Multivariate Verfahren 2

discriminant analysis

Helmut Waldl

May 21st and 22nd 2012

discriminant analysis

preface

Problem: The population is composed of several subpopulations (groups), each object is member of **exactly one** of these groups.

Objective: Based on a vector of observed variates an object from the population with unknown class membership should be assigned to the correct class.

General decision theory ansatz

Given: $\Omega = \bigcup_{k=1}^{g} \Omega_k$ with pairwise disjoint subpopulations $\Omega_1, \ldots, \Omega_g;\ g \geq 2$

To each object ω there are assigned the values of $\mathbf{x} = (x_1, \ldots, x_p)$ and an index k of the subpopulation Ω_k where ω comes from.

discriminant analysis

preface

Problem: The population is composed of several subpopulations (groups), each object is member of **exactly one** of these groups.

Objective: Based on a vector of observed variates an object from the population with unknown class membership should be assigned to the correct class.

General decision theory ansatz

Given: $\Omega = \bigcup_{k=1}^{g} \Omega_k$ with pairwise disjoint subpopulations $\Omega_1, \ldots, \Omega_g;\ g \geq 2$

To each object ω there are assigned the values of $\mathbf{x} = (x_1, \ldots, x_p)$ and an index k of the subpopulation Ω_k where ω comes from.

discriminant analysis

preface

Problem: The population is composed of several subpopulations (groups), each object is member of **exactly one** of these groups.

Objective: Based on a vector of observed variates an object from the population with unknown class membership should be assigned to the correct class.

General decision theory ansatz

Given: $\Omega = \bigcup_{k=1}^{g} \Omega_k$ with pairwise disjoint subpopulations $\Omega_1, \ldots, \Omega_g;\ g \geq 2$

To each object ω there are assigned the values of $\mathbf{x} = (x_1, \ldots, x_p)$ and an index k of the subpopulation Ω_k where ω comes from.

discriminant analysis

preface

Problem: The population is composed of several subpopulations (groups), each object is member of **exactly one** of these groups.

Objective: Based on a vector of observed variates an object from the population with unknown class membership should be assigned to the correct class.

General decision theory ansatz
Given: $\Omega = \bigcup_{k=1}^{g} \Omega_k$ with pairwise disjoint subpopulations $\Omega_1, \ldots, \Omega_g;\ g \geq 2$

To each object ω there are assigned the values of $\mathbf{x} = (x_1, \ldots, x_p)$ and an index k of the subpopulation Ω_k where ω comes from.

discriminant analysis

decision theory ansatz

Problem:

The unknown index k should be uniquely determined because of $\mathbf{x}$:

We have to find a **decision function**

$$\begin{aligned} e: \Omega_x &\rightarrow \{1, \ldots, g\} \\ \mathbf{x} &\mapsto \hat{k} = e(\mathbf{x}) \end{aligned}$$

$\hat{k}$ is the estimate for the unknown class index k.

e should be such that there are as few wrong decisions as possible.

discriminant analysis

decision theory ansatz

Problem:

The unknown index k should be uniquely determined because of $\mathbf{x}$:

We have to find a **decision function**

$$\begin{aligned} e: \Omega_x &\rightarrow \{1, \ldots, g\} \\ \mathbf{x} &\mapsto \hat{k} = e(\mathbf{x}) \end{aligned}$$

$\hat{k}$ is the estimate for the unknown class index k.

e should be such that there are as few wrong decisions as possible.

discriminant analysis

decision theory ansatz

Problem:

The unknown index k should be uniquely determined because of $\mathbf{x}$:

We have to find a **decision function**

$$\begin{aligned} e: \Omega_x &\rightarrow \{1, \ldots, g\} \\ \mathbf{x} &\mapsto \hat{k} = e(\mathbf{x}) \end{aligned}$$

$\hat{k}$ is the estimate for the unknown class index k.

e should be such that there are as few wrong decisions as possible.

discriminant analysis

decision theory ansatz

Problem:

The unknown index k should be uniquely determined because of $\mathbf{x}$:

We have to find a **decision function**

$$\begin{aligned} e: \Omega_x &\rightarrow \{1, \ldots, g\} \\ \mathbf{x} &\mapsto \hat{k} = e(\mathbf{x}) \end{aligned}$$

$\hat{k}$ is the estimate for the unknown class index k.

e should be such that there are as few wrong decisions as possible.

discriminant analysis

decision theory ansatz

Assumption: $\mathbf{x}$ and k are random variates.

Let $p(k) = P(\omega \in \Omega_k)$ be the **a priori** probability for $\omega \in \Omega_k$

$f(\mathbf{x}|k)$ the pdf of the distribution of $\mathbf{x}$ in Ω_k

The pdf of the distribution of $\mathbf{x}$ in Ω then is a mixture of distributions:

$$f(\mathbf{x}) = \sum_{k=1}^{g} f(\mathbf{x}|k) \cdot p(k)$$

discriminant analysis

decision theory ansatz

Assumption: $\mathbf{x}$ and k are random variates.

Let $p(k) = P(\omega \in \Omega_k)$ be the **a priori** probability for $\omega \in \Omega_k$

$f(\mathbf{x}|k)$ the pdf of the distribution of $\mathbf{x}$ in Ω_k

The pdf of the distribution of $\mathbf{x}$ in Ω then is a mixture of distributions:

$$f(\mathbf{x}) = \sum_{k=1}^{g} f(\mathbf{x}|k) \cdot p(k)$$

discriminant analysis

decision theory ansatz

Assumption: $\mathbf{x}$ and k are random variates.

Let $p(k) = P(\omega \in \Omega_k)$ be the **a priori** probability for $\omega \in \Omega_k$

$f(\mathbf{x}|k)$ the pdf of the distribution of $\mathbf{x}$ in Ω_k

The pdf of the distribution of $\mathbf{x}$ in Ω then is a mixture of distributions:

$$f(\mathbf{x}) = \sum_{k=1}^{g} f(\mathbf{x}|k) \cdot p(k)$$

discriminant analysis

decision theory ansatz

Assumption: $\mathbf{x}$ and k are random variates.

Let $p(k) = P(\omega \in \Omega_k)$ be the **a priori** probability for $\omega \in \Omega_k$

$f(\mathbf{x}|k)$ the pdf of the distribution of $\mathbf{x}$ in Ω_k

The pdf of the distribution of $\mathbf{x}$ in Ω then is a mixture of distributions:

$$f(\mathbf{x}) = \sum_{k=1}^{g} f(\mathbf{x}|k) \cdot p(k)$$

discriminant analysis

decision theory ansatz

For the classification problem we are first of all interested in the **a posteriori** probability $f(k|\mathbf{x})$, the distribution of the class index when the data $\mathbf{x}$ are given.

According to the Bayes theorem we get:

$$\frac{f(k|\mathbf{x})}{p(k|\mathbf{x})} = \frac{f(\mathbf{x}|k) \cdot p(k)}{f(\mathbf{x})}$$

In practice we may assume that $p(k)$ and $f(\mathbf{x}|k)$ are unknown and first have to be estimated from a so called "learning sample".

In the learning sample we know the class index k as well as the data $\mathbf{x}$ of the objects.

We now assume that the distributions $p(k)$ and $f(\mathbf{x}|k)$ and their parameters are already estimated and thus $p(k)$ and $f(\mathbf{x}|k)$ are known.

discriminant analysis

decision theory ansatz

For the classification problem we are first of all interested in the **a posteriori** probability $f(k|\mathbf{x})$, the distribution of the class index when the data $\mathbf{x}$ are given.

According to the Bayes theorem we get:

$$\begin{matrix} f(k|\mathbf{x}) \\ p(k|\mathbf{x}) \end{matrix} = \frac{f(\mathbf{x}|k) \cdot p(k)}{f(\mathbf{x})}$$

In practice we may assume that $p(k)$ and $f(\mathbf{x}|k)$ are unknown and first have to be estimated from a so called "learning sample".

In the learning sample we know the class index k as well as the data $\mathbf{x}$ of the objects.

We now assume that the distributions $p(k)$ and $f(\mathbf{x}|k)$ and their parameters are already estimated and thus $p(k)$ and $f(\mathbf{x}|k)$ are known.

discriminant analysis

decision theory ansatz

For the classification problem we are first of all interested in the **a posteriori** probability $f(k|\mathbf{x})$, the distribution of the class index when the data $\mathbf{x}$ are given.

According to the Bayes theorem we get:

$$\frac{f(k|\mathbf{x})}{p(k|\mathbf{x})} = \frac{f(\mathbf{x}|k) \cdot p(k)}{f(\mathbf{x})}$$

In practice we may assume that $p(k)$ and $f(\mathbf{x}|k)$ are unknown and first have to be estimated from a so called "learning sample".

In the learning sample we know the class index k as well as the data $\mathbf{x}$ of the objects.

We now assume that the distributions $p(k)$ and $f(\mathbf{x}|k)$ and their parameters are already estimated and thus $p(k)$ and $f(\mathbf{x}|k)$ are known.

discriminant analysis

decision theory ansatz

For the classification problem we are first of all interested in the **a posteriori** probability $f(k|\mathbf{x})$, the distribution of the class index when the data $\mathbf{x}$ are given.

According to the Bayes theorem we get:

$$\begin{matrix} f(k|\mathbf{x}) \\ p(k|\mathbf{x}) \end{matrix} = \frac{f(\mathbf{x}|k) \cdot p(k)}{f(\mathbf{x})}$$

In practice we may assume that $p(k)$ and $f(\mathbf{x}|k)$ are unknown and first have to be estimated from a so called "learning sample".

In the learning sample we know the class index k as well as the data $\mathbf{x}$ of the objects.

We now assume that the distributions $p(k)$ and $f(\mathbf{x}|k)$ and their parameters are already estimated and thus $p(k)$ and $f(\mathbf{x}|k)$ are known.

discriminant analysis

decision theory ansatz

For the classification problem we are first of all interested in the **a posteriori** probability $f(k|\mathbf{x})$, the distribution of the class index when the data $\mathbf{x}$ are given.

According to the Bayes theorem we get:

$$\begin{matrix} f(k|\mathbf{x}) \\ p(k|\mathbf{x}) \end{matrix} = \frac{f(\mathbf{x}|k) \cdot p(k)}{f(\mathbf{x})}$$

In practice we may assume that $p(k)$ and $f(\mathbf{x}|k)$ are unknown and first have to be estimated from a so called "learning sample".

In the learning sample we know the class index k as well as the data $\mathbf{x}$ of the objects.

We now assume that the distributions $p(k)$ and $f(\mathbf{x}|k)$ and their parameters are already estimated and thus $\underline{p(k) \text{ and } f(\mathbf{x}|k) \text{ are known.}}$

discriminant analysis

denotation for misclassification probabilities

k . . .unknown group index

Mix-up probability - **individual error rate**:

$$\varepsilon_{k\hat{k}}(e) = P\left(e(\mathbf{x}) = \hat{k}|k\right),\ k \neq \hat{k}$$

Conditional error rate:

$$\varepsilon(e|\mathbf{x}) = P\left(k \neq e(\mathbf{x})|\mathbf{x}\right)$$

Total error rate:

$$\varepsilon(e) = P\left(e(\mathbf{x}) \neq k\right)$$

Success rate: $1 - \varepsilon(e)$. We get

$$\varepsilon(e) = \sum_{k=1}^{g} \sum_{\hat{k} \neq k=1}^{g} \varepsilon_{k\hat{k}}(e) \cdot p(k) = \int_{\Omega_x} \varepsilon(e|\mathbf{x}) \cdot f(\mathbf{x})\, d\mathbf{x}$$

discriminant analysis

denotation for misclassification probabilities

k . . .unknown group index

Mix-up probability - **individual error rate**:

$$\varepsilon_{k\hat{k}}(e) = P\left(e(\mathbf{x}) = \hat{k}|k\right),\ k \neq \hat{k}$$

Conditional error rate:

$$\varepsilon(e|\mathbf{x}) = P\left(k \neq e(\mathbf{x})|\mathbf{x}\right)$$

Total error rate:

$$\varepsilon(e) = P\left(e(\mathbf{x}) \neq k\right)$$

Success rate: $1 - \varepsilon(e)$. We get

$$\varepsilon(e) = \sum_{k=1}^{g} \sum_{\hat{k} \neq k=1}^{g} \varepsilon_{k\hat{k}}(e) \cdot p(k) = \int_{\Omega_x} \varepsilon(e|\mathbf{x}) \cdot f(\mathbf{x})\ d\mathbf{x}$$

discriminant analysis

denotation for misclassification probabilities

k . . .unknown group index

Mix-up probability - **individual error rate**:

$$\varepsilon_{k\hat{k}}(e) = P\left(e(\mathbf{x}) = \hat{k}|k\right),\ k \neq \hat{k}$$

Conditional error rate:

$$\varepsilon(e|\mathbf{x}) = P\left(k \neq e(\mathbf{x})|\mathbf{x}\right)$$

Total error rate:

$$\varepsilon(e) = P\left(e(\mathbf{x}) \neq k\right)$$

Success rate: $1 - \varepsilon(e)$. We get

$$\varepsilon(e) = \sum_{k=1}^{g} \sum_{\hat{k} \neq k=1}^{g} \varepsilon_{k\hat{k}}(e) \cdot p(k) = \int_{\Omega_x} \varepsilon(e|\mathbf{x}) \cdot f(\mathbf{x})\, d\mathbf{x}$$

discriminant analysis

denotation for misclassification probabilities

k . . .unknown group index

Mix-up probability - **individual error rate**:

$$\varepsilon_{k\hat{k}}(e) = P\left(e(\mathbf{x}) = \hat{k}|k\right),\ k \neq \hat{k}$$

Conditional error rate:

$$\varepsilon(e|\mathbf{x}) = P\left(k \neq e(\mathbf{x})|\mathbf{x}\right)$$

Total error rate:

$$\varepsilon(e) = P\left(e(\mathbf{x}) \neq k\right)$$

Success rate: $1 - \varepsilon(e)$. We get

$$\varepsilon(e) = \sum_{k=1}^{g} \sum_{\hat{k} \neq k=1}^{g} \varepsilon_{k\hat{k}}(e) \cdot p(k) = \int_{\Omega_x} \varepsilon(e|\mathbf{x}) \cdot f(\mathbf{x})\ d\mathbf{x}$$

discriminant analysis

denotation for misclassification probabilities

k . . .unknown group index

Mix-up probability - **individual error rate**:

$$\varepsilon_{k\hat{k}}(e) = P\left(e(\mathbf{x}) = \hat{k}|k\right),\ k \neq \hat{k}$$

Conditional error rate:

$$\varepsilon(e|\mathbf{x}) = P\left(k \neq e(\mathbf{x})|\mathbf{x}\right)$$

Total error rate:

$$\varepsilon(e) = P\left(e(\mathbf{x}) \neq k\right)$$

Success rate: $1 - \varepsilon(e)$. We get

$$\varepsilon(e) = \sum_{k=1}^{g} \sum_{\hat{k}\neq k=1}^{g} \varepsilon_{k\hat{k}}(e) \cdot p(k) = \int_{\Omega_x} \varepsilon(e|\mathbf{x}) \cdot f(\mathbf{x})\ d\mathbf{x}$$

discriminant analysis

Bayes- and ML decision rules

Definition: **Bayes decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $p(\hat{k}|\mathbf{x}) = \max_{k=1,\ldots,g} \{p(k|\mathbf{x})\}$ or rather

$p(\hat{k}) \cdot f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{p(k) \cdot f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class with the maximum a posteriori probability.

The maximum may be ambiguous but that does not affect the optimality property of this decision rule.

Special case: equal a priori probabilities $p(1) = p(2) = \cdots = p(g)$ $\longrightarrow$ **ML decision rule**

discriminant analysis

Bayes- and ML decision rules

Definition: **Bayes decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $p(\hat{k}|\mathbf{x}) = \max_{k=1,\ldots,g} \{p(k|\mathbf{x})\}$ or rather

$p(\hat{k}) \cdot f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{p(k) \cdot f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class with the maximum a posteriori probability.

The maximum may be ambiguous but that does not affect the optimality property of this decision rule.

Special case: equal a priori probabilities $p(1) = p(2) = \cdots = p(g)$ $\longrightarrow$ **ML decision rule**

discriminant analysis

Bayes- and ML decision rules

Definition: **Bayes decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $p(\hat{k}|\mathbf{x}) = \max_{k=1,\ldots,g} \{p(k|\mathbf{x})\}$ or rather

$p(\hat{k}) \cdot f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{p(k) \cdot f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class with the maximum a posteriori probability.

The maximum may be ambiguous but that does not affect the optimality property of this decision rule.

Special case: equal a priori probabilities $p(1) = p(2) = \cdots = p(g)$ $\longrightarrow$ **ML decision rule**

discriminant analysis

Bayes- and ML decision rules

Definition: **Bayes decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $p(\hat{k}|\mathbf{x}) = \max_{k=1,\ldots,g} \{p(k|\mathbf{x})\}$ or rather

$p(\hat{k}) \cdot f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{p(k) \cdot f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class with the maximum a posteriori probability.

The maximum may be ambiguous but that does not affect the optimality property of this decision rule.

Special case: equal a priori probabilities $p(1) = p(2) = \cdots = p(g)$ $\longrightarrow$ **ML decision rule**

discriminant analysis

Bayes- and ML decision rules

Definition: **Bayes decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $p(\hat{k}|\mathbf{x}) = \max_{k=1,\ldots,g} \{p(k|\mathbf{x})\}$ or rather

$p(\hat{k}) \cdot f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{p(k) \cdot f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class with the maximum a posteriori probability.

The maximum may be ambiguous but that does not affect the optimality property of this decision rule.

Special case: equal a priori probabilities $p(1) = p(2) = \cdots = p(g)$ $\longrightarrow$ **ML decision rule**

discriminant analysis

Bayes- and ML decision rules

Definition: **ML decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class index $\hat{k}$ such that the likelihood $L(\mathbf{x}|\hat{k}) = f(\mathbf{x}|\hat{k})$ is maximum.

The ML decision rule is particularly appropriate to classification if we regard the class membership not as variate but as unknown parameter (no a posteriori probabilities).

discriminant analysis

Bayes- and ML decision rules

Definition: **ML decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class index $\hat{k}$ such that the likelihood $L(\mathbf{x}|\hat{k}) = f(\mathbf{x}|\hat{k})$ is maximum.

The ML decision rule is particularly appropriate to classification if we regard the class membership not as variate but as unknown parameter (no a posteriori probabilities).

discriminant analysis

Bayes- and ML decision rules

Definition: **ML decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class index $\hat{k}$ such that the likelihood $L(\mathbf{x}|\hat{k}) = f(\mathbf{x}|\hat{k})$ is maximum.

The ML decision rule is particularly appropriate to classification if we regard the class membership not as variate but as unknown parameter (no a posteriori probabilities).

discriminant analysis

Bayes- and ML decision rules

Definition: **ML decision rule**

choose $\hat{k} = e(\mathbf{x})$ such that $f(\mathbf{x}|\hat{k}) = \max_{k=1,\ldots,g} \{f(\mathbf{x}|k)\}$

i.e. with given $\mathbf{x}$ we choose the class index $\hat{k}$ such that the likelihood $L(\mathbf{x}|\hat{k}) = f(\mathbf{x}|\hat{k})$ is maximum.

The ML decision rule is particularly appropriate to classification if we regard the class membership not as variate but as unknown parameter (no a posteriori probabilities).

Theorem: **Optimality of the Bayes decision rule**

Among all decision rules the Bayes decision rule has the smallest conditional error rate for all $\mathbf{x}$ and thus also the smallest total error rate.

Proof: the conditional error rate is

$$\varepsilon(e|\mathbf{x}) = P(k \neq e(\mathbf{x})|\mathbf{x}) = 1 - P(k = e(\mathbf{x})|\mathbf{x}) = 1 - \underbrace{p(k|\mathbf{x})}_{\substack{\text{a posteriori} \\ \text{probability}}}$$

With the Bayes decision rule we have $p(\hat{k}_B|\mathbf{x}) \geq p(k|\mathbf{x}) \quad \Longrightarrow$

$$\Longrightarrow \quad \varepsilon(e|\mathbf{x}) \geq 1 - p(\hat{k}_B|\mathbf{x}) = \varepsilon(e_B|\mathbf{x})$$

$\varepsilon(e_B|\mathbf{x})$ is the conditional error rate of the Bayes decision rule.

As we have $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, also the total error rate is minimum:

$$\varepsilon(e) = \int_{\Omega_x} \underbrace{\varepsilon(e|\mathbf{x})}_{\geq \varepsilon(e_B|\mathbf{x})} \cdot f(\mathbf{x})\, dx \;\geq\; \int_{\Omega_x} \varepsilon(e_B|\mathbf{x}) \cdot f(\mathbf{x})\, dx = \varepsilon(e_B)$$

Theorem: **Optimality of the Bayes decision rule**

Among all decision rules the Bayes decision rule has the smallest conditional error rate for all $\mathbf{x}$ and thus also the smallest total error rate.

Proof: the conditional error rate is

$$\varepsilon(e|\mathbf{x}) = P(k \neq e(\mathbf{x})|\mathbf{x}) = 1 - P(k = e(\mathbf{x})|\mathbf{x}) = 1 - \underbrace{p(k|\mathbf{x})}_{\substack{\text{a posteriori} \\ \text{probability}}}$$

With the Bayes decision rule we have $p(\hat{k}_B|\mathbf{x}) \geq p(k|\mathbf{x}) \quad \Longrightarrow$

$$\Longrightarrow \quad \varepsilon(e|\mathbf{x}) \geq 1 - p(\hat{k}_B|\mathbf{x}) = \varepsilon(e_B|\mathbf{x})$$

$\varepsilon(e_B|\mathbf{x})$ is the conditional error rate of the Bayes decision rule.

As we have $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, also the total error rate is minimum:

$$\varepsilon(e) = \int_{\Omega_x} \underbrace{\varepsilon(e|\mathbf{x})}_{\geq \varepsilon(e_B|\mathbf{x})} \cdot f(\mathbf{x})\, dx \;\geq\; \int_{\Omega_x} \varepsilon(e_B|\mathbf{x}) \cdot f(\mathbf{x})\, dx = \varepsilon(e_B)$$

Theorem: **Optimality of the Bayes decision rule**

Among all decision rules the Bayes decision rule has the smallest conditional error rate for all $\mathbf{x}$ and thus also the smallest total error rate.

Proof: the conditional error rate is

$$\varepsilon(e|\mathbf{x}) = P(k \neq e(\mathbf{x})|\mathbf{x}) = 1 - P(k = e(\mathbf{x})|\mathbf{x}) = 1 - \underbrace{p(k|\mathbf{x})}_{\substack{\text{a posteriori} \\ \text{probability}}}$$

With the Bayes decision rule we have $p(\hat{k}_B|\mathbf{x}) \geq p(k|\mathbf{x}) \implies$

$$\implies \varepsilon(e|\mathbf{x}) \geq 1 - p(\hat{k}_B|\mathbf{x}) = \varepsilon(e_B|\mathbf{x})$$

$\varepsilon(e_B|\mathbf{x})$ is the conditional error rate of the Bayes decision rule.

As we have $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, also the total error rate is minimum:

$$\varepsilon(e) = \int_{\Omega_x} \underbrace{\varepsilon(e|\mathbf{x})}_{\geq \varepsilon(e_B|\mathbf{x})} \cdot f(\mathbf{x})\, dx \;\geq\; \int_{\Omega_x} \varepsilon(e_B|\mathbf{x}) \cdot f(\mathbf{x})\, dx = \varepsilon(e_B)$$

Theorem: **Optimality of the Bayes decision rule**

Among all decision rules the Bayes decision rule has the smallest conditional error rate for all $\mathbf{x}$ and thus also the smallest total error rate.

Proof: the conditional error rate is

$$\varepsilon(e|\mathbf{x}) = P(k \neq e(\mathbf{x})|\mathbf{x}) = 1 - P(k = e(\mathbf{x})|\mathbf{x}) = 1 - \underbrace{p(k|\mathbf{x})}_{\substack{\text{a posteriori} \\ \text{probability}}}$$

With the Bayes decision rule we have $p(\hat{k}_B|\mathbf{x}) \geq p(k|\mathbf{x}) \quad \Longrightarrow$

$$\Longrightarrow \quad \varepsilon(e|\mathbf{x}) \geq 1 - p(\hat{k}_B|\mathbf{x}) = \varepsilon(e_B|\mathbf{x})$$

$\varepsilon(e_B|\mathbf{x})$ is the conditional error rate of the Bayes decision rule.

As we have $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, also the total error rate is minimum:

$$\varepsilon(e) = \int_{\Omega_x} \underbrace{\varepsilon(e|\mathbf{x})}_{\geq \varepsilon(e_B|\mathbf{x})} \cdot f(\mathbf{x})\, dx \;\geq\; \int_{\Omega_x} \varepsilon(e_B|\mathbf{x}) \cdot f(\mathbf{x})\, dx = \varepsilon(e_B)$$

Theorem: **Optimality of the Bayes decision rule**

Among all decision rules the Bayes decision rule has the smallest conditional error rate for all $\mathbf{x}$ and thus also the smallest total error rate.

Proof: the conditional error rate is

$$\varepsilon(e|\mathbf{x}) = P(k \neq e(\mathbf{x})|\mathbf{x}) = 1 - P(k = e(\mathbf{x})|\mathbf{x}) = 1 - \underbrace{p(k|\mathbf{x})}_{\substack{\text{a posteriori} \\ \text{probability}}}$$

With the Bayes decision rule we have $p(\hat{k}_B|\mathbf{x}) \geq p(k|\mathbf{x}) \quad \Longrightarrow$

$$\Longrightarrow \quad \varepsilon(e|\mathbf{x}) \geq 1 - p(\hat{k}_B|\mathbf{x}) = \varepsilon(e_B|\mathbf{x})$$

$\varepsilon(e_B|\mathbf{x})$ is the conditional error rate of the Bayes decision rule.

As we have $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, also the total error rate is minimum:

$$\varepsilon(e) = \int_{\Omega_x} \underbrace{\varepsilon(e|\mathbf{x})}_{\geq \varepsilon(e_B|\mathbf{x})} \cdot f(\mathbf{x})\, dx \;\geq\; \int_{\Omega_x} \varepsilon(e_B|\mathbf{x}) \cdot f(\mathbf{x})\, dx = \varepsilon(e_B)$$

Theorem: **Optimality of the Bayes decision rule**

Among all decision rules the Bayes decision rule has the smallest conditional error rate for all $\mathbf{x}$ and thus also the smallest total error rate.

Proof: the conditional error rate is

$$\varepsilon(e|\mathbf{x}) = P(k \neq e(\mathbf{x})|\mathbf{x}) = 1 - P(k = e(\mathbf{x})|\mathbf{x}) = 1 - \underbrace{p(k|\mathbf{x})}_{\substack{\text{a posteriori} \\ \text{probability}}}$$

With the Bayes decision rule we have $p(\hat{k}_B|\mathbf{x}) \geq p(k|\mathbf{x}) \quad \Longrightarrow$

$$\Longrightarrow \quad \varepsilon(e|\mathbf{x}) \geq 1 - p(\hat{k}_B|\mathbf{x}) = \varepsilon(e_B|\mathbf{x})$$

$\varepsilon(e_B|\mathbf{x})$ is the conditional error rate of the Bayes decision rule.

As we have $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, also the total error rate is minimum:

$$\varepsilon(e) = \int_{\Omega_x} \underbrace{\varepsilon(e|\mathbf{x})}_{\geq \varepsilon(e_B|\mathbf{x})} \cdot f(\mathbf{x})\, dx \;\geq\; \int_{\Omega_x} \varepsilon(e_B|\mathbf{x}) \cdot f(\mathbf{x})\, dx = \varepsilon(e_B)$$

discriminant analysis

Bayes- and ML decision rules

Corollary: **Optimality of the ML decision rule**

The ML decision rule is optimal if the a priori probabilities are equal, i.e. if no advance information is available.

Cost function

In Minimizing the error rate all misclassifications have been evaluated equal, no matter from which class the object came from and to which class it was assigned.

Different evaluations of misclassifications may be carried out with cost functions.

This is important in practice, e.g. the classification of a sick person as healthy is more fatal than the classification of a healthy person as sick.

discriminant analysis

Bayes- and ML decision rules

Corollary: **Optimality of the ML decision rule**

The ML decision rule is optimal if the a priori probabilities are equal, i.e. if no advance information is available.

Cost function

In Minimizing the error rate all misclassifications have been evaluated equal, no matter from which class the object came from and to which class it was assigned.

Different evaluations of misclassifications may be carried out with cost functions.

This is important in practice, e.g. the classification of a sick person as healthy is more fatal than the classification of a healthy person as sick.

discriminant analysis

Bayes- and ML decision rules

Corollary: **Optimality of the ML decision rule**

The ML decision rule is optimal if the a priori probabilities are equal, i.e. if no advance information is available.

Cost function

In Minimizing the error rate all misclassifications have been evaluated equal, no matter from which class the object came from and to which class it was assigned.

Different evaluations of misclassifications may be carried out with cost functions.

This is important in practice, e.g. the classification of a sick person as healthy is more fatal than the classification of a healthy person as sick.

discriminant analysis

Bayes- and ML decision rules

Corollary: **Optimality of the ML decision rule**

The ML decision rule is optimal if the a priori probabilities are equal, i.e. if no advance information is available.

Cost function

In Minimizing the error rate all misclassifications have been evaluated equal, no matter from which class the object came from and to which class it was assigned.

Different evaluations of misclassifications may be carried out with cost functions.

This is important in practice, e.g. the classification of a sick person as healthy is more fatal than the classification of a healthy person as sick.

Corollary: **Optimality of the ML decision rule**

The ML decision rule is optimal if the a priori probabilities are equal, i.e. if no advance information is available.

Cost function

In Minimizing the error rate all misclassifications have been evaluated equal, no matter from which class the object came from and to which class it was assigned.

Different evaluations of misclassifications may be carried out with cost functions.

This is important in practice, e.g. the classification of a sick person as healthy is more fatal than the classification of a healthy person as sick.

Corollary: **Optimality of the ML decision rule**

The ML decision rule is optimal if the a priori probabilities are equal, i.e. if no advance information is available.

Cost function

In Minimizing the error rate all misclassifications have been evaluated equal, no matter from which class the object came from and to which class it was assigned.

Different evaluations of misclassifications may be carried out with cost functions.

This is important in practice, e.g. the classification of a sick person as healthy is more fatal than the classification of a healthy person as sick.

discriminant analysis

cost functions

Definition:

$C(k, \hat{k})$ are the costs that arise if the rule decides on $\hat{k}$ but the true class index is k. We have $C(k, k) = 0$

Let $e(\mathbf{x}) = \hat{k}$, out of it the conditional costs arise:

$$C(\hat{k}|\mathbf{x}) = \sum_{k=1}^{g} C(k, \hat{k}) \cdot p(k|\mathbf{x}) \qquad \text{or rather equivalent}$$

$$\bar{C}(\hat{k}|\mathbf{x}) = \sum_{k=1}^{g} C(k, \hat{k}) \cdot p(k) \cdot f(\mathbf{x}|k)$$

i.e. the costs weighted with the a posteriori probabilities.

discriminant analysis

cost functions

Definition:

$C(k, \hat{k})$ are the costs that arise if the rule decides on $\hat{k}$ but the true class index is k. We have $C(k, k) = 0$

Let $e(\mathbf{x}) = \hat{k}$, out of it the conditional costs arise:

$$C(\hat{k}|\mathbf{x}) = \sum_{k=1}^{g} C(k, \hat{k}) \cdot p(k|\mathbf{x}) \qquad \text{or rather equivalent}$$

$$\bar{C}(\hat{k}|\mathbf{x}) = \sum_{k=1}^{g} C(k, \hat{k}) \cdot p(k) \cdot f(\mathbf{x}|k)$$

i.e. the costs weighted with the a posteriori probabilities.

discriminant analysis

cost functions

optimal cost decision rule (analog with Bayes decision rule)
Choose $\hat{k} = e(\mathbf{x})$ such that the conditional costs C and $\bar{C}$ are minimized respectively.

Conclusion: If the conditional costs are minimal for all $\mathbf{x}$ then also the total costs are minimal.

We are using the same argumentation as in the proof of the optimality of the Bayes decision rule: We have the total costs

$$C(\hat{k}) = \int_{\Omega_x} C(\hat{k}|\mathbf{x}) \cdot f(\mathbf{x})\, dx \qquad \bar{C}(\hat{k}) = \int_{\Omega_x} \bar{C}(\hat{k}|\mathbf{x}) \cdot f(\mathbf{x})\, dx$$

discriminant analysis

cost functions

optimal cost decision rule (analog with Bayes decision rule)
Choose $\hat{k} = e(\mathbf{x})$ such that the conditional costs C and $\bar{C}$ are minimized respectively.

Conclusion: If the conditional costs are minimal for all $\mathbf{x}$ then also the total costs are minimal.

We are using the same argumentation as in the proof of the optimality of the Bayes decision rule: We have the total costs

$$C(\hat{k}) = \int_{\Omega_x} C(\hat{k}|\mathbf{x}) \cdot f(\mathbf{x})\, dx \qquad \bar{C}(\hat{k}) = \int_{\Omega_x} \bar{C}(\hat{k}|\mathbf{x}) \cdot f(\mathbf{x})\, dx$$

discriminant analysis

cost functions

optimal cost decision rule (analog with Bayes decision rule)
Choose $\hat{k} = e(\mathbf{x})$ such that the conditional costs C and $\bar{C}$ are minimized respectively.

Conclusion: If the conditional costs are minimal for all $\mathbf{x}$ then also the total costs are minimal.

We are using the same argumentation as in the proof of the optimality of the Bayes decision rule: We have the total costs

$$C(\hat{k}) = \int_{\Omega_x} C(\hat{k}|\mathbf{x}) \cdot f(\mathbf{x})\ dx \qquad \bar{C}(\hat{k}) = \int_{\Omega_x} \bar{C}(\hat{k}|\mathbf{x}) \cdot f(\mathbf{x})\ dx$$

discriminant analysis

cost functions

Important **special cost functions**:

$$C_e(k, \hat{k}) = \begin{cases} 0 & \text{for } k = \hat{k} \\ c > 0 & \text{for } k \neq \hat{k} \end{cases}$$

ordinary symmetric cost function, all misclassifications are evaluated with equal costs. The cost optimal decision rule with this cost function is the Bayes decision rule.

$$C_p(k, \hat{k}) = \begin{cases} 0 & \text{for } k = \hat{k} \\ \frac{c}{p(k)} & \text{for } k \neq \hat{k} \end{cases}$$

inversely proportional cost function. Here the costs for misclassifications of objects from classes with low a priori probabilities are strongly increased (e.g. rarely occurring severe diseases). The cost optimal decision rule with this cost function is the ML decision rule.

discriminant analysis

cost functions

Important **special cost functions**:

$$C_e(k, \hat{k}) = \begin{cases} 0 & \text{for } k = \hat{k} \\ c > 0 & \text{for } k \neq \hat{k} \end{cases}$$

ordinary symmetric cost function, all misclassifications are evaluated with equal costs. The cost optimal decision rule with this cost function is the Bayes decision rule.

$$C_p(k, \hat{k}) = \begin{cases} 0 & \text{for } k = \hat{k} \\ \frac{c}{p(k)} & \text{for } k \neq \hat{k} \end{cases}$$

inversely proportional cost function. Here the costs for misclassifications of objects from classes with low a priori probabilities are strongly increased (e.g. rarely occurring severe diseases). The cost optimal decision rule with this cost function is the ML decision rule.

discriminant analysis

discriminant function

The above decision rules are based on the following principle:
For given $\mathbf{x}$ compute the g discriminant functions
$d_1(\mathbf{x}), \ldots, d_g(\mathbf{x}) : \Omega_x \to \mathbb{R}$. ω is then assigned to $\Omega_{\hat{k}}$ with

$$d_{\hat{k}(\mathbf{x})} = \max_{k=1,\ldots,g} \{d_k(\mathbf{x})\} \qquad (e(\mathbf{x}) = \hat{k})$$

Each decision rule has its own discriminant function:

- Bayes decision rule: $d_k(\mathbf{x}) = p(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k))$
- ML decision rule: $d_k(\mathbf{x}) = f(\mathbf{x}|k)$
- cost optimal decision rule: $d_k(\mathbf{x}) = -C(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = -\bar{C}(k|\mathbf{x}))$

discriminant analysis

discriminant function

The above decision rules are based on the following principle:
For given $\mathbf{x}$ compute the g discriminant functions
$d_1(\mathbf{x}), \ldots, d_g(\mathbf{x}) : \Omega_x \to \mathbb{R}$. ω is then assigned to $\Omega_{\hat{k}}$ with

$$d_{\hat{k}(\mathbf{x})} = \max_{k=1,\ldots,g} \{d_k(\mathbf{x})\} \qquad (e(\mathbf{x}) = \hat{k})$$

Each decision rule has its own discriminant function:

- Bayes decision rule: $d_k(\mathbf{x}) = p(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k))$
- ML decision rule: $d_k(\mathbf{x}) = f(\mathbf{x}|k)$
- cost optimal decision rule: $d_k(\mathbf{x}) = -C(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = -\bar{C}(k|\mathbf{x}))$

discriminant analysis

discriminant function

The above decision rules are based on the following principle:
For given $\mathbf{x}$ compute the g discriminant functions
$d_1(\mathbf{x}), \ldots, d_g(\mathbf{x}) : \Omega_x \to \mathbb{R}$. ω is then assigned to $\Omega_{\hat{k}}$ with

$$d_{\hat{k}(\mathbf{x})} = \max_{k=1,\ldots,g} \{d_k(\mathbf{x})\} \qquad (e(\mathbf{x}) = \hat{k})$$

Each decision rule has its own discriminant function:

- Bayes decision rule: $d_k(\mathbf{x}) = p(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k))$
- ML decision rule: $d_k(\mathbf{x}) = f(\mathbf{x}|k)$
- cost optimal decision rule: $d_k(\mathbf{x}) = -C(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = -\bar{C}(k|\mathbf{x}))$

discriminant analysis

discriminant function

The above decision rules are based on the following principle:
For given $\mathbf{x}$ compute the g discriminant functions
$d_1(\mathbf{x}), \ldots, d_g(\mathbf{x}) : \Omega_x \to \mathbb{R}$. ω is then assigned to $\Omega_{\hat{k}}$ with

$$d_{\hat{k}(\mathbf{x})} = \max_{k=1,\ldots,g} \{d_k(\mathbf{x})\} \qquad (e(\mathbf{x}) = \hat{k})$$

Each decision rule has its own discriminant function:

- Bayes decision rule: $d_k(\mathbf{x}) = p(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k))$
- ML decision rule: $d_k(\mathbf{x}) = f(\mathbf{x}|k)$
- cost optimal decision rule: $d_k(\mathbf{x}) = -C(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = -\bar{C}(k|\mathbf{x}))$

discriminant analysis

discriminant function

The above decision rules are based on the following principle:
For given $\mathbf{x}$ compute the g discriminant functions
$d_1(\mathbf{x}), \ldots, d_g(\mathbf{x}) : \Omega_x \to \mathbb{R}$. ω is then assigned to $\Omega_{\hat{k}}$ with

$$d_{\hat{k}(\mathbf{x})} = \max_{k=1,\ldots,g} \{d_k(\mathbf{x})\} \qquad (e(\mathbf{x}) = \hat{k})$$

Each decision rule has its own discriminant function:

- Bayes decision rule: $d_k(\mathbf{x}) = p(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k))$
- ML decision rule: $d_k(\mathbf{x}) = f(\mathbf{x}|k)$
- cost optimal decision rule: $d_k(\mathbf{x}) = -C(k|\mathbf{x}) \qquad (d_k(\mathbf{x}) = -\bar{C}(k|\mathbf{x}))$

discriminant analysis

discriminant function

For some applications it is more favorable to work with functions $f(d_1(\mathbf{x})), \ldots, f(d_g(\mathbf{x}))$ instead of the discriminant functions $d_1(\mathbf{x}), \ldots, d_g(\mathbf{x})$ themselves. Here f has to be strictly monotonic (increasing)

Example: $(\mathbf{x}|k) \sim \mathbf{N}(\mu, \sigma^2)$
the Bayes discriminant function is $d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k)$
equivalent is $d'_k(\mathbf{x}) = \ln(p(k)) + \underbrace{\ln(f(\mathbf{x}|k))}_{\substack{\text{advantageous} \\ \text{for computation}}}$

discriminant analysis

discriminant function

For some applications it is more favorable to work with functions $f(d_1(\mathbf{x})), \ldots, f(d_g(\mathbf{x}))$ instead of the discriminant functions $d_1(\mathbf{x}), \ldots, d_g(\mathbf{x})$ themselves. Here f has to be strictly monotonic (increasing)

Example: $(\mathbf{x}|k) \sim \mathbf{N}(\mu, \sigma^2)$
the Bayes discriminant function is $d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k)$
equivalent is $d'_k(\mathbf{x}) = \ln(p(k)) + \underbrace{\ln(f(\mathbf{x}|k))}_{\substack{\text{advantageous} \\ \text{for computation}}}$