Risk

Risk

Conditional risk

Expected loss associated to each action $\alpha_i \in \mathcal{A}$ given $\mathbf{x}$:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{|\mathcal{W}|} \lambda(\alpha_i|w_j) \cdot P(w_j|\mathbf{x})$$

If I see feature vector $\mathbf{x}$*, what is my expected loss if I decide* α_i*?*
How risky is this α_i *at this particular* $\mathbf{x}$

Risk

Conditional risk

Expected loss associated to each action $\alpha_i \in \mathcal{A}$ given $\mathbf{x}$:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{|\mathcal{W}|} \lambda(\alpha_i|w_j) \cdot P(w_j|\mathbf{x})$$

If I see feature vector $\mathbf{x}$*, what is my expected loss if I decide* α_i*?*
How risky is this α_i *at this particular* $\mathbf{x}$

Risk (or loss)

Average over all possible observations of conditional risk

$$R = \int_{\mathbb{R}^d} R\left(\gamma(\mathbf{x})|\mathbf{x}\right) \cdot p(\mathbf{x}) d\mathbf{x}$$

where decision rule $\gamma : \mathbb{R}^d \rightarrow \mathcal{A}$ selects the appropriate action to take

Minimum-risk Bayes rule

$$R = \int R\left(\gamma(\mathbf{x})|\mathbf{x}\right) \cdot p(\mathbf{x}) d\mathbf{x}$$

Minimum-risk Bayes rule

$$R = \int R\left(\gamma(\mathbf{x})|\mathbf{x}\right) \cdot p(\mathbf{x})d\mathbf{x}$$

- If $\gamma(\mathbf{x})$ is chosen so that $R(\gamma(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$

Minimum-risk Bayes rule

$$R = \int R\left(\gamma(\mathbf{x})|\mathbf{x}\right) \cdot p(\mathbf{x}) d\mathbf{x}$$

- If $\gamma(\mathbf{x})$ is chosen so that $R(\gamma(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$
  - $\rightarrow$ Overall risk will be minimized

Minimum-risk Bayes rule

$$R = \int R\left(\gamma(\mathbf{x})|\mathbf{x}\right) \cdot p(\mathbf{x})d\mathbf{x}$$

- If $\gamma(\mathbf{x})$ is chosen so that $R(\gamma(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$
  $\rightarrow$ Overall risk will be minimized
- Bayes decision rule: Select the action $\alpha_i \in \mathcal{A}$ that minimizes $R(\alpha_i|\mathbf{x})$

Minimum-risk Bayes rule

$$R = \int R\left(\gamma(\mathbf{x})|\mathbf{x}\right) \cdot p(\mathbf{x})d\mathbf{x}$$

- If $\gamma(\mathbf{x})$ is chosen so that $R(\gamma(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$
  $\rightarrow$ Overall risk will be minimized
- Bayes decision rule: Select the action $\alpha_i \in \mathcal{A}$ that minimizes $R(\alpha_i|\mathbf{x})$

Minimum-risk Bayes rule

$$\gamma(\mathbf{x}) = \arg\min_{\alpha_i \in \mathcal{A}} R(\alpha_i|\mathbf{x})$$

Minimum-risk Bayes rule

$$R = \int R\left(\gamma(\mathbf{x})|\mathbf{x}\right) \cdot p(\mathbf{x}) d\mathbf{x}$$

- If $\gamma(\mathbf{x})$ is chosen so that $R(\gamma(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$
  $\rightarrow$ Overall risk will be minimized
- Bayes decision rule: Select the action $\alpha_i \in \mathcal{A}$ that minimizes $R(\alpha_i|\mathbf{x})$

Minimum-risk Bayes rule

$$\gamma(\mathbf{x}) = \arg\min_{\alpha_i \in \mathcal{A}} R(\alpha_i|\mathbf{x})$$

- Bayes risk: Best performance that can be achieved
  $\rightarrow$ Usually denoted as R^*

Classification: the two-category case

- $\mathcal{A} = \{\alpha_1, \alpha_2\}$
- $\mathcal{W} = \{\omega_1, \omega_2\}$
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ with $i, j \in \{1, 2\}$

Classification: the two-category case

- $\mathcal{A} = \{\alpha_1, \alpha_2\}$
- $\mathcal{W} = \{\omega_1, \omega_2\}$
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ with $i, j \in \{1, 2\}$

$$\omega_1 \Rightarrow R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$$

Classification: the two-category case

- $\mathcal{A} = \{\alpha_1, \alpha_2\}$
- $\mathcal{W} = \{\omega_1, \omega_2\}$
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ with $i, j \in \{1, 2\}$

$$\omega_1 \Rightarrow \overbrace{\lambda_{11} \cdot P(\omega_1|\mathbf{x}) + \lambda_{12} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_1|\mathbf{x})} < \overbrace{\lambda_{21} \cdot P(\omega_1|\mathbf{x}) + \lambda_{22} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_2|\mathbf{x})}$$

Classification: the two-category case

- $\mathcal{A} = \{\alpha_1, \alpha_2\}$
- $\mathcal{W} = \{\omega_1, \omega_2\}$
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ with $i, j \in \{1, 2\}$

$$\omega_1 \Rightarrow \overbrace{\lambda_{11} \cdot P(\omega_1|\mathbf{x}) + \lambda_{12} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_1|\mathbf{x})} < \overbrace{\lambda_{21} \cdot P(\omega_1|\mathbf{x}) + \lambda_{22} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_2|\mathbf{x})}$$

$$\Downarrow \; P(\omega|\mathbf{x}) \propto p(\mathbf{x}|\omega) \cdot P(\omega)$$

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

Classification: the two-category case

- $\mathcal{A} = \{\alpha_1, \alpha_2\}$
- $\mathcal{W} = \{\omega_1, \omega_2\}$
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ with $i, j \in \{1, 2\}$

$$\omega_1 \Rightarrow \overbrace{\lambda_{11} \cdot P(\omega_1|\mathbf{x}) + \lambda_{12} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_1|\mathbf{x})} < \overbrace{\lambda_{21} \cdot P(\omega_1|\mathbf{x}) + \lambda_{22} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_2|\mathbf{x})}$$

$$\Downarrow \; P(\omega|\mathbf{x}) \propto p(\mathbf{x}|\omega) \cdot P(\omega)$$

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

$$\Downarrow$$

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Classification: the two-category case

- $\mathcal{A} = \{\alpha_1, \alpha_2\}$
- $\mathcal{W} = \{\omega_1, \omega_2\}$
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ with $i, j \in \{1, 2\}$

$$\omega_1 \Rightarrow \overbrace{\lambda_{11} \cdot P(\omega_1|\mathbf{x}) + \lambda_{12} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_1|\mathbf{x})} < \overbrace{\lambda_{21} \cdot P(\omega_1|\mathbf{x}) + \lambda_{22} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_2|\mathbf{x})}$$

$$\Downarrow \; P(\omega|\mathbf{x}) \propto p(\mathbf{x}|\omega) \cdot P(\omega)$$

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

$$\Downarrow$$

$$\text{Likelihood ratio} \Rightarrow \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

Classification: the two-category case

- $\mathcal{A} = \{\alpha_1, \alpha_2\}$
- $\mathcal{W} = \{\omega_1, \omega_2\}$
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ with $i, j \in \{1, 2\}$

$$\omega_1 \Rightarrow \overbrace{\lambda_{11} \cdot P(\omega_1|\mathbf{x}) + \lambda_{12} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_1|\mathbf{x})} < \overbrace{\lambda_{21} \cdot P(\omega_1|\mathbf{x}) + \lambda_{22} \cdot P(\omega_2|\mathbf{x})}^{R(\alpha_2|\mathbf{x})}$$

$$\Downarrow \; P(\omega|\mathbf{x}) \propto p(\mathbf{x}|\omega) \cdot P(\omega)$$

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

$$\Downarrow$$

$$\text{Likelihood ratio} \Rightarrow \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \underbrace{\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}}_{\text{Independent from } \mathbf{x}} = \theta$$

Classification: likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Classification: likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Consider a medical scenario where $\omega_1 \Rightarrow$healthy and $\omega_2 \Rightarrow$sick

Classification: likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Consider a medical scenario where ω_1 ⇒healthy and ω_2 ⇒sick

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases}$$

Classification: likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Consider a medical scenario where ω_1 ⇒healthy and ω_2 ⇒sick

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases}$$

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i = 2 \wedge j = 1 \\ 2 & \text{if} \quad i = 1 \wedge j = 2 \end{cases}$$

Classification: likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Consider a medical scenario where ω_1 ⇒healthy and ω_2 ⇒sick

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases}$$

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i = 2 \wedge j = 1 \\ 2 & \text{if} \quad i = 1 \wedge j = 2 \end{cases}$$

$$\mathrm{LR}_A = \frac{1-0}{1-0} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

Classification: likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Consider a medical scenario where $\omega_1 \Rightarrow$healthy and $\omega_2 \Rightarrow$sick

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases}$$

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i = 2 \wedge j = 1 \\ 2 & \text{if} \quad i = 1 \wedge j = 2 \end{cases}$$

$$\text{LR}_A = \frac{1-0}{1-0} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

$$\text{LR}_B = \frac{2-0}{1-0} \cdot \frac{P(\omega_2)}{P(\omega_1)} = 2\theta$$

Classification: likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

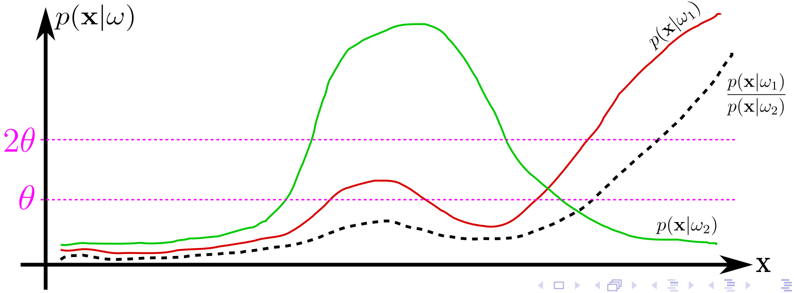
Consider a medical scenario where ω_1 ⇒healthy and ω_2 ⇒sick

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases}$$

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i = 2 \wedge j = 1 \\ 2 & \text{if} \quad i = 1 \wedge j = 2 \end{cases}$$

$$\text{LR}_A = \frac{1-0}{1-0} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

$$\text{LR}_B = \frac{2-0}{1-0} \cdot \frac{P(\omega_2)}{P(\omega_1)} = 2\theta$$



Classification: Bayes error

General classification case:

- $\mathcal{A} = \{\alpha_1, \ldots, \alpha_c\}$
- $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$

Classification: Bayes error

General classification case:

- $\mathcal{A} = \{\alpha_1, \ldots, \alpha_c\}$
- $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$

Symmetrical/zero-one loss

- No loss to correct decision
- Unit loss to any error

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Classification: Bayes error

General classification case:

- $\mathcal{A} = \{\alpha_1, \dots, \alpha_c\}$
- $\mathcal{W} = \{\omega_1, \dots, \omega_c\}$

Symmetrical/zero-one loss

- No loss to correct decision
- Unit loss to any error

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases} \qquad i, j = 1, \dots, c$$

Conditional risk:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\mathbf{x}) = \sum_{j \neq i}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

Classification: Bayes error

General classification case:

- $\mathcal{A} = \{\alpha_1, \ldots, \alpha_c\}$
- $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$

Symmetrical/zero-one loss

- No loss to correct decision
- Unit loss to any error

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Conditional risk:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\mathbf{x}) = \sum_{j \neq i}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

$\rightarrow$ Bayes rule $\Rightarrow$ Minimize conditional risk $\Rightarrow$ Maximize posterior probability

Classification: Bayes error

General classification case:

- $\mathcal{A} = \{\alpha_1, \ldots, \alpha_c\}$
- $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$

Symmetrical/zero-one loss

- No loss to correct decision
- Unit loss to any error

$$\lambda_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Conditional risk:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\mathbf{x}) = \sum_{j \neq i}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

$\rightarrow$ Bayes rule $\Rightarrow$ Minimize conditional risk $\Rightarrow$ Maximize posterior probability

$\rightarrow$ Bayes error:

$\Rightarrow$ Bayes risk for classification problems with zero-one loss

$\Rightarrow$ Irreducible misclassification probability

Summary

- **Loss** (λ): cost of one action (α_i) VS a state of nature (ω_j)
- **Conditional risk** ($R(\alpha_i|\mathbf{x})$): expected loss for observation $\mathbf{x}$ and action (α_i)
- **Risk** (R): also *expected loss*; overall expected loss for a decision rule (α_i) across the whole distribution of data
- **Bayes risk**: the minimum possible risk (the theoretical lower bound)
- **Bayes rule**: Action selection function (γ) that selects the optimal action that minimizes the risk (Bayes risk)
- **Bayes error**: Bayes risk for classification problems with zero-one loss

Discriminant functions

- General manner for representing classifiers

Discriminant functions

- General manner for representing classifiers
- Notation $\Rightarrow$ $g : \mathbb{R}^d \rightarrow \mathcal{R}$, where $\mathcal{R}$ denotes the *acceptance region*

Discriminant functions

- General manner for representing classifiers
- Notation $\Rightarrow$ $g : \mathbb{R}^d \rightarrow \mathcal{R}$, where $\mathcal{R}$ denotes the *acceptance region*
- For $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$ states of nature:

Discriminant functions

- General manner for representing classifiers
- Notation $\Rightarrow$ $g : \mathbb{R}^d \rightarrow \mathcal{R}$, where $\mathcal{R}$ denotes the *acceptance region*
- For $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$ states of nature:
  - $\rightarrow$ $g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})$
  - $\rightarrow$ $\mathcal{R}_1, \ldots, \mathcal{R}_c$

Discriminant functions

- General manner for representing classifiers
- Notation $\Rightarrow$ $g : \mathbb{R}^d \rightarrow \mathcal{R}$, where $\mathcal{R}$ denotes the *acceptance region*
- For $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$ states of nature:
  - $\rightarrow$ $g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})$
  - $\rightarrow$ $\mathcal{R}_1, \ldots, \mathcal{R}_c$
- The classifier is described as:

$$\alpha_a(\mathbf{x}) : a = \underset{i=1,\ldots,c}{\arg\max}\; g_i(\mathbf{x})$$

Discriminant functions

- General manner for representing classifiers
- Notation $\Rightarrow$ $g : \mathbb{R}^d \rightarrow \mathcal{R}$, where $\mathcal{R}$ denotes the *acceptance region*
- For $\mathcal{W} = \{\omega_1, \ldots, \omega_c\}$ states of nature:
  - $\rightarrow$ $g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})$
  - $\rightarrow$ $\mathcal{R}_1, \ldots, \mathcal{R}_c$
- The classifier is described as:

$$\alpha_a(\mathbf{x}) : a = \underset{i=1,\ldots,c}{\arg\max}\; g_i(\mathbf{x})$$

$\rightarrow$ For the Bayes classifier: $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x}) = P(\omega_i|\mathbf{x})$

Discriminant functions

- Acceptance region:

$$\mathcal{R}_i = \{\mathbf{x} : g_i(\mathbf{x}) > g_j(\mathbf{x}) \; \forall i, j \in \{1, \ldots, c\}, j \neq i\}$$

Discriminant functions

- Acceptance region:

$$\mathcal{R}_i = \{\mathbf{x} : g_i(\mathbf{x}) > g_j(\mathbf{x}) \; \forall i, j \in \{1, \ldots, c\}, j \neq i\}$$

- Decision boundaries:

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

Discriminant functions

- Acceptance region:

$$\mathcal{R}_i = \{\mathbf{x} : g_i(\mathbf{x}) > g_j(\mathbf{x})\ \forall i, j \in \{1, \ldots, c\}, j \neq i\}$$

- Decision boundaries:

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

- When $c = 2$ (binary classification), only one function is considered:

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

Outline

Introduction

- Optimal classifier: requires $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$

Introduction

- Optimal classifier: requires $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$
  - $\rightarrow$ Unfortunately, this complete knowledge is rarely available

Introduction

- Optimal classifier: requires $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$
  - $\rightarrow$ Unfortunately, this complete knowledge is rarely available
- Most typical scenario:

Introduction

- Optimal classifier: requires $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$
  - $\rightarrow$ Unfortunately, this complete knowledge is rarely available
- Most typical scenario:
  - $\rightarrow$ Some vague and general knowledge about the situation
  - $\rightarrow$ A number of *design samples* or *training data*

Introduction

- Optimal classifier: requires $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$
  - $\rightarrow$ Unfortunately, this complete knowledge is rarely available
- Most typical scenario:
  - $\rightarrow$ Some vague and general knowledge about the situation
  - $\rightarrow$ A number of *design samples* or *training data*

$\Rightarrow$ **Task**: Estimate distributions using this piece of information

Introduction

- Optimal classifier: requires $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$
  - $\rightarrow$ Unfortunately, this complete knowledge is rarely available
- Most typical scenario:
  - $\rightarrow$ Some vague and general knowledge about the situation
  - $\rightarrow$ A number of *design samples* or *training data*

$\Rightarrow$ **Task**: Estimate distributions using this piece of information

- $\rightarrow$ **Parametric Pattern Recognition**: Particular case in which $p(\mathbf{x}|\omega_i)$ can be characterized by few parameters

Introduction

- Optimal classifier: requires $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$
  - → Unfortunately, this complete knowledge is rarely available
- Most typical scenario:
  - → Some vague and general knowledge about the situation
  - → A number of *design samples* or *training data*

⇒ **Task**: Estimate distributions using this piece of information

  - → **Parametric Pattern Recognition**: Particular case in which $p(\mathbf{x}|\omega_i)$ can be characterized by few parameters
  - → Simplifies the problem of *estimating an unknown function* $p(\mathbf{x}|\omega_i)$ to one of estimating the parameters of a distribution

Parameter estimation

- Suppose the existence of a dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_c\}$

Parameter estimation

- Suppose the existence of a dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_c\}$
  - $\rightarrow$ $\mathcal{D}_i \sim p(\mathbf{x}|\omega_i)$ with $i = 1, \ldots, c$ (i.i.d.)
  - $\rightarrow$ $\mathcal{D} = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{D}|}\right\}$

Parameter estimation

- Suppose the existence of a dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_c\}$
  - $\rightarrow$ $\mathcal{D}_i \sim p(\mathbf{x}|\omega_i)$ with $i = 1, \ldots, c$ (i.i.d.)
  - $\rightarrow$ $\mathcal{D} = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{D}|}\right\}$
- Assume $p(\mathbf{x}|\omega_i)$ has a known parametric form $\Rightarrow$ determined by $\boldsymbol{\theta}_i$

Parameter estimation

- Suppose the existence of a dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_c\}$
  - $\rightarrow$ $\mathcal{D}_i \sim p(\mathbf{x}|\omega_i)$ with $i = 1, \ldots, c$ (i.i.d.)
  - $\rightarrow$ $\mathcal{D} = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{D}|}\right\}$
- Assume $p(\mathbf{x}|\omega_i)$ has a known parametric form $\Rightarrow$ determined by $\boldsymbol{\theta}_i$
  - $\rightarrow$ Distribution is written as $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$
  - $\rightarrow$ Samples in $\mathcal{D}_i$ give no information about $\boldsymbol{\theta}_j$ if $j \neq i$

Parameter estimation

- Suppose the existence of a dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_c\}$
  - $\rightarrow$ $\mathcal{D}_i \sim p(\mathbf{x}|\omega_i)$ with $i = 1, \ldots, c$ (i.i.d.)
  - $\rightarrow$ $\mathcal{D} = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{D}|}\right\}$
- Assume $p(\mathbf{x}|\omega_i)$ has a known parametric form $\Rightarrow$ determined by $\boldsymbol{\theta}_i$
  - $\rightarrow$ Distribution is written as $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$
  - $\rightarrow$ Samples in $\mathcal{D}_i$ give no information about $\boldsymbol{\theta}_j$ if $j \neq i$
- We have c separate problems: use a set $\mathcal{D}$ of training samples drawn independently from the probability density $p(\mathbf{x}|\boldsymbol{\theta})$ to estimate the unknown parameter vector $\boldsymbol{\theta}$

Parameter estimation

- Suppose the existence of a dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_c\}$
  - $\rightarrow$ $\mathcal{D}_i \sim p(\mathbf{x}|\omega_i)$ with $i = 1, \ldots, c$ (i.i.d.)
  - $\rightarrow$ $\mathcal{D} = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{D}|}\right\}$
- Assume $p(\mathbf{x}|\omega_i)$ has a known parametric form $\Rightarrow$ determined by $\boldsymbol{\theta}_i$
  - $\rightarrow$ Distribution is written as $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$
  - $\rightarrow$ Samples in $\mathcal{D}_i$ give no information about $\boldsymbol{\theta}_j$ if $j \neq i$
- We have c separate problems: use a set $\mathcal{D}$ of training samples drawn independently from the probability density $p(\mathbf{x}|\boldsymbol{\theta})$ to estimate the unknown parameter vector $\boldsymbol{\theta}$

$\Rightarrow$ What about **prior** $P(\omega)$?

Parameter estimation

Likelihood

The function that measures how likely is to obtain precisely this dataset, given the parameters is the so-called **likelihood** of $\boldsymbol{\theta}$ with regard to $\mathcal{D}$.

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{|\mathcal{D}|} p(\mathbf{x}_n|\boldsymbol{\theta})$$

Recall: Samples are drawn i.i.d. from the original data distribution

Parameter estimation

Likelihood

The function that measures how likely is to obtain precisely this dataset, given the parameters is the so-called **likelihood** of $\boldsymbol{\theta}$ with regard to $\mathcal{D}$.

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{|\mathcal{D}|} p(\mathbf{x}_n|\boldsymbol{\theta})$$

Recall: Samples are drawn i.i.d. from the original data distribution

- **Task**: Find $\hat{\boldsymbol{\theta}}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$

Parameter estimation

Likelihood

The function that measures how likely is to obtain precisely this dataset, given the parameters is the so-called **likelihood** of $\boldsymbol{\theta}$ with regard to $\mathcal{D}$.

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{|\mathcal{D}|} p(\mathbf{x}_n|\boldsymbol{\theta})$$

Recall: Samples are drawn i.i.d. from the original data distribution

- **Task**: Find $\hat{\boldsymbol{\theta}}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$
  $\rightarrow$ Maximum Likelihood Estimation (MLE)

Parameter estimation

Likelihood

The function that measures how likely is to obtain precisely this dataset, given the parameters is the so-called **likelihood** of $\boldsymbol{\theta}$ with regard to $\mathcal{D}$.

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{|\mathcal{D}|} p(\mathbf{x}_n|\boldsymbol{\theta})$$

Recall: Samples are drawn i.i.d. from the original data distribution

- **Task**: Find $\hat{\boldsymbol{\theta}}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$
  - $\rightarrow$ Maximum Likelihood Estimation (MLE)
  - $\rightarrow$ In practise *log-likelihood* function: $l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta})$

Maximum Likelihood Estimation

- The MLE problem may be formally described as:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \ln p(\mathcal{D}|\boldsymbol{\theta})$$

Maximum Likelihood Estimation

- The MLE problem may be formally described as:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \ln p(\mathcal{D}|\boldsymbol{\theta})$$

- The set of necessary conditions for the MLE of $\boldsymbol{\theta}$ can be obtained from the set of $|\boldsymbol{\theta}|$ equations:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta}) = 0 \ \text{ where } \ \boldsymbol{\nabla}_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \dots \\ \frac{\partial}{\partial \theta_{|\boldsymbol{\theta}|}} \end{bmatrix}$$

Gaussian case

Univariate case

Gaussian distribution with $\boldsymbol{\theta} = (\mu, \sigma^2)$ where $\mathbf{x} \in \mathbb{R} \equiv x$:

$$\hat{\mu} = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{D}|} x_n \qquad \hat{\sigma^2} = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{D}|} (x_n - \hat{\mu})^2$$

Multivariate case

Gaussian distribution with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathbf{x} \in \mathbb{R}^d$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{D}|} \mathbf{x}_n \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{D}|} (\mathbf{x}_n - \hat{\boldsymbol{\mu}})\,(\mathbf{x}_n - \hat{\boldsymbol{\mu}})^t$$