Chapter 2
Bayesian Decision Theory
Bayesian Decision Theory

- Bayesian decision theory is a statistical approach to data mining/pattern recognition

- Mathematical foundation for decision making

- Using probabilistic approach to help making decision so as to minimize the risk (cost).
Bayesian Decision Theory

- Basic Assumptions
  - The decision problem is posed (formalized) in **probabilistic terms**
  - All the relevant probability values are known

- Key Principle
  - Bayes Theorem
Preliminaries and Notations

\[ \omega_i \in \{\omega_1, \omega_2, \ldots, \omega_c\} : \text{a state of nature} \]

\[ P(\omega_i) : \text{prior probability} \]

\[ x : \text{feature vector} \]

\[ p(x) : \text{evidence probability} \]

\[ p(x | \omega_i) : \text{class-conditional density / likelihood} \]

\[ P(\omega_i | x) : \text{posterior probability} \]
Decision Before Observation

The Problem
- To make a decision where
  - Prior probability is known
  - No observation is allowed

Naïve Decision Rule

Decide $\omega_1$ if $P(\omega_1) > P(\omega_2)$, otherwise $\omega_2$

This is the best we can do without observation
Fixed prior probabilities -> Same decisions all time
Bayes Theorem

\[ P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)} \]

\[ p(x) = \sum_{j=1}^{c} p(x \mid \omega_i)P(\omega_i) \]

Thomas Bayes
(1702-1761)
Decision After Observation

\[ P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)} \]

\[ D(x) = \arg \max_{\omega_i} P(\omega_i \mid x) \]

unimportant in making decision
Decision After Observation

\[ P(\omega_i | x) = \frac{p(x | \omega_i) P(\omega_i)}{p(x)} \]

**Known**
- Prior probability
- Class-conditional pdf
- Observation

**Unknown**
- Posterior probability

\[ P(\omega_i | x) : \]
Special Cases

\[ P(\omega_i | x) = \frac{p(x | \omega_i)P(\omega_i)}{p(x)} \] (posterior = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}})

- Case I: Equal prior probability
  - \( P(\omega_1) = P(\omega_2) = \cdots = P(\omega_c) = 1/c \)
  - Depends on the likelihood \( p(x | \omega_j) \)

- Case II: Equal likelihood
  - \( p(x | \omega_1) = p(x | \omega_2) = \cdots = p(x | \omega_c) \)
  - Degenerate to naïve decision rule

- Normally, prior probability and likelihood function together in Bayesian decision process
An example

$\omega_1$: sea bass

$\omega_2$: salmon

$P(\omega_1) = \frac{2}{3}$

$P(\omega_2) = \frac{1}{3}$

What will the posterior probability for either type of fish look like?

class-conditional pdf for \textit{lightness}

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

h-axis: lightness of fish scales
v-axis: posterior probability for each type of fish

Black curve: sea bass
Red curve: salmon

- For each value of x, the higher curve yields the output of Bayesian decision
- For each value of x, the posteriors of either curve sum to 1.0
Another Example

Problem statement

- A new medical test is used to detect whether a patient has a certain cancer or not, whose test result is either + (positive) or – (negative)
- For patient with this cancer, the probability of returning positive test result is 0.98
- For patient without this cancer, the probability of returning negative test result is 0.97
- The probability for any person to have this cancer is 0.008

Question

- If positive test result is returned, does she/he have cancer?
Another Example (Cont.)

\( \omega_1 : \text{cancer} \quad \omega_2 : \text{no cancer} \quad x \in \{+, -\} \)

\[
P(\omega_1) = 0.008 \quad \quad P(\omega_2) = 1 - P(\omega_1) = 0.992
\]

\[
P(+) | \omega_1 = 0.98 \quad \quad P(- | \omega_1) = 1 - P(+) | \omega_1 = 0.02
\]

\[
P(- | \omega_2) = 0.97 \quad \quad P(+) | \omega_2 = 1 - P(- | \omega_2) = 0.03
\]

\[
P(\omega_1 | +) = \frac{P(\omega_1)P(+ | \omega_1)}{P(+)} = \frac{P(\omega_1)P(+ | \omega_1)}{P(\omega_1)P(+ | \omega_1) + P(\omega_2)P(+ | \omega_2)}
\]

\[
= \frac{0.008 \times 0.98}{0.008 \times 0.98 + 0.992 \times 0.03} = 0.2085
\]

\[
P(\omega_2 | +) = 1 - P(\omega_1 | +) = 0.7915
\]

\[
P(\omega_2 | +) > P(\omega_1 | +) \quad \quad \text{No cancer!}
\]
Feasibility of Bayes Formula

\[ P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)} \quad \left( \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \right) \]

- To compute posterior probability, we need to know prior probability and likelihood

How do we know these probabilities?

- A simple solution: Counting Relative frequencies
- An advanced solution: Conduct Density estimation
A Further Example

Problem

Based on the height of a car in some campus, decide whether it costs more than $50,000 or not

\[ \omega_1 : \text{price} > $50,000 \]
\[ \omega_2 : \text{price} \leq $50,000 \]
\[ x : \text{height of a car} \]

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

Quantities to know:

\[ P(\omega_1) \quad P(\omega_2) \quad P(x | \omega_1) \quad P(x | \omega_2) \]

How to get them?

Counting relative frequencies via collected samples
A Further Example (Cont.)

- Collecting samples
  - Suppose we have randomly picked 1209 cars in the campus, got prices from their owners, and measured their heights

- Compute $P(\omega_1)$ and $P(\omega_2)$

$\# \text{ cars in } \omega_1 : 221$

$\# \text{ cars in } \omega_2 : 988$

$P(\omega_1) = \frac{221}{1209} = 0.183$

$P(\omega_2) = \frac{988}{1209} = 0.817$
A Further Example (Cont.)

- Compute $P(x|\omega_1)$, $P(x|\omega_2)$
  - Discretize the height spectrum (say [0.5m, 2.5m]) into 20 intervals each with length 0.1m, and then count the number of cars falling into each interval for either class.

- Suppose $x = 1.05$, which means that $x$ falls into interval $I_x = [1.0m, 1.1m]$

  For $\omega_1$, # cars in $I_x$ is 46,
  For $\omega_2$, # cars in $I_x$ is 59,

  $P(x = 1.05 | \omega_1) = \frac{46}{221} = 0.2081$
  $P(x = 1.05 | \omega_2) = \frac{59}{988} = 0.0597$
A Further Example (Cont.)

Question

For a car with height 1.05m, is its price greater than $50,000?

\[
P(\omega_1) = \frac{221}{1209} = 0.183
\]

\[
P(\omega_2) = \frac{988}{1209} = 0.817
\]

\[
P(x = 1.05 | \omega_1) = \frac{46}{221} = 0.2081
\]

\[
P(x = 1.05 | \omega_2) = \frac{59}{988} = 0.0597
\]

\[
P(\omega_2 | x = 1.05) = \frac{P(\omega_2)P(x = 1.05 | \omega_2)}{P(\omega_1)P(x = 1.05 | \omega_1)} = \frac{0.817 \times 0.0597}{0.183 \times 0.2081}
\]

\[
P(\omega_1 | x) < P(\omega_2 | x), \quad \text{price} \leq $50,000
\]
Is Bayes Decision Rule Optimal

Consider two categories

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

When we observe $x$, the probability of error is:

$$P(\text{error} | x) = \begin{cases} P(\omega_2 | x) & \text{if we decide } \omega_1 \\ P(\omega_1 | x) & \text{if we decide } \omega_2 \end{cases}$$

Thus, under Bayes decision rule, we have

$$P(\text{error} | x) = \min[P(\omega_1 | x), P(\omega_2 | x)]$$

For every $x$, we ensure that $P(\text{error} | x)$ is as small as possible
Is Bayes Decision Rule Optimal

- Consider two categories
  - Decide $\omega_1$ if $P(\omega_1 | x) > P(\omega_2 | x)$
  - When we observe $x$, the probability of error is:
    \[
P(error | x) = \begin{cases} 
    P(\omega_2 | x) & \text{if we decide } \omega_1 \\
    P(\omega_1 | x) & \text{if we decide } \omega_2
    \end{cases}
    \]

Thus, under Bayes decision rule, we have

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

For every $x$, we ensure that $P(error|x)$ is as small as possible
Generalized Bayes Decision Rule

- Allowing to use more than one feature

\[ x \in \mathbb{R} \implies x \in \mathbb{R}^d : \text{d-dimensional Euclidean Space} \]

- Allowing more than two states of nature

\[ \Omega = \{\omega_1, \omega_2, \ldots, \omega_c\} : \text{a set of } c \text{ states of nature} \]

- Allowing actions other than merely deciding the state of nature

\[ A = \{\alpha_1, \alpha_2, \ldots, \alpha_a\} : \text{a set of } a \text{ possible actions} \]

Note that \( c \neq a \)
Generalized Bayes Decision Rule (cont.)

- Introducing a loss function more general than the probability of error

\[ \lambda : \Omega \times A \rightarrow R \] (loss function)

\[ \lambda_{ij} = \lambda(\omega_j, \alpha_i) : \text{the loss incurred for taking action } \alpha_i \]
when the state of nature is \( \omega_j \)

For ease of reference, it is usually written as:

\[ \lambda_{ij} = \lambda(\alpha_i \mid \omega_j) : \]

We want to minimize the expected loss in making decision.
Introducing a loss function more general than the probability of error function ($\lambda : \mathcal{R}_A \times \mathcal{\Omega} \rightarrow \mathcal{\Omega}$) is nature of state the when action for taking incurred loss ($\alpha$, $\omega$).

For ease of reference, it is usually written as:

$$\lambda_{ij} = \lambda(\omega_j, \alpha_i) : \text{the loss incurred for taking action } \alpha_i \text{ when the state of nature is } \omega_j$$

We want to minimize the expected loss in making decision.
Generalized Bayes Decision Rule (cont.)

<table>
<thead>
<tr>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Given a particular x, we have to decide which action to take</td>
</tr>
</tbody>
</table>

To do this, we need to know the loss of taking each action \( \alpha_i \) \( (1 \leq i \leq a) \)

\[
\lambda_{ij} = \lambda(\alpha_i | \omega_j): \\
\text{The action being taken } \alpha_i \quad \text{True state of nature } \omega_j
\]

We want to minimize the expected loss in making decision.
Generalized Bayes Decision Rule (cont.)

Expected loss:

\[ R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)P(\omega_j \mid x) = \sum_{j=1}^{c} \lambda_{ij}P(\omega_j \mid x) \]

- The incurred loss of taking action \( \alpha_i \) in case of true state of nature being \( \omega_j \)
- The probability of \( \omega_j \) being the true state of nature

The expected loss is also named as “conditional risk”
Suppose we have:

<table>
<thead>
<tr>
<th>Class</th>
<th>Action</th>
<th>$\alpha_1 = \text{“Recipe A”}$</th>
<th>$\alpha_2 = \text{“Recipe B”}$</th>
<th>$\alpha_3 = \text{“No Recipe”}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega_1 = \text{“cancer”}$</td>
<td>5</td>
<td>50</td>
<td>10,000</td>
<td></td>
</tr>
<tr>
<td>$\omega_2 = \text{“no cancer”}$</td>
<td>60</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

For a particular $x$:

$P(\omega_1 | x) = 0.01$

$P(\omega_2 | x) = 0.99$

$$R(\alpha_1 | x) = \sum_{j=1}^{2} \lambda(\alpha_1 | \omega_j) \cdot P(\omega_j | x)$$

$$= \lambda(\alpha_1 | \omega_1) \cdot P(\omega_1 | x) + \lambda(\alpha_1 | \omega_2) \cdot P(\omega_2 | x)$$

$$= 5 \times 0.01 + 60 \times 0.99 = 59.45$$

Similarly, we can get:

$R(\alpha_2 | x) = 3.47$

$R(\alpha_3 | x) = 100$
Generalized Bayes Decision Rule (cont.)

- 0/1 Loss Function

\[ R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)P(\omega_j \mid x) = \sum_{j=1}^{c} \lambda_{ij}P(\omega_j \mid x) \]

\[ \lambda(\alpha_i \mid \omega_j) = \begin{cases} 
0 & \text{if } \alpha_i \text{ is a correct decision associated with } \omega_j \\
1 & \text{otherwise} 
\end{cases} \]

\[ R(\alpha_i \mid x) = P(error \mid x) \]
Generalized Bayes Decision Rule (cont.)

- Bayes decision rule (general case)

\[
\alpha(x) = \arg \min_{\alpha_i \in A} R(\alpha_i \mid x) = \arg \min_{\alpha_i \in A} \sum_{j=1}^c \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)
\]

- Overall risk

\[
R = \int R(\alpha(x) \mid x) \cdot p(x) dx
\]

For every \( x \), we ensure that the conditional risk \( R(a(x) \mid x) \) is as small as possible; Thus, the overall risk over all possible \( x \) must be as small as possible.

The optimal one to minimize the overall risk. Its resulting overall risk is called the Bayesian risk.
General Case: Two-Category

\[ \Omega = \{\omega_1, \omega_2\} \]
\[ A = \{\alpha_1, \alpha_2\} \]

<table>
<thead>
<tr>
<th>State of Nature</th>
<th>( \omega_1 )</th>
<th>( \omega_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>( \alpha_1 )</td>
<td>( \lambda_{11} )</td>
</tr>
<tr>
<td></td>
<td>( \alpha_2 )</td>
<td>( \lambda_{21} )</td>
</tr>
</tbody>
</table>

\[
R(\alpha_1 \mid x) = \lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x) 
\]
\[
R(\alpha_2 \mid x) = \lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x) 
\]
General Case: Two-Category

Perform \( \alpha_1 \) if \( R(\alpha_2 | x) > R(\alpha_1 | x) \); otherwise perform \( \alpha_2 \)

\[ \lambda_{21} P(\omega_1 | x) + \lambda_{22} P(\omega_2 | x) > \lambda_{11} P(\omega_1 | x) + \lambda_{12} P(\omega_2 | x) \]

\[ (\lambda_{21} - \lambda_{11}) P(\omega_1 | x) > (\lambda_{12} - \lambda_{22}) P(\omega_2 | x) \]

\[ R(\alpha_1 | x) = \lambda_{11} P(\omega_1 | x) + \lambda_{12} P(\omega_2 | x) \]

\[ R(\alpha_2 | x) = \lambda_{21} P(\omega_1 | x) + \lambda_{22} P(\omega_2 | x) \]
General Case: Two-Category

Perform $\alpha_1$ if $R(\alpha_2|x) > R(\alpha_1|x)$; otherwise perform $\alpha_2$

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

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

*positive*  *positive*

_Posterior probabilities are scaled before comparison._
General Case: Two-Category

Perform $\alpha_1$ if $R(\alpha_2|x) > R(\alpha_1|x)$; otherwise perform $\alpha_2$

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

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

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

$\frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \frac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{(\lambda_{21} - \lambda_{11})P(\omega_1)}$
Perform \( \alpha_1 \) if

\[
\frac{p(x | \omega_1)}{p(x | \omega_2)} > \frac{(\lambda_{12} - \lambda_{22}) P(\omega_2)}{(\lambda_{21} - \lambda_{11}) P(\omega_1)}
\]
Discriminant Function

- Discriminant functions for multicategory

\[ g_i(x) : R^d \rightarrow R \quad (1 \leq i \leq c) \]

- One function per category

\[
\begin{align*}
&g_1(x) \\
g_2(x) \\
\vdots \\
g_c(x) \\
\end{align*}
\]

Action (e.g., classification)

Assign \( x \) to \( \omega_i \) if

\[ g_i(x) > g_j(x) \quad \text{for all} \ j \neq i. \]
 Discriminant Function

- Minimum Risk Case:
  \[ g_i(x) = -R(\alpha_i | x) \]

- Minimum Error-Rate Case:
  \[ g_i(x) = P(\omega_i | x) \]
  \[ g_i(x) = p(x | \omega_i)P(\omega_i) \]
  \[ g_i(x) = \ln p(x | \omega_i) + \ln P(\omega_i) \]
Discriminant Function

- Relationship between minimum risk and minimum error rate

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

\[
= \sum_{j \neq i} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid x) + \lambda(\alpha_i \mid \omega_i) \cdot P(\omega_i \mid x)
\]

\[
= \sum_{j \neq i} P(\omega_j \mid x)
\]

\[
= 1 - P(\omega_i \mid x)
\]

**Error Rate** (误差率/错误率)

the probability that action \( \alpha_i \) (decide \( \omega_i \)) is wrong
Discriminant Function

- Various discriminant function
- Identical classification results

If $f(.)$ is a monotonically increasing function, then $f(g_i(.))$’s are also be discriminant functions.

Example

\[ f(x) = k \cdot x \quad (k > 0) \quad \Rightarrow \quad f(g_i(x)) = k \cdot g_i(x) \quad (1 \leq i \leq c) \]

\[ f(x) = \ln x \quad \Rightarrow \quad f(g_i(x)) = \ln g_i(x) \quad (1 \leq i \leq c) \]
Decision Regions

- $c$ discriminant functions result in $c$ decision regions.

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

where $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ ($i \neq j$) and $\bigcap_{i=1}^{c} \mathcal{R}_i = \mathcal{R}^d$

- Decision boundary
  - Decision regions are separated by decision boundaries

*Two-category example*
The Normal Distribution

Discrete random variable \((X)\) — Assume integer

Probability mass function (pmf):
\[
p(x) = P(X = x)
\]

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

Continuous random variable \((X)\)

Probability density function (pdf):
\[
p(x) \text{ or } f(x) \quad \text{not a probability}
\]

Cumulative distribution function (cdf):
\[
F(x) = P(X \leq x) = \int_{-\infty}^{x} p(t)dt
\]
Expectations

- a.k.a. expected value, mean or average of a random variable
- $x$ is a random variable, the expectation of $x$

\[
E[x] = \begin{cases} 
\sum_{x=-\infty}^{\infty} xp(x) & \text{x is discrete} \\
\int_{-\infty}^{\infty} xp(x)dx & \text{x is continuous}
\end{cases}
\]

The $k^{th}$ moment \( E[X^k] \)

The $1^{st}$ moment \( \mu_X = E[X] \)

The $k^{th}$ central moment \( E[(X - \mu_X)^k] \)
Important Expectations

- **Mean**

\[
\mu_X = E[X] = \begin{cases} 
\sum_{x=-\infty}^{\infty} xp(x) & X \text{ is discrete} \\
\int_{-\infty}^{\infty} xp(x) dx & X \text{ is continuous}
\end{cases}
\]

- **Variance**

\[
\sigma_X^2 = Var[X] = E[(X - \mu_X)^2] = \begin{cases} 
\sum_{x=-\infty}^{\infty} (x - \mu_X)^2 p(x) & X \text{ is discrete} \\
\int_{-\infty}^{\infty} (x - \mu_X)^2 p(x) dx & X \text{ is continuous}
\end{cases}
\]

**Notation:** \( \sigma^2 = Var[x] \) (\( \sigma \): standard deviation ?)

**Fact:** \( \sigma^2 = Var[x] = E[x^2] - (E[x])^2 \)
The entropy measures the fundamental uncertainty in the value of points selected randomly from a distribution.

\[
H[X] = \begin{cases} 
- \sum_{x=-\infty}^{\infty} p(x) \log p(x) & X \text{ is discrete} \\
- \int_{-\infty}^{\infty} p(x) \log p(x) \, dx & X \text{ is continuous}
\end{cases}
\]
Univariate Gaussian Distribution

Gaussian distribution, a.k.a. Gaussian density, normal density.

\[ X \sim N(\mu, \sigma^2) \]

\[ p(x) = \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ E[X] = \mu \]

\[ \text{Var}[X] = \sigma^2 \]
Univariate Gaussian Distribution

Gaussian distribution, a.k.a. Gaussian density, normal density.

\[ X \sim \mathcal{N}(\mu, \sigma^2) \]

\[ \mu = 0, \quad \sigma^2 = 0.2, \]
\[ \mu = 0, \quad \sigma^2 = 1.0, \]
\[ \mu = 0, \quad \sigma^2 = 5.0, \]
\[ \mu = -2, \quad \sigma^2 = 0.5, \]

\( \mathbb{E}[X] = \mu \)
\( \text{Var}[X] = \sigma^2 \)

\[ \varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Random Vectors

- A d-dimensional random vector is:

\[ X = (x_1, x_2, \ldots, x_d)^T \quad X : \Omega \rightarrow \mathbb{R}^d \]

\[ X \sim p(X) = p(x_1, x_2, \ldots, x_d) \quad \text{(joint pdf)} \]

- Expected vector

\[ E[X] = \begin{pmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{pmatrix} \quad E[x_i] = \int_{-\infty}^{+\infty} x_i p(x_i) dx_i \quad (1 \leq i \leq d) \]

\[ \mathbf{\mu} = E[X] = (\mu_1, \mu_2, \ldots, \mu_d)^T \]

Marginal pdf on the \(i^{th}\) component.
Random Vectors

- Covariance matrix

\[ \Sigma = E[(X - \mu)(X - \mu)^T] = \begin{pmatrix}
\sigma_1^2 & \sigma_{12} & \Lambda & \sigma_{1d} \\
\sigma_{21} & \sigma_2^2 & \Lambda & \sigma_{2d} \\
M & M & O & M \\
\sigma_{d1} & \sigma_{d2} & \Lambda & \sigma_d^2
\end{pmatrix} \]

\[ \sigma_{ij} = \sigma_{ji} = E[(x_i - \mu_i)(x_j - \mu_j)] = \int_{-\infty}^{+\infty} (x_i - \mu_i)(x_j - \mu_j) p(x_i, x_j) dx_i dx_j \]

Properties:
Symmetric, Positive semidefinite

Marginal pdf on a pair of random variables \((x_i, x_j)\)
Multivariate Gaussian Distribution

- $X$ is a $d$-dimensional random vector

$$X \sim N(\mu, \Sigma)$$

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

$$E[X] = \mu$$

$$E[(X - \mu)(X - \mu)^T] = \Sigma$$
Properties of $N(\mu, \Sigma)$

- $X$ is a $d$-dimensional random vector, and
  
  $$X \sim N(\mu, \Sigma)$$

- If $Y = A^T X$, where $A$ is a $d \times k$ matrix, then

  $$Y \sim N(A^T \mu, A^T \Sigma A)$$
On Covariance Matrix

- As mentioned before, $\Sigma$ is *symmetric* and *positive semidefinite*.

\[
\Sigma = \Phi \Lambda \Phi^T = \Phi \Lambda^{1/2} \Lambda^{1/2} \Phi^T
\]

$\Phi$: *orthonormal* matrix, whose columns are *eigenvectors* of $\Sigma$.

$\Lambda$: *diagonal* matrix (eigenvalues).

- Thus,

\[
\Sigma = (\Phi \Lambda^{1/2})(\Phi \Lambda^{1/2})^T
\]
Mahalanobis Distance

- Mahalanobis distance

\[
r^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)
\]

\[X \sim N(\mu, \Sigma)\]

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

*depends on the value of \(r^2\)*

*constant*

*\(r^2\)
Discriminant Functions for Gaussian Density

- Minimum-error-rate classification

\[ g_i(x) = P(\omega_i \mid x) \quad (1 \leq i \leq c) \]

\[ g_i(x) = \ln P(\omega_i \mid x) \]

\[ p(x \mid \omega_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right] \]

\[ g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i) \]
Discriminant Functions for Gaussian Density

Three cases

- Case 1 \( \Sigma_i = \sigma^2 I \)
  - Classes are centered at different mean, and their feature components are pairwisely independent have the same variance.

- Case 2 \( \Sigma_i = \Sigma \)
  - Classes are centered at different mean, but have the same variation.

- Case 3 \( \Sigma_i \neq \Sigma_j \)
  - Arbitrary
Case 1: $\Sigma_i = \sigma^2 I$

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

$$g_i(x) = -\frac{1}{2\sigma^2} \| x - \mu_i \|^2 + \ln P(\omega_i)$$

$$\Sigma_i^{-1} = \frac{1}{\sigma^2} I$$

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

irrelevant

$$g_i(x) = \frac{1}{\sigma^2} \mu_i^T x + \left[ -\frac{1}{2\sigma^2} \mu_i^T \mu_i + \ln P(\omega_i) \right]$$
Case 1: $\sum_i = \sigma^2 I$

$$g_i(x) = \frac{1}{\sigma^2} \mu_i^T x + \left[ -\frac{1}{2\sigma^2} \mu_i^T \mu_i + \ln P(\omega_i) \right]$$

- It is a linear discriminant function

$$g_i(x) = w_i^T x + w_{i0}$$

- where
  - Weight vector
    $$w_i = \frac{1}{\sigma^2} \mu_i$$
  - Threshold/bias
    $$w_{i0} = -\frac{1}{2\sigma^2} \mu_i^T \mu_i + \ln P(\omega_i)$$
Case 1: $\Sigma_i = \sigma^2 I$

\[
g_i(x) = w_i^T x + w_{i0}
\]

\[
w_{i0} = -\frac{1}{2\sigma^2} \mu_i^T \mu_i + \ln P(\omega_i)
\]

\[
w_i = \frac{1}{\sigma^2} \mu_i
\]

\[
w_i^T x + w_{i0} = w_j^T x + w_{j0}
\]

\[(w_i^T - w_j^T)x = w_{j0} - w_{i0}\]

\[
\mu_i^T - \mu_j^T = \frac{1}{2} (\mu_i^T \mu_i - \mu_j^T \mu_j) - \sigma^2 \ln \frac{P(\omega_i)}{P(\omega_j)}
\]

\[
\mu_i^T - \mu_j^T = \frac{1}{2} (\mu_i^T - \mu_j^T)(\mu_i + \mu_j) - \sigma^2 \frac{(\mu_i^T - \mu_j^T)(\mu_i - \mu_j)}{\|\mu_i - \mu_j\|} \ln \frac{P(\omega_i)}{P(\omega_j)}
\]
Case 1: $\Sigma_i = \sigma^2 I$

- The decision boundary will be a **hyperplane** perpendicular to the line btw. the means at somewhere.

\[
w^T (x - x_0) = 0
\]

\[
w = \mu_i - \mu_j
\]

\[
x_0 = \frac{1}{2} (\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)
\]

**midpoint**

0 if $P(\omega_i) = P(\omega_j)$

\[
(w^T) x = \frac{1}{2} (\mu_i^T - \mu_j^T)(\mu_i + \mu_j) - \sigma^2 \frac{\|\mu_i - \mu_j\|^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)
\]

Boundary btw. $\omega_i$ and $\omega_j$

$g_i(x) = g_j(x)$
Case 1: $\Sigma_i = \sigma^2 I$

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

Minimum distance classifier (template matching)
Case 1: $\Sigma_i = \sigma^2 I$

$$P(\omega_1) > P(\omega_2)$$
Case 1: $\Sigma_i = \sigma^2 I$

$P(\omega_1) > P(\omega_2)$
Case 1: $\Sigma_i = \sigma^2 I$

$P(\omega_1) > P(\omega_2)$
Case 2: $\Sigma_i = \Sigma$

\[
g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)
\]

\[
g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma^{-1} (x - \mu_i) + \ln P(\omega_i)
\]

Irrelevant if $P(\omega_i) = P(\omega_j) \ \forall i, j$

Irrelevant

\[
g_i(x) = w_i^T x + w_{i0}
\]

\[
\begin{cases}
w_i = \Sigma^{-1} \mu_i \\
w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln P(\omega_i)
\end{cases}
\]
Case 2: \[ \Sigma_i = \Sigma \]

\[ g_i(x) = w_i^T x + w_{i0} \]

\[ w_i = \Sigma^{-1} \mu_i \]

\[ w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln P(\omega_i) \]

\[ w^T (x - x_0) = 0 \]

\[ x_0 = \frac{1}{2} (\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)} (\mu_i - \mu_j) \]
Case 2: $\sum_i = \sum$
Case 2: $\sum_i = \sum$
Case 3: $\Sigma_i \neq \Sigma_j$

\[
g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)
\]

\[
g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)
\]

\[
g_i(x) = x^T W_i x + w_i^T x + w_{i0}
\]

Without this term

In Case 1 and 2

\[
W_i = -\frac{1}{2} \Sigma_i^{-1}
\]

\[
w_i = \Sigma_i^{-1} \mu_i
\]

\[
w_{i0} = -\frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i - \frac{1}{2} \ln |\Sigma_i^{-1}| + \ln P(\omega_i)
\]

Decision surfaces are hyperquadrics, e.g.,
- Hyperplanes
- Hyperspheres
- Hyperellipsoids
- Hyperhyperboloids
Case 3: $\sum_i \neq \sum_j$

Non-simply connected decision regions can arise in one dimension for Gaussians having unequal variance.
Case 3: $\sum_i \neq \sum_j$
Case 3: $\Sigma_i \neq \Sigma_j$
Case 3: \( \Sigma_i \neq \Sigma_j \)
Case 3: $\Sigma_i \neq \Sigma_j$
Summary

- Bayesian Decision Theory
  - Basic concepts
  - Bayes theorem
  - Bayes decision rule
- Feasibility of Bayes Decision Rule
  - Prior probability + likelihood
  - Solution I: counting relative frequencies
  - Solution II: conduct density estimation
Summary

- Bayes decision rule: The general scenario
  - Allowing more than one feature
  - Allowing more than two states of nature
  - Allowing actions than merely deciding state of nature
  - Loss function
- Expected loss (conditional risk)
- General Bayes decision rule
- Minimum-error-rate classification
- Discriminant functions
- Gaussian density
- Discriminant functions for Gaussian pdf.
**k-means**

Start

Number of cluster K

Centroid

Distance objects to centroids

Grouping based on minimum distance

No object move group?

end

+ -
Thank You!

Any Question?