Pattern Recognition

Bayesian Decision Theory

- **Bayes Rule**
- **Bayes Error**
- **Loss Function**
- **Discriminant Function**
- **Normal Distribution**
- **Maximum Likelihood**

An Example: Prior Probability

- One day, Peter shows swine flu symptoms

  He went to see a doctor if he is OK

$$\omega \in \{ill, healthy\}$$

- The doctor said, according to past data,
  - ✓ 85% of people was healthy

    Prior Probability $p(\omega = healthy) = 0.85$
  - ✓ 15% of people was ill

    Prior Probability $p(\omega = ill) = 0.15$
  - ✓ $p(\omega = healthy) > p(\omega = ill)$

    therefore, Peter was healthy

An Example: Conditional Probability

- The problem of using Prior Probability to determine our health condition:

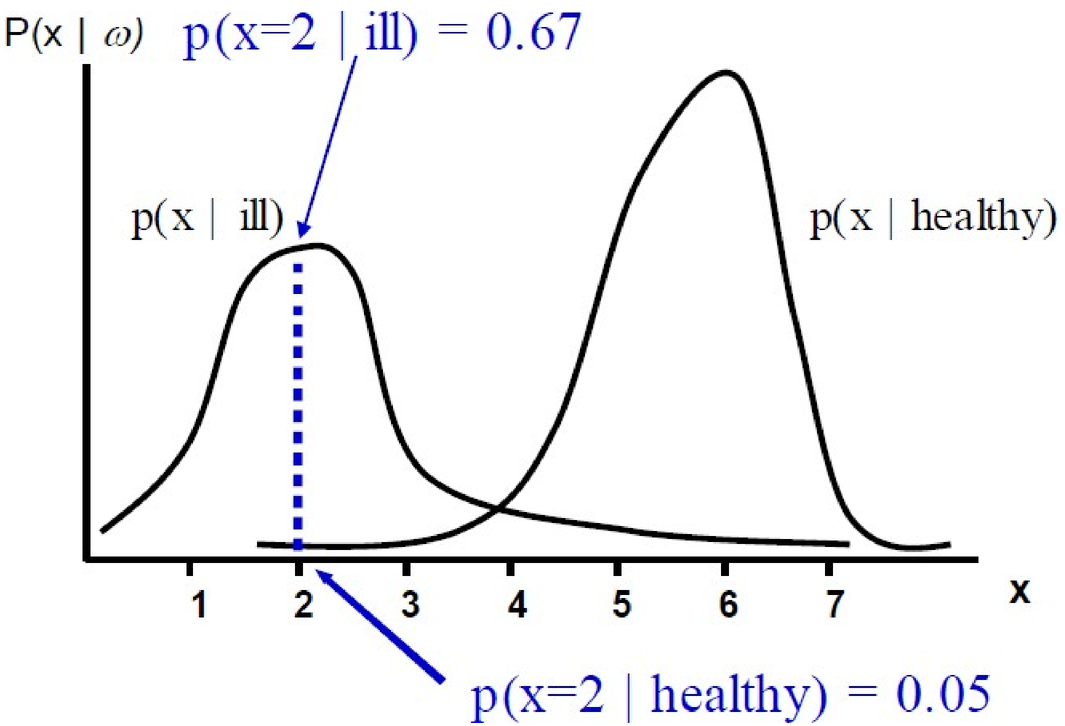
  Ignore our present state of health

- Need to measure some patterns of our physical condition

Example: Measure the amount of red blood cells denoted by x

An Example: Conditional Probability

- According to the previous knowledge (Conditional Probability, also called likelihood):



An Example: Posterior Probability

- Now, we know the following information:

$$p(ill) = 0.15 \qquad p(healthy) = 0.85$$

$$p(x = 2|ill) = 0.67 \qquad p(x = 2|healthy) = 0.05$$

- How can they help the doctor make a decision?
  - ✓ Bayes Formula

$$\underbrace{p(\omega|x)}_{\text{posterior}} = \frac{\overbrace{p(x|\omega)}^{\text{likelihood}}\overbrace{p(\omega)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

Bayes Rule

$$p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)}$$

- Bayes Decision Rule:
  - ✓ Decide ω_1 if $p(\omega_1|x) > p(\omega_2|x)$
  - ✓ Decide ω_2 if $p(\omega_2|x) > p(\omega_1|x)$
- Or equivalent to
  - ✓ Decide ω_1 if $p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$
  - ✓ Decide ω_2 if $p(x|\omega_2)p(\omega_2) > p(x|\omega_1)p(\omega_1)$

$p(x)$ is ignored since it is the same for both classes

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

Maximum Likelihood (ML) Rule

$$p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)}$$

- It is a special case of Bayes Rule

  When $p(\omega_1) = p(\omega_2)$, the decision is based entirely on the likelihood $p(x|\omega_j)$

$$p(\omega|x) \propto p(x|\omega)$$

- Decision Rule:
  - ✓ Decide ω_1 if $p(x|\omega_1) > p(x|\omega_2)$
  - ✓ Decide ω_2 if $p(x|\omega_2) > p(x|\omega_1)$

An Example: Final Decision by Doctor

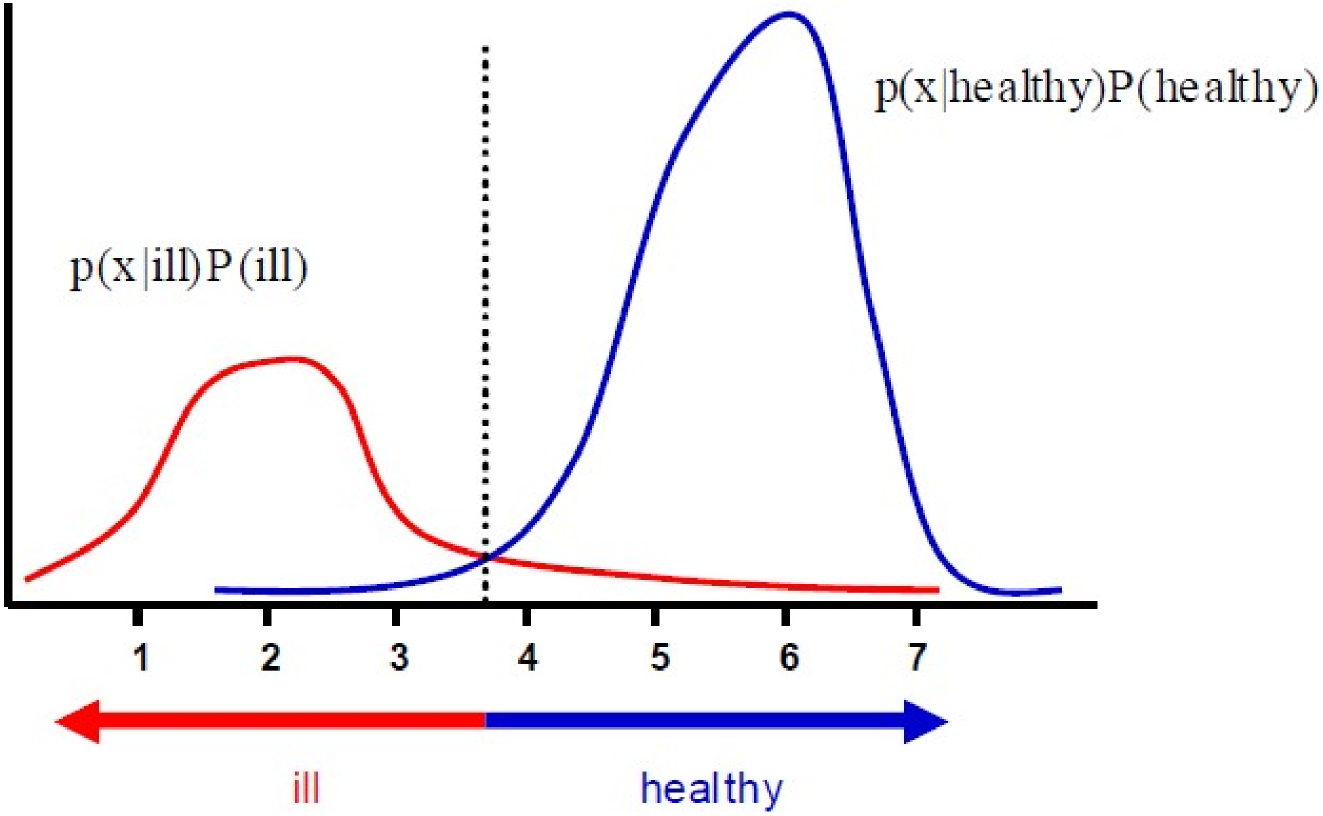
$$p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)}$$

- Recall:
  - ✓ $p(ill) = 0.15$, $p(healthy) = 0.85$
  - ✓ $p(x = 2|ill) = 0.67, p(x = 2|healthy) = 0.05$
- $p(\omega|x = 2) \propto p(x = 2|\omega) \times p(\omega)$
  - ✓ $p(healthy|x = 2) \propto 0.05 \times 0.85 = 0.0425$
  - ✓ $p(ill|x = 2) \propto 0.67 \times 0.15 = 0.1005$
- According to Bayes Decision Rule

$$p(ill|x = 2) > p(healthy|x = 2),$$

therefore, Peter was ill.

Decision Boundary



Probability of Error

- There are two possible errors

True	Decision: Class 1	Decision: Class 2
Class 1	Correct	Error
Class 2	Error	Correct

- Probability of error is

$p(error|x) = p(\omega_1|x)$ if we decide ω_2

$p(error|x) = p(\omega_2|x)$ if we decide ω_1

Classification Error

- Bayes rule or Bayes classifier
  - ✓ If $p(\omega_1|x) > p(\omega_2|x)$, decide ω_1
  - ✓ Otherwise, decide ω_2

- Therefore, the classification error is

$$p(error|x) = min[p(\omega_1|x), p(\omega_2|x)]$$

Extension

- Extend to

  Multi-dimensional features

$$x \rightarrow X$$

  Multi-class problem (3 or more classes)

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

- Bayes rule or Bayes classifier for multi-class problems

  Select ω_i if $p(\omega_i|x) > p(\omega_j|x)$ for all $j \neq i$

  Probability of error for multi-class problems

$$p(error|x) = 1 - max[p(\omega_1|x), p(\omega_2|x), \ldots, p(\omega_c|x)]$$

Bayes Error Rate

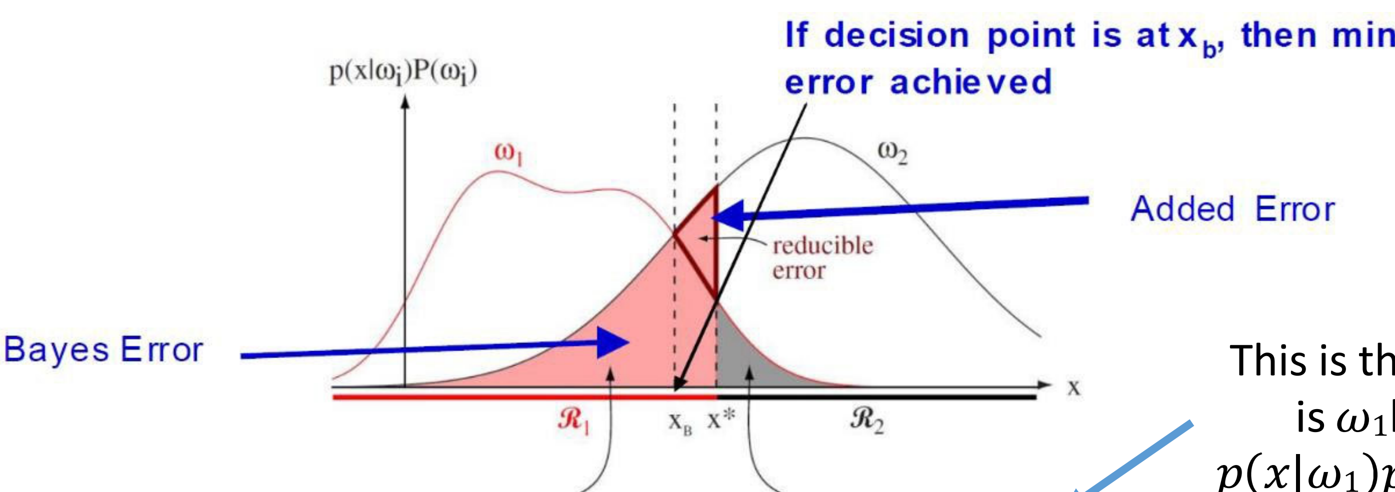
- Error = Bayes Error + Added Error

$$p(error) = p(x \in R_2, \omega_1) + p(x \in R_1, \omega_2)$$

$$= p(x \in R_2|\omega_1)p(\omega_1) + p(x \in R_1|\omega_1)p(\omega_2)$$

$$= \quad p(x|\omega_1)p(\omega_1)dx + \quad p(x|\omega_2)p(\omega_2)dx$$

R_2 R_1



This is the error since true state is $\omega_1$but decide ω_2 since $p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$

$p(x|\omega_1)p(\omega_1)dx$ $p(x|\omega_2)p(\omega_2)dx$

R_2 R_1

Cost Consideration

- **Salmon/Sea bass Illustration**

  Costs of different errors should be considered

  Case 1: Company's view

    Salmon is more expensive than sea bass. Selling salmon with the price of sea bass will be a loss

  Case 2: Customer's view

    Customers who buy salmon will be very upset if they get sea bass

Loss Function

- Measures the cost for each action taken
- Convert a probability determination into a decision
- Let

  $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c categories

  $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of a possible actions

  $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ be the loss incurred for taking action α_i when the category is ω_j.

Loss Function

- Given observation x, if the true state is ω_j and we take action α_i, then loss is λ_{ij}

The expected loss associated with taking action α_i is Conditional risk (expected loss of taking action α_i)

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)\, p(\omega_j|x)$$

And overall risk R (expected loss)

$$R = \int R(a(x)|x)p(x)dx$$

Loss Function

- Bayes decision procedure provides the optimal performance on an overall risk

  Choose α_i so that $R(\alpha_i|x)$ is as small as possible for every x, then the overall risk will be minimized.

  So compute $R(\alpha_i|x)$ for all i and select the action α_i for which $R(\alpha_i|x)$ is the minimum.

- R in this case is called the Bayes Risk

Loss Function

- For example, a two-class problem:

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

Where $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ be the loss incurred for taking action α_i when the category ω_j.

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

Minimum Risk Decision Rule

- The fundamental rule is to decide ω_1 if

$$R(\alpha_1|x) < R(\alpha_2|x)$$

- This result in the equivalent rule:

Decide ω_1 if

$$(\lambda_{21} - \lambda_{11})p(x|\omega_1)p(\omega_1) > (\lambda_{12} - \lambda_{22})p(x|\omega_2)p(\omega_2)$$

Otherwise, decide ω_2

Minimum Risk Decision Rule

- If we assume $\lambda_{21} > \lambda_{11}$, the preceding rule is equivalent to the following rule:

If $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{p(\omega_2)}{p(\omega_1)}$, then take action α_1 (decide ω_1)

Otherwise, take action α_2 (decide ω_2)

Minimum Risk Decision Rule

- Example:

$$\lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 1 & 5 \\ 10 & 3 \end{pmatrix}$$

If $p(\omega_1|x) = 0.1$ and $p(\omega_2|x) = 0.9$

$$R(\alpha_1|x) = 1 \times 0.1 + 5 \times 0.9 = 4.6$$

$$R(\alpha_2|x) = 10 \times 0.1 + 3 \times 0.9 = 3.7$$

Then the action α_2 is selected

If $p(\omega_1|x) = 0.8$ and $p(\omega_2|x) = 0.2$

$$R(\alpha_1|x) = 1 \times 0.8 + 5 \times 0.2 = 1.8$$

$$R(\alpha_2|x) = 10 \times 0.8 + 3 \times 0.2 = 8.6$$

Then the action α_1 is selected

Minimum-Error-Rate Classification

- Assume the following situation

  If

  The action α_i is taken

  The true state of nature is ω_j

  Then

  The decision is correct if $i = j$ (i.e., error=0)

  Otherwise

  The decision is wrong if $i \neq j$ (i.e., error=1)

Minimum-Error-Rate Classification

- Zero-one loss function

$$\lambda(\alpha_i|\omega_j) = \begin{matrix} 0 & i = j \\ 1 & i \neq j \end{matrix} \quad i, j = 1, 2, \dots, c$$

✓ No loss to a correct decision

$$\lambda = 0$$

✓ A unit loss to any error

$$\lambda = 1$$

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

All errors are equally costly

Minimum-Error-Rate Classification

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

- Risk:

$$R(\alpha_i|x) = \quad {}_{j=1}^{c} \lambda(\alpha_i|\omega_j)\, p(\omega_j|x)$$

$$= \quad {}_{j \neq i}\, p(\omega_j|x)$$

$$= 1 - p(\omega_i|x)$$

Note: zero-one loss function is used here

Minimizing the average probability of error requires maximizing $p(\omega_i|x)$

For Minimum error rate

✓ Decide ω_i if $p(\omega_i|x) > p(\omega_j|x), \forall j \neq i$

Minimum-Error-Rate Classification

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{p(\omega_2)}{p(\omega_1)}$$

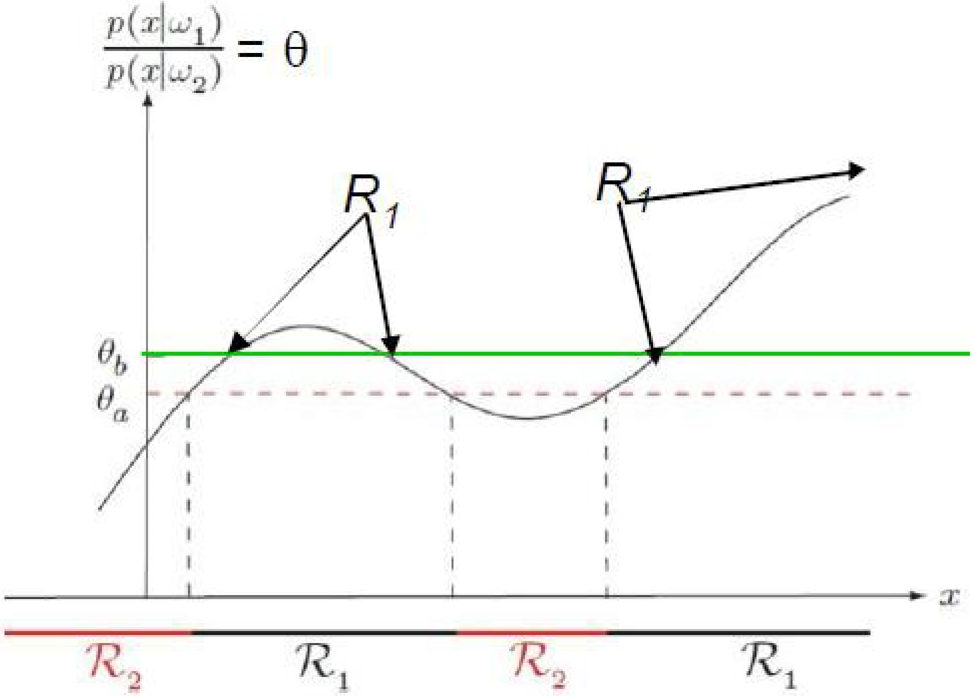
$$\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{p(\omega_2)}{p(\omega_1)} = \theta_\lambda$$

Other loss function $\lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$

$$\theta_b = \frac{2p(\omega_2)}{p(\omega_1)}$$

Zero-one loss function $\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

$$\theta_a = \frac{p(\omega_2)}{p(\omega_1)}$$



If loss function penalizes mis-categorizing ω_2 as ω_1 ($\lambda_{12} = 2$) more than the converse (penalize mis-categorizing ω_1 as ω_2 ($\lambda_{21} = 1$))
We get larger threshold $\theta_b > \theta_a$
Hence R_1 becomes smaller

Minimum-Error-Rate Classification

Salmon/Sea bass Illustration

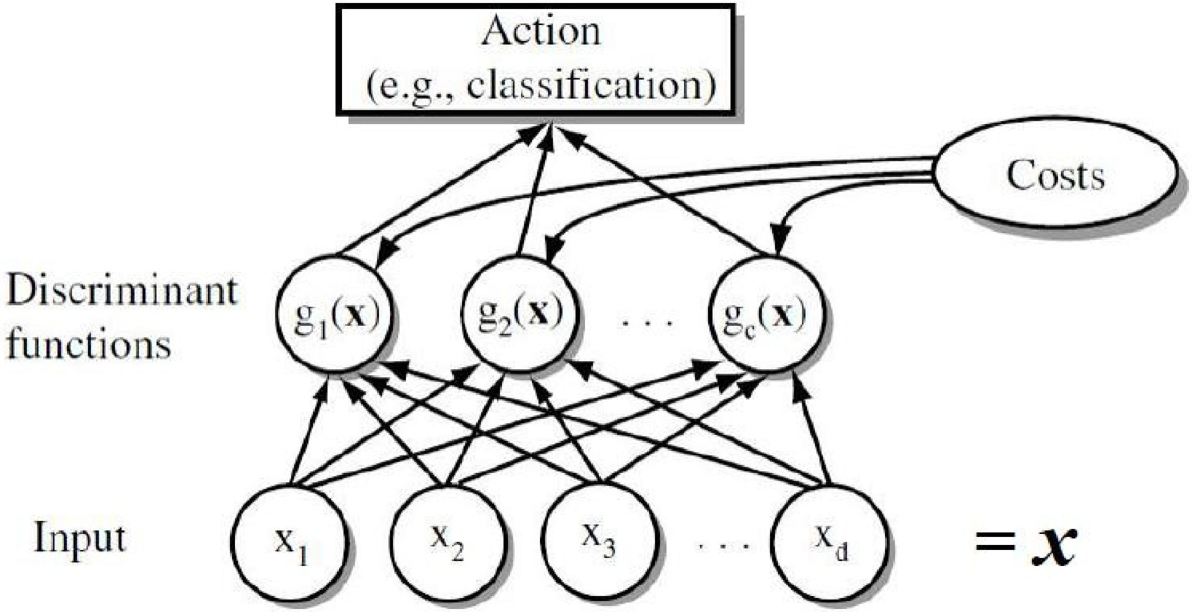
- Risk R_1: misclassifying salmon as sea bass
- Risk R_2: misclassifying sea bass as salmon
- In last example, we penalize misclassifying salmon as sea bass ($\lambda_{12} = 2$) more than the converse (penalize misclassifying sea bass as salmon ($\lambda_{21} = 1$)). This is from the view point of the fish producer. Hence the risk R_1 would become smaller
- Recall that $R_1 = R_2$ if we make $\lambda_{12} = \lambda_{21}$

Classifiers

- One of the most useful ways to represent classifier is in terms of a set of discriminant functions:

$$g_i(x), i = 1, \ldots, c$$

- x will be assigned to class ω_i if $g_i(x) > g_j(x)$ for all $j \neq i$



Discrimination Function

- For Minimum-Error-Rate

$$g_i(x) = p(\omega_i|x)$$

- Choice of discriminant functions not unique!

If G is a monotonically increasing function, then

$G(g_i(x)) > G(g_j(x))$ if $g_i(x) > g_j(x)$ for all $j \neq i$

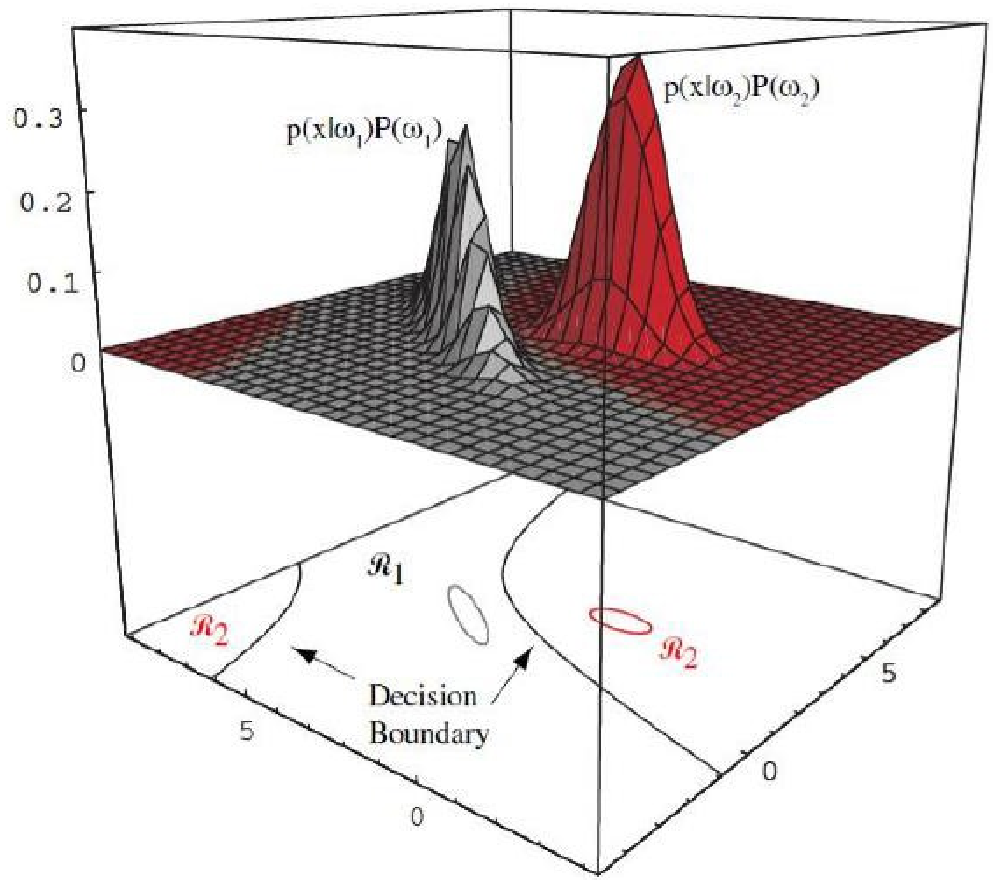
Example, the natural log function

$$G(g_i(x)) = \ln(g_i(x))$$
$$= \ln\big(p(x|\omega_i)\big) + \ln\big(p(\omega_i)\big) - \ln(p(x))$$

Decision Surfaces

- Feature space divided into c decision regions
- If $g_i(x) > g_j(x)\ \forall j \neq i$ then x is in R_i
- The decision boundary consists of two hyperbolas (p being Gaussian)

  The decision region R_2 is not simply connected
- Ellipses mark where density is $1/e$ times that of peak distribution



Two-Class Problem

- A classifier that places a pattern in one of only two classes

  It has two discriminant functions g_1 and g_2

  Re-define a single discriminant function

$$g(x) = g_1(\mathrm{x}) - g_2(\mathrm{x})$$

  Decide ω_1, if $g(x) > 0$;

  Otherwise, decide ω_2

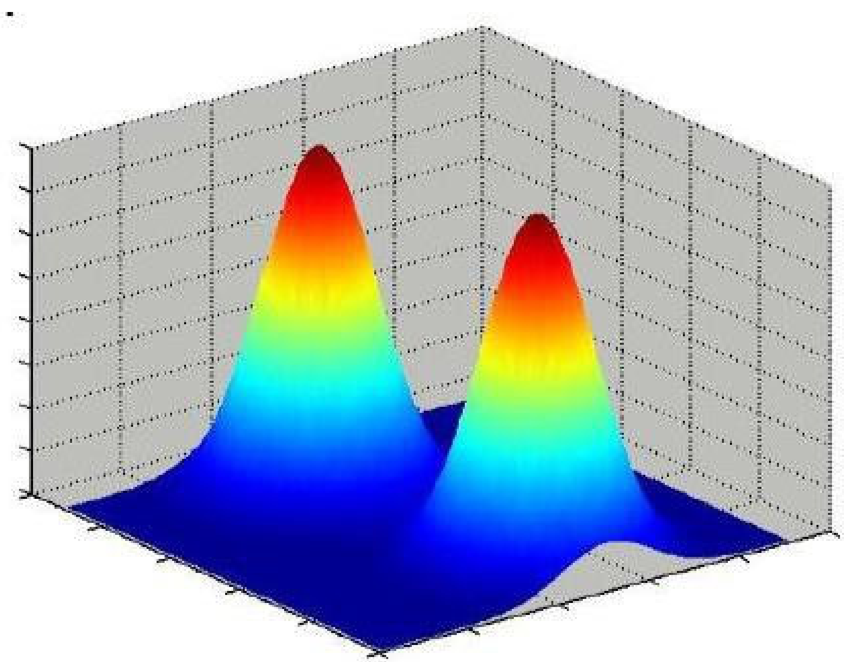
  Two commonly used minimum-error-rate discriminant functions:

$$g(x) = p(\omega_1|x) - p(\omega_2|\mathrm{x})$$

$$\text{or } g(x) = ln\frac{p(x|\omega_1)}{p(x|\omega_2)} + ln\frac{p(\omega_1)}{p(\omega_2)}$$

Special Case: Normal Distribution

- Advantages of Normal Distribution:
  - ✓ Analytically tractable, continuous
  - ✓ A lot of processes are asymptotically Gaussian
  - ✓ Handwritten characters, speech sounds are ideal or prototype corrupted by random process



Normal Distribution

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

- Probability that X falls into the interval (a, b):

$${}_a^b\, p(x)dx$$

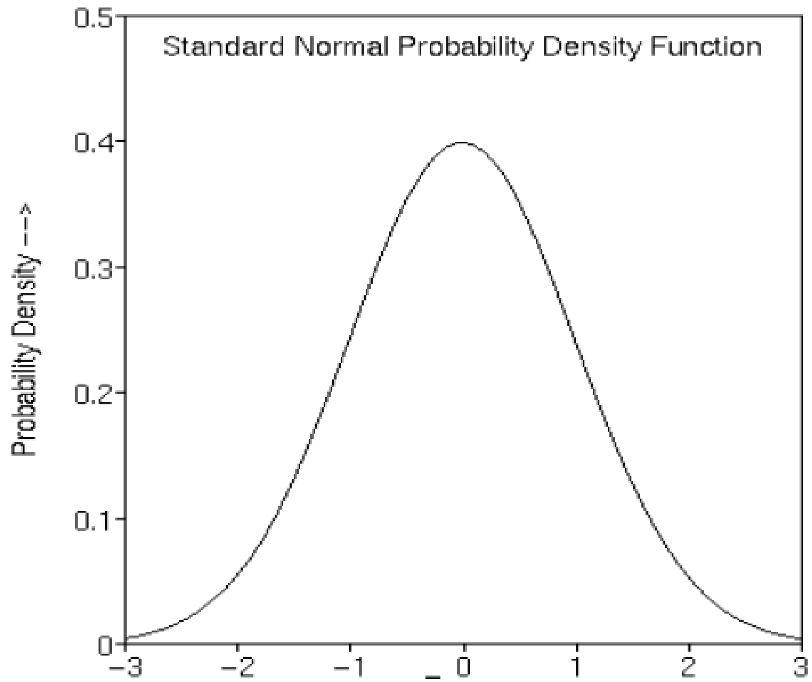
- Mean:

$$\mu = {}_{-\infty}^{\infty}\, xp(x)dx$$

- Variance:

$$\sigma^2 = {}_{-\infty}^{\infty}\, (x-\mu)^2 p\ (x)dx$$

- Standard Deviation: σ



Normal Distribution

- Multivariate Normal Density in d dimensions is:

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

Where

$$x = (x_1, x_2, \ldots, x_d)^T$$

$$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^T$$

$$\Sigma = \quad (x-\mu)(x-\mu)^T p(x)dx$$

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

Discriminant Functions for the Normal Density

- Recall, a convenient discriminant function

$$g_i(x) = ln\big(p(x|\omega_i)\big) + ln(p(\omega_i))$$

For the case of multivariate normal

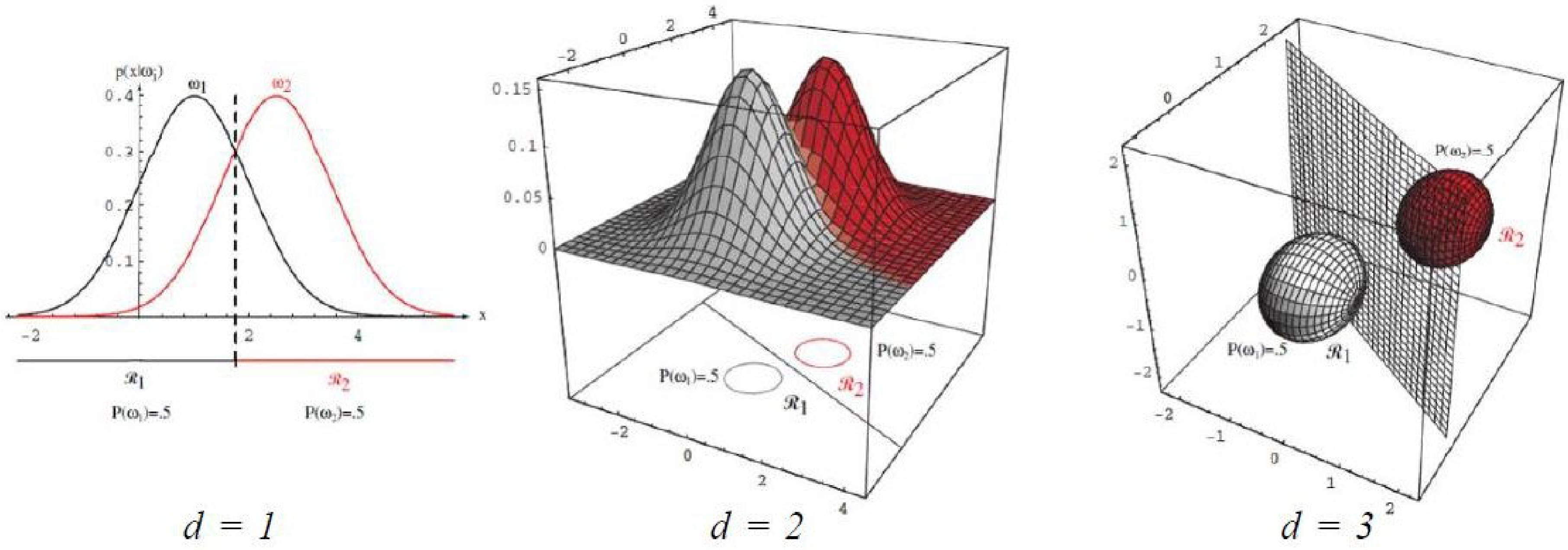
$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_i|) + \ln\big(p(\omega_i)\big)$$

If $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$

Discriminant Functions for Multivariate Normal Density

- Case $\Sigma_i = \sigma^2 I$ (I stands for the identity matrix)

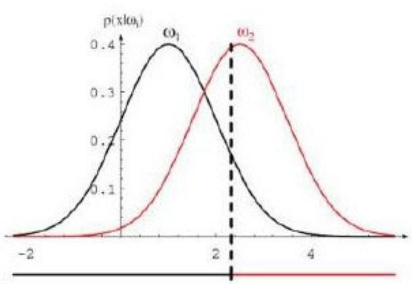
  Examples ($p(\omega_1) = 0.5, p(\omega_2) = 0.5$)



Discriminant Functions for Multivariate Normal Density

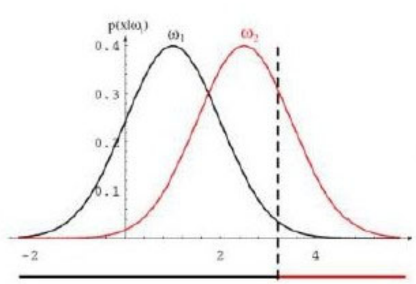
- Influence of Priors on Decision Boundaries
- As priors change, the decision boundary shifts. For sufficiently disparate priors the boundary will not lie between the means

$p(\omega_1) = 0.7$
$p(\omega_2) = 0.3$

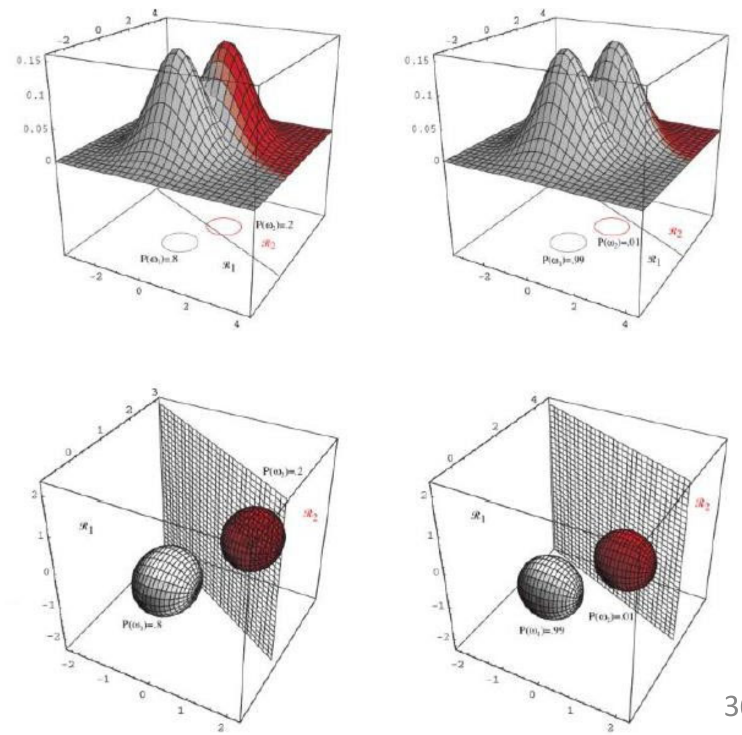


$p(\omega_1)$ $p(\omega_2)$

$p(\omega_1) = 0.9$
$p(\omega_2) = 0.1$



$p(\omega_1)$ $p(\omega_2)$



Discriminant Functions for Multivariate Normal Density

- Case $\Sigma_i = \Sigma$

  Covariance of all classes are identical but arbitrary

  Geometrically, the samples fall in hyper-ellipsoidal clusters of equal size and shape

  The discriminant function:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) + ln p(\omega_i)$$