtesy of K. Bowyer at U. South Florida.

Need for Prior Info

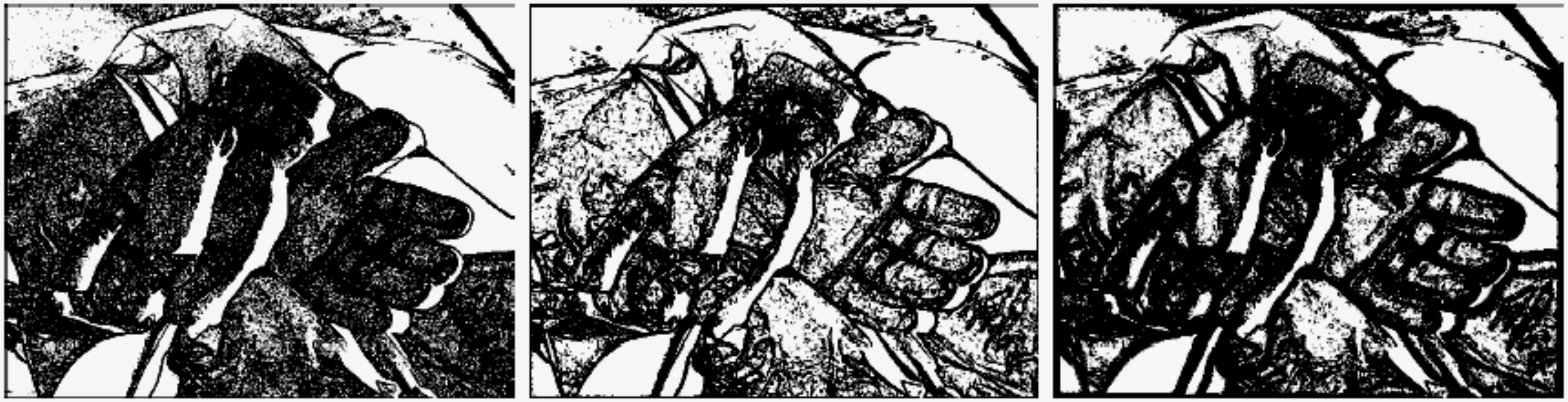


Figure 2.6 The edge estimated on the glove image by ML with the filter at scale 0 (left), filter at scale 1 (centre), and filter with scales 0, 1, 2, 4 (right). Observe that ML significantly overestimates the number of edges in this image.

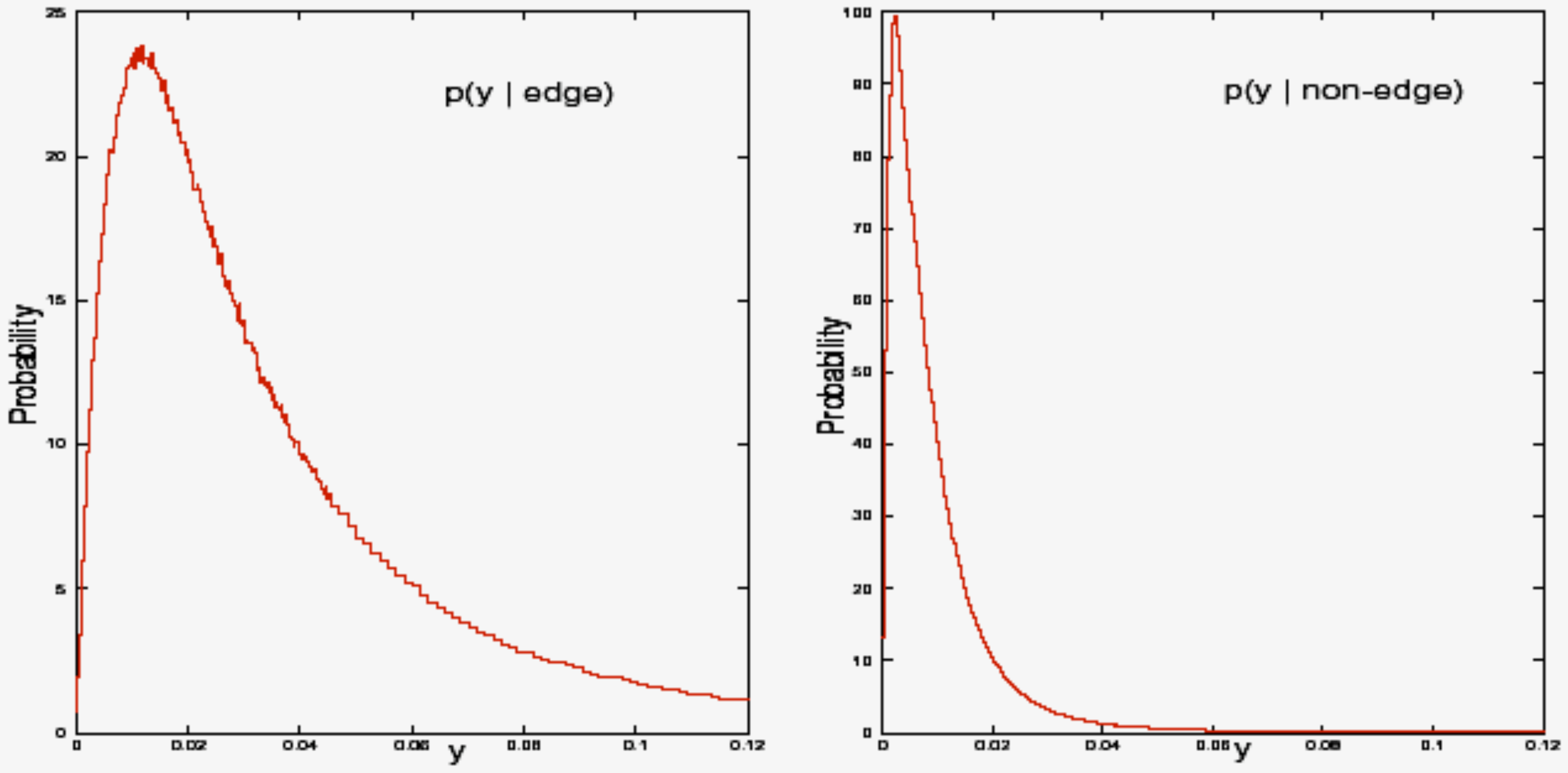


Figure 2.5 Empirical distributions for a gradient filter response on boundaries (left) and off boundaries (right).

Minimal Error Decisions

- Make the best guess in terms of error rate given data and category probabilities

- Decision given the posterior probabilities

Given an observation X, Decide (Guess):

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ True state of nature = ω_1

if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ True state of nature = ω_2

Therefore:

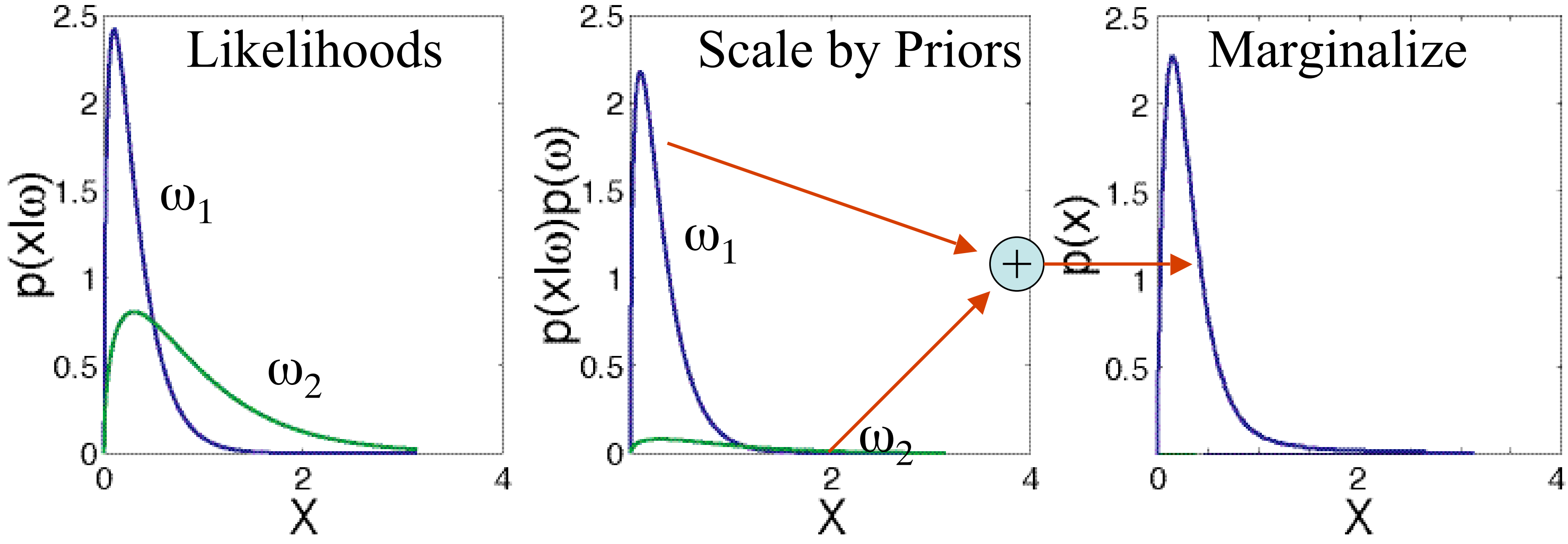
For a particular x, the probability of error is :

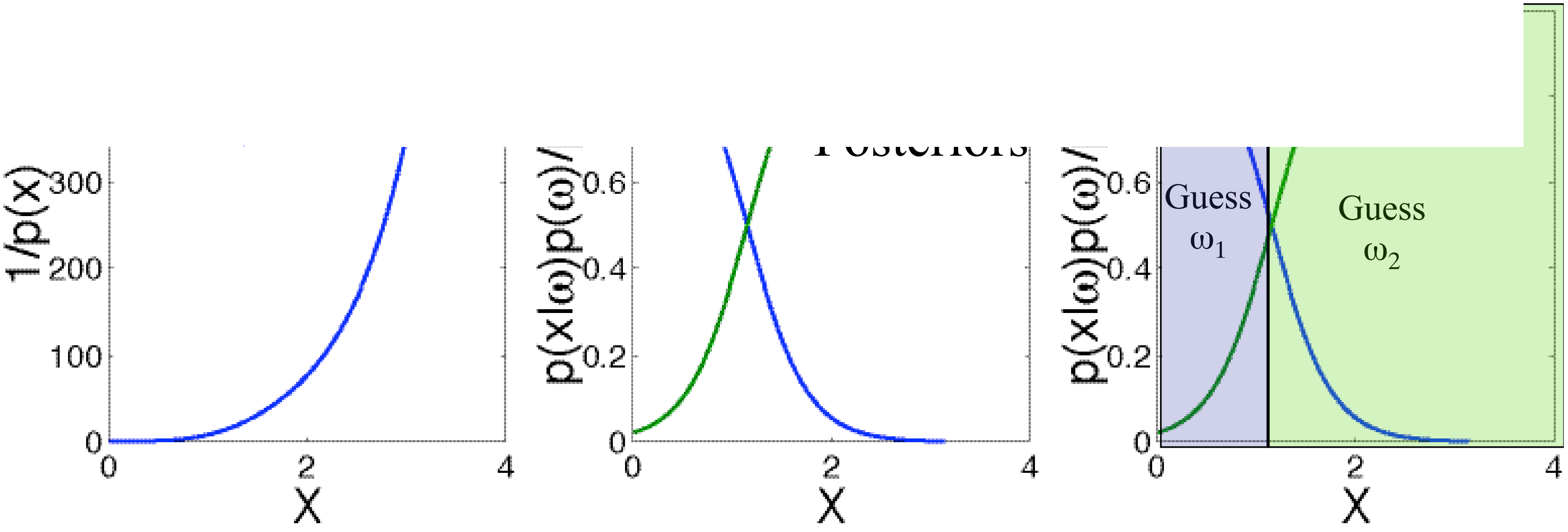
$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$

$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

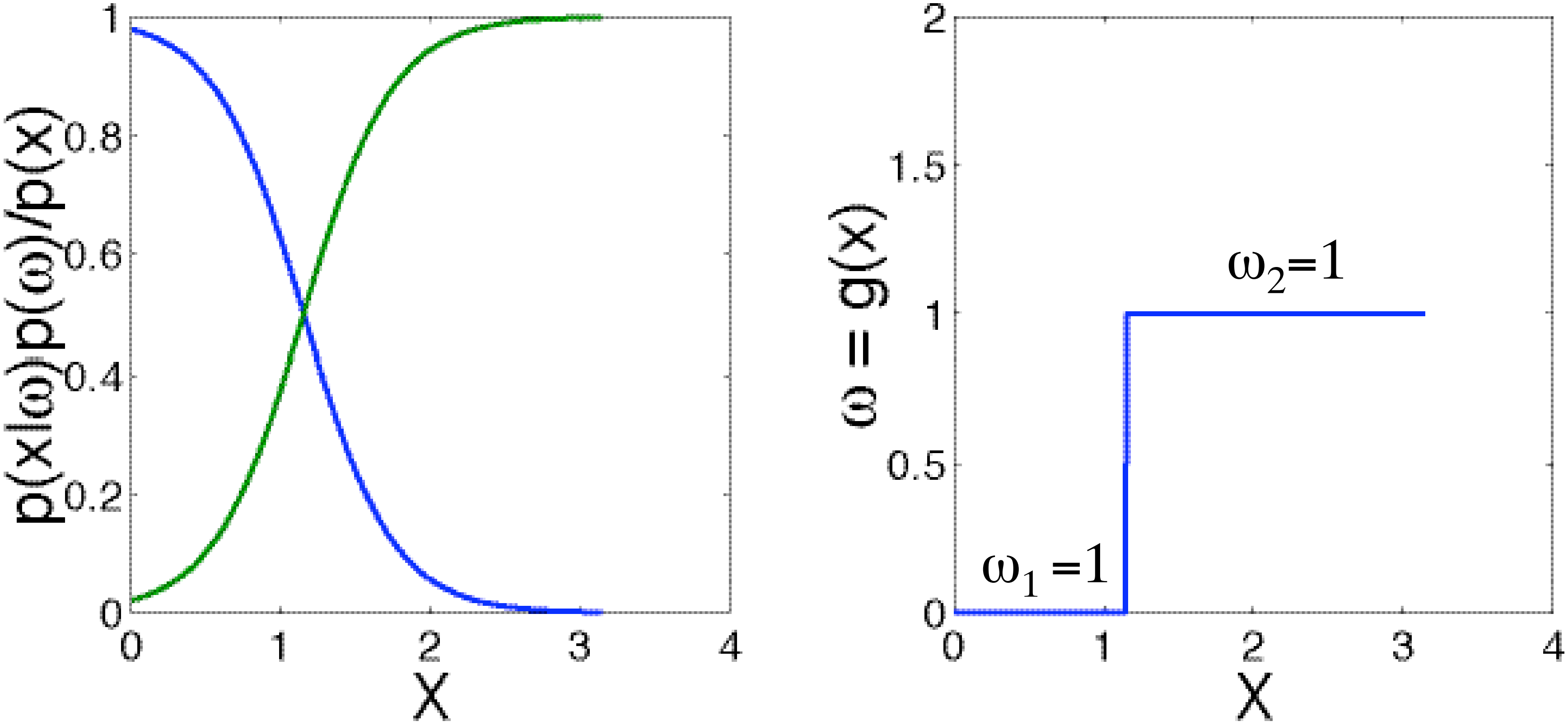
$$P(error \mid x) = min\ [P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

(Bayes decision)


Likelihoods
p(x|ω)
ω_1
ω_2
0
0.5
1
1.5
2
2.5
0
2
4
X
Scale by Priors
p(x|ω)p(ω)
ω_1
ω_2
+
0
0.5
1
1.5
2
2.5
0
2
4
X
Marginalize
p(x)
0.5
1
1.5
2
2.5
0
2
4
X


1/p(x)
0
100
200
300
0
2
4
X
Posteriors
p(x|ω)p(ω)/
0
0.2
0.4
0.6
0
2
4
X
p(x|ω)p(ω)/
Guess ω_1
Guess ω_2
0
0.2
0.4
0.6
0
2
4
X

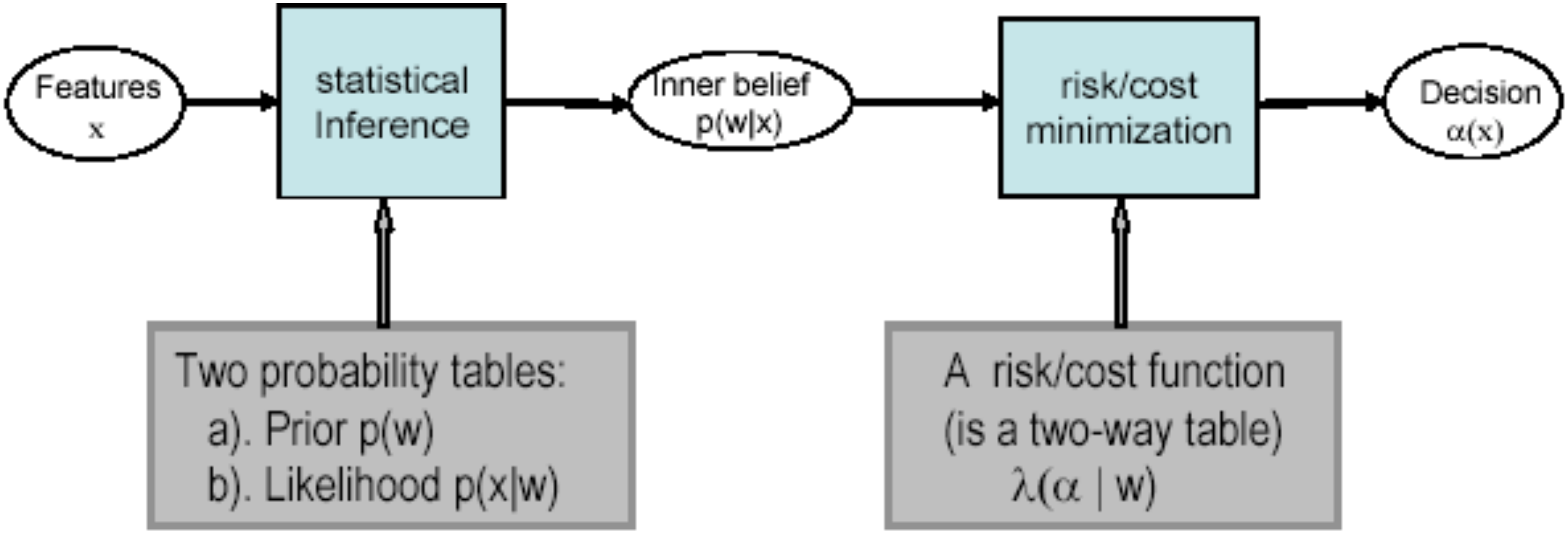
Decision Functions



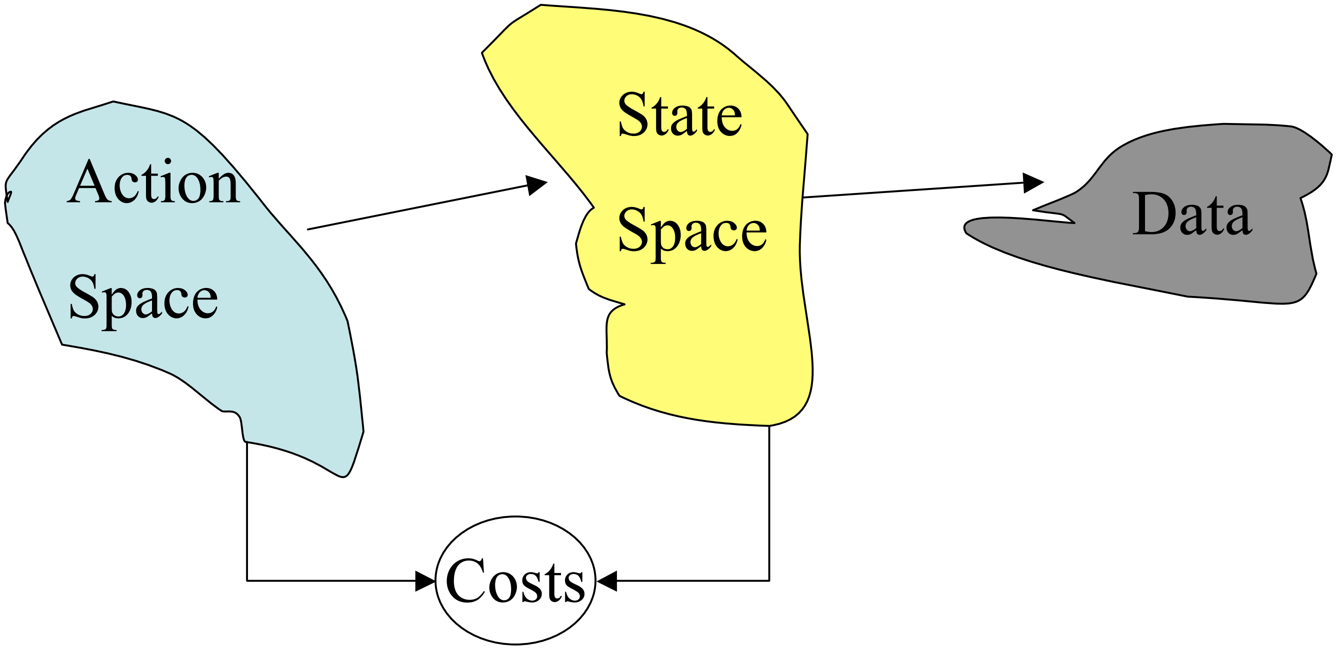
Bayesian Decision Theory – Continuous Features

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allow for actions more general than deciding the state of nature:
    - Introduce a loss (cost) function which is more general than the probability of error (measure of the space of outcomes).

Bayesian Decisions



- $\{\omega_1, \omega_2, \ldots, \omega_c\}$: the set of c states of nature ("categories")
- Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions
- Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j



- The loss function encodes the relative cost of each action. (Negate the cost? Gain function)

- Classification: Action = class choice.
  - Allowing actions other than class choice includes data rejection (e.g. unknown class).

Decision Rules

- A decision rule is a mapping function from feature space to the set of actions:

  g: $\alpha = g(x)$

  What is the optimal decision rule?
  Minimize risk.

Overall risk:

$$R = \text{Sum of all } \underbrace{R(\alpha_i \mid x)}_{\textbf{Conditional risk}} \text{ for } i = 1,\ldots,a$$

Minimizing R ⟺ Minimizing $R(\alpha_i \mid x)$ *for* $i = 1,\ldots, a$

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

for i = 1,…,a

- Select the action α_i for which $R(\alpha_i \mid x)$ is minimum

Minimum($R(\alpha_i \mid x)$*) is called the Bayes risk:*

The best performance that can be achieved!

- Two-category classification

$$\alpha_1 : deciding\ \omega_1$$

$$\alpha_2 : deciding\ \omega_2$$

$$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$$

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

decide ω_1 if:

$$(\lambda_{21} - \lambda_{11})\, P(x \mid \omega_1)\, P(\omega_1) >$$
$$(\lambda_{12} - \lambda_{22})\, P(x \mid \omega_2)\, P(\omega_2)$$

and decide ω_2 otherwise

Likelihood ratio:

The preceding rule is equivalent to the following rule:

$$if\ \frac{P(x\,|\,\omega_1)}{P(x\,|\,\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)
Otherwise take action α_2 (decide ω_2)

Optimal decision property

“If the likelihood ratio exceeds a threshold value independent of the input pattern x, we can take optimal actions”

Exercise

Select the optimal decision where:

$\Omega = \{\omega_1, \omega_2\}$

N(μ , σ)

$P(x \mid \omega_1)$ → N(2, 0.5) (Normal distribution)

$P(x \mid \omega_2)$ → N(1.5, 0.2)

$P(\omega_1) = 2/3$

$P(\omega_2) = 1/3$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

Minimum-Error-Rate Classification

- Action: Choose class

  If action α_i is taken and the true state of nature is ω_j then:

  the decision is correct if $i = j$ and in error if $i \neq j$

- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

- Introduction of the zero-one loss function:

$$\lambda(\alpha_i,\omega_j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases}\qquad i,j=1,\dots,c$$

Therefore, the conditional risk is:

$$R(\alpha_i\mid x)=\sum_{j=1}^{j=c}\lambda(\alpha_i\mid\omega_j)P(\omega_j\mid x)$$

$$=\sum_{j\neq 1}P(\omega_j\mid x)=1-P(\omega_i\mid x)$$

“The risk corresponding to this loss function is the average probability error”

- Minimize the risk requires maximize $P(\omega_i \mid x)$ (since $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$)

- For Minimum error rate
  - Decide $\boldsymbol{\omega_i}$ *if* $P\ (\boldsymbol{\omega_i} \mid x) > P(\boldsymbol{\omega_j} \mid x)\ \forall j \neq i$

- Regions of decision and zero-one loss function, therefore:

$$Let\ \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\cdot\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda\ then\ decide\ \omega_1\ \ if:\ \frac{P(x|\omega_1)}{P(x|\omega_2)}>\theta_\lambda$$

- If λ is the zero-one loss function which means:

$$\lambda=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$

$$then\ \ \theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}=\theta_a$$

$$if\ \ \lambda=\begin{pmatrix}0 & 2\\ 1 & 0\end{pmatrix}then\ \ \theta_\lambda=\frac{2P(\omega_2)}{P(\omega_1)}=\theta_b$$

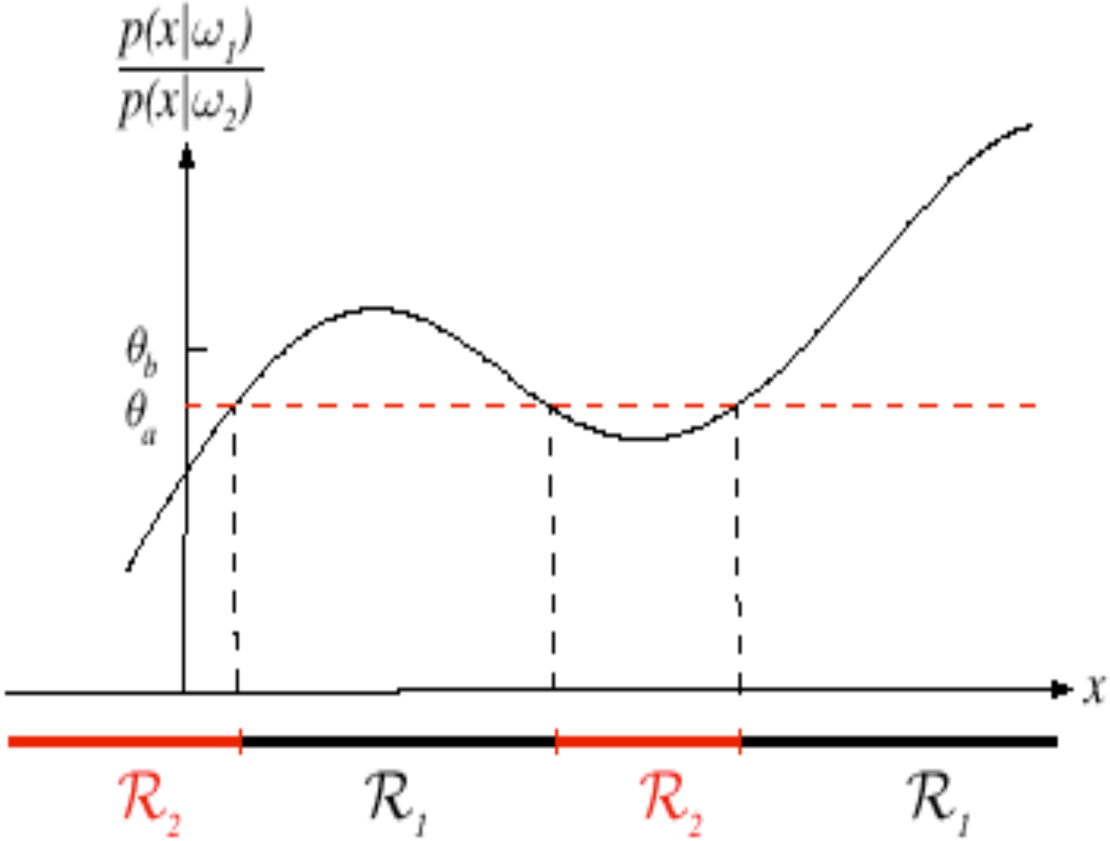


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classifiers, Discriminant Functions and Decision Surfaces

- The multi-category case
  - Set of discriminant functions $g_i(x), i = 1,\ldots, c$
  - The classifier assigns a feature vector x to class $\boldsymbol{\omega_i}$ if:

$$g_i(x) > g_j(x) \;\forall j \neq i$$

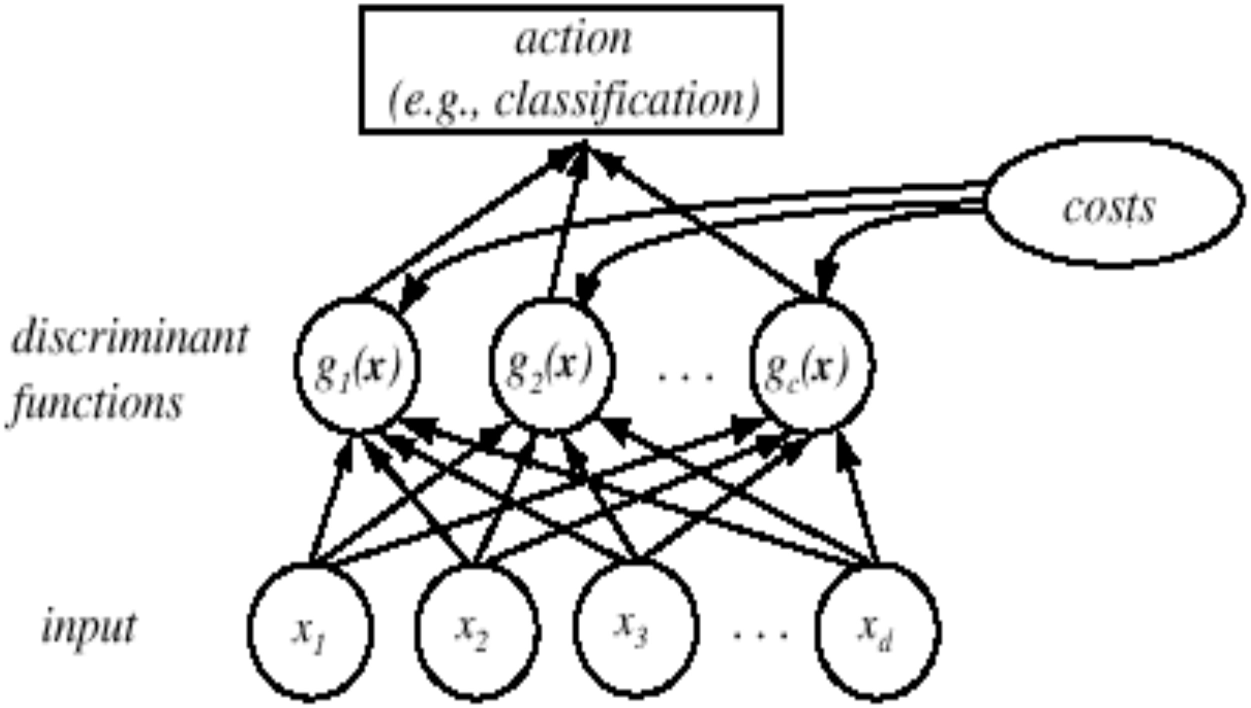


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- *Let* $g_i(x) = - R(\alpha_i \mid x)$
  (max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take

$$g_i(x) = P(\omega_i \mid x)$$

(max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$$

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

(ln: natural logarithm)

- Feature space divided into c decision regions

$$if\ g_i(x) > g_j(x)\ \forall j \neq i\ then\ x\ is\ in\ \mathcal{R}_i$$

($\mathcal{R}_i$ means assign x to ω_i)

- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions g_1 and g_2

    Let $g(x) \equiv g_1(x) - g_2(x)$

$$\text{Decide } \omega_1\ if\ g(x) > 0\ \text{; Otherwise decide } \omega_2$$

– The computation of g(x)

$$g(x) = \ln \frac{P(\omega_1 | x)}{P(\omega_2 | x)}$$

$$\begin{aligned} g(x) &= \ln P(\omega_1 | x) - \ln P(\omega_2 | x) \\ &= \ln \frac{P(x | \omega_1) P(\omega_1)}{P(x)} - \ln \frac{P(x | \omega_2) P(\omega_2)}{P(x)} \\ &= \ln \frac{P(x | \omega_1)}{P(x | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \end{aligned}$$

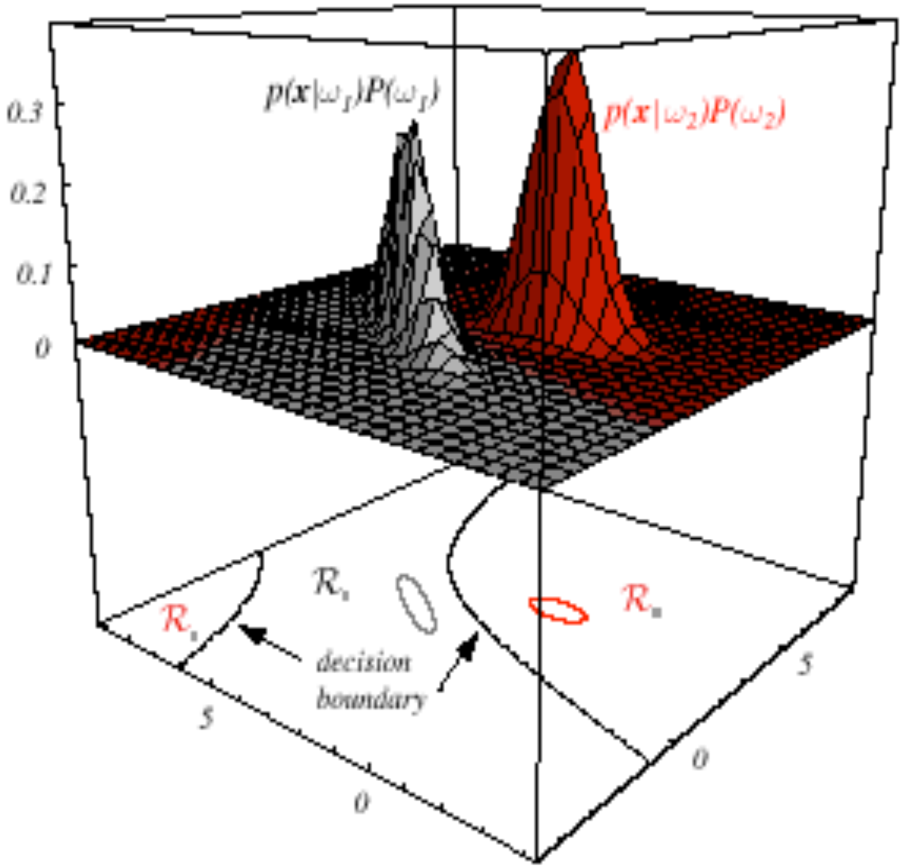


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

The Normal Density

- Univariate density
  - Density which is analytically tractable
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

Where:

μ = mean (or expected value) of x

σ^2 = expected squared deviation or variance

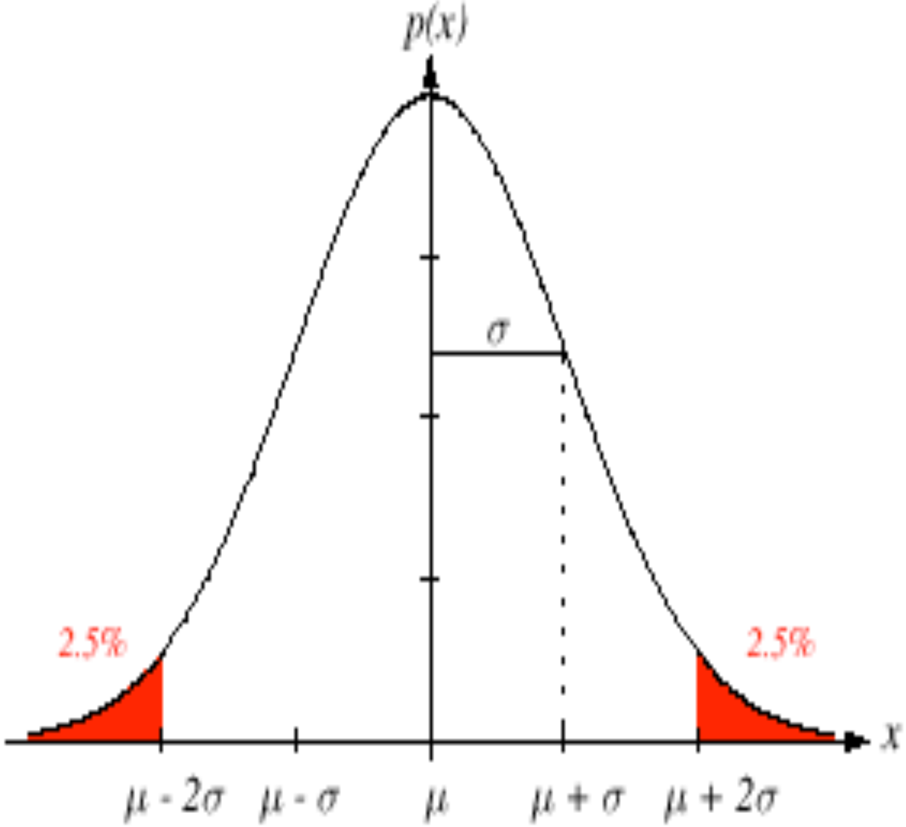


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Multivariate density
  - Multivariate normal density in d dimensions is:

$$P(x) = \frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right]$$

where:

$x = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose vector form)

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d*d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively