Bayes Classification

Naresh Manwani

Machine Learning Lab, IIIT Hyderabad

Bayes Decision Theory

- Design classifiers that make decisions using the Bayes rule.
- "Best" decision is chosen by minimizing some expected "risk".
  - The simplest risk is the classification error.

Bayes Decision Theory

- Design classifiers that make decisions using the Bayes rule.
- "Best" decision is chosen by minimizing some expected "risk".
  - The simplest risk is the classification error.
  - Risk can also include different types of actions as well costs associated with different misclassification errors.

Bayes Decision Theory

- Design classifiers that make decisions using the Bayes rule.
- "Best" decision is chosen by minimizing some expected "risk".
  - The simplest risk is the classification error.
  - Risk can also include different types of actions as well costs associated with different misclassification errors.
  - This is critical when some errors are more serious than others (e.g., intruder detection).

Bayes Decision Theory

Bayes Decision Theory

- **Bayes Formula**

$$P(\omega_j \mid x) = \frac{p(\omega_j, x)}{p(x)} = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

$$=. \frac{likelihood \times prior}{evidence}$$

Bayes Decision Theory

- **Bayes Formula**

$$P(\omega_j \mid x) = \frac{p(\omega_j, x)}{p(x)} = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

$$=. \frac{likelihood \times prior}{evidence}$$

Reflects our knowledge of the problem, which comes from "subject matter expertise

Bayes Decision Theory

- **Bayes Formula**

$$P(\omega_j \mid x) = \frac{p(\omega_j, x)}{p(x)} = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

$$=. \frac{likelihood \times prior}{evidence}$$

Reflects our knowledge of the problem, which comes from "subject matter expertise

probability that feature vector x could have occurred from class ω_j.

Bayes Decision Theory

- **Bayes Formula**

$$P(\omega_j \mid x) = \frac{p(\omega_j, x)}{p(x)} = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

$$=. \frac{likelihood \times prior}{evidence}$$

Reflects our knowledge of the problem, which comes from "subject matter expertise

Probability that class ω_j occurred given feature vector x.

probability that feature vector x could have occurred from class ω_j.

Bayes Decision Theory

Bayes Decision Theory

- Evidence act as normalizing term as:

$$p(x) = \sum_{i=1,2} p(\omega_i, x) = \sum_{i=1,2} p(x \mid \omega_i) P(\omega_i)$$

Bayes Decision Theory

- Evidence act as normalizing term as:

$$p(x) = \sum_{i=1,2} p(\omega_i, x) = \sum_{i=1,2} p(x \mid \omega_i) P(\omega_i)$$

- Bayes Formula (Two Category)

Bayes Decision Theory

- Evidence act as normalizing term as:

$$p(x) = \sum_{i=1,2} p(\omega_i, x) = \sum_{i=1,2} p(x \mid \omega_i) P(\omega_i)$$

- Bayes Formula (Two Category)

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)} = \frac{p(x \mid \omega_j) P(\omega_j)}{\sum_{i=1,2} p(x \mid \omega_i) P(\omega_i)}$$

Bayes Decision Rule

- Using Bayes' rule:

Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise **decide** ω_2

Or

Decide ω_1 if $p(x \mid \omega_1)P(\omega_1) > p(x \mid \omega_2)P(\omega_2)$; otherwise **decide** ω_2

Bayes Decision Rule

- Using Bayes' rule:

Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise **decide** ω_2

Or

Decide ω_1 if $p(x \mid \omega_1)P(\omega_1) > p(x \mid \omega_2)P(\omega_2)$; otherwise **decide** ω_2

Or

Decide ω_1 if $\dfrac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2

Bayes Decision Rule

- Using Bayes' rule:

Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise **decide** ω_2

Or

Decide ω_1 if $p(x \mid \omega_1)P(\omega_1) > p(x \mid \omega_2)P(\omega_2)$; otherwise **decide** ω_2

Or

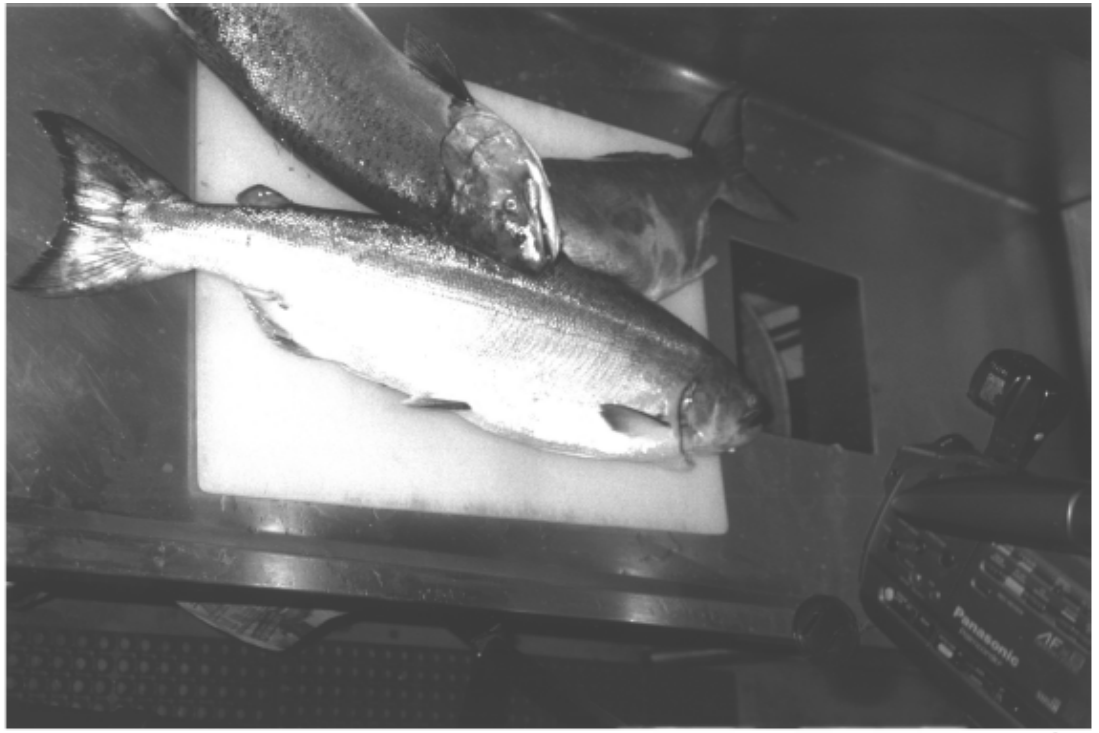
Decide ω_1 if $\dfrac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2

Likelihood Ratio

Threshold

Toy Example Walkthrough

- Image Based Fish Classification (Salmon v/s Sea Bass)



Prior knowledge based classification

Prior knowledge based classification

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2 (i.e., solmon)
- **Prior knowledge based classification**

Prior knowledge based classification

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2 (i.e., solmon)
- **Prior knowledge based classification**
  - Let $P(\omega_1)$ and $P(\omega_2)$ be the class prior probabilities of the next fish on conveyer belt being sea bass and solmon, respectively. .

Prior knowledge based classification

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2 (i.e., solmon)
- **Prior knowledge based classification**
  - Let $P(\omega_1)$ and $P(\omega_2)$ be the class prior probabilities of the next fish on conveyer belt being sea bass and solmon, respectively. .
  - $P(\omega_1)+ P(\omega_2) = 1$ i.e., only if two types of fishes are caught

Prior knowledge based classification

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2 (i.e., solmon)
- **Prior knowledge based classification**
  - Let $P(\omega_1)$ and $P(\omega_2)$ be the class prior probabilities of the next fish on conveyer belt being sea bass and solmon, respectively. .
  - $P(\omega_1)+ P(\omega_2) = 1$ i.e., only if two types of fishes are caught
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ else ω_2. (**Decision Rule**)

Prior knowledge based classification

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2 (i.e., solmon)
- **Prior knowledge based classification**
  - Let $P(\omega_1)$ and $P(\omega_2)$ be the class prior probabilities of the next fish on conveyer belt being sea bass and solmon, respectively. .
  - $P(\omega_1)+ P(\omega_2) = 1$ i.e., only if two types of fishes are caught
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ else ω_2. (**Decision Rule**)
  - Will be making the same decision all times!

Prior knowledge based classification

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2 (i.e., solmon)
- **Prior knowledge based classification**
  - Let $P(\omega_1)$ and $P(\omega_2)$ be the class prior probabilities of the next fish on conveyer belt being sea bass and solmon, respectively. .
  - $P(\omega_1)+ P(\omega_2) = 1$ i.e., only if two types of fishes are caught
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ else ω_2. (**Decision Rule**)
  - Will be making the same decision all times!
  - Favours the most likely class.

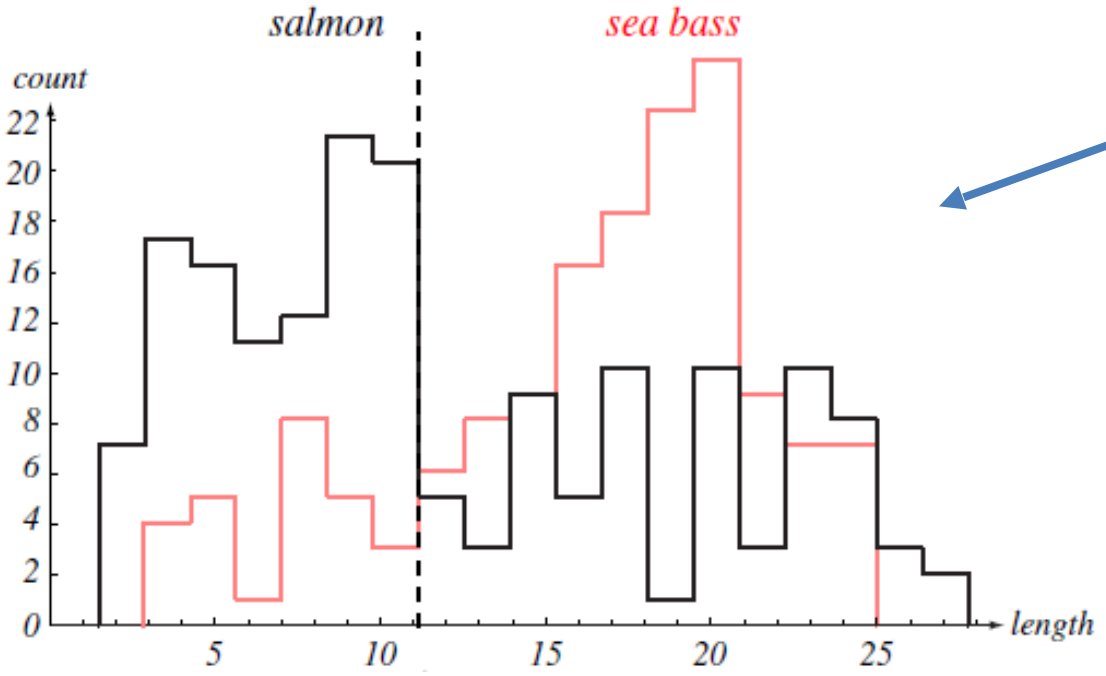
Prior knowledge based classification

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2 (i.e., solmon)
- **Prior knowledge based classification**
  - Let $P(\omega_1)$ and $P(\omega_2)$ be the class prior probabilities of the next fish on conveyer belt being sea bass and solmon, respectively. .
  - $P(\omega_1)+ P(\omega_2) = 1$ i.e., only if two types of fishes are caught
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ else ω_2. (**Decision Rule**)
  - Will be making the same decision all times!
  - Favours the most likely class.
- But we can use more information !!

Features and Feature Spaces

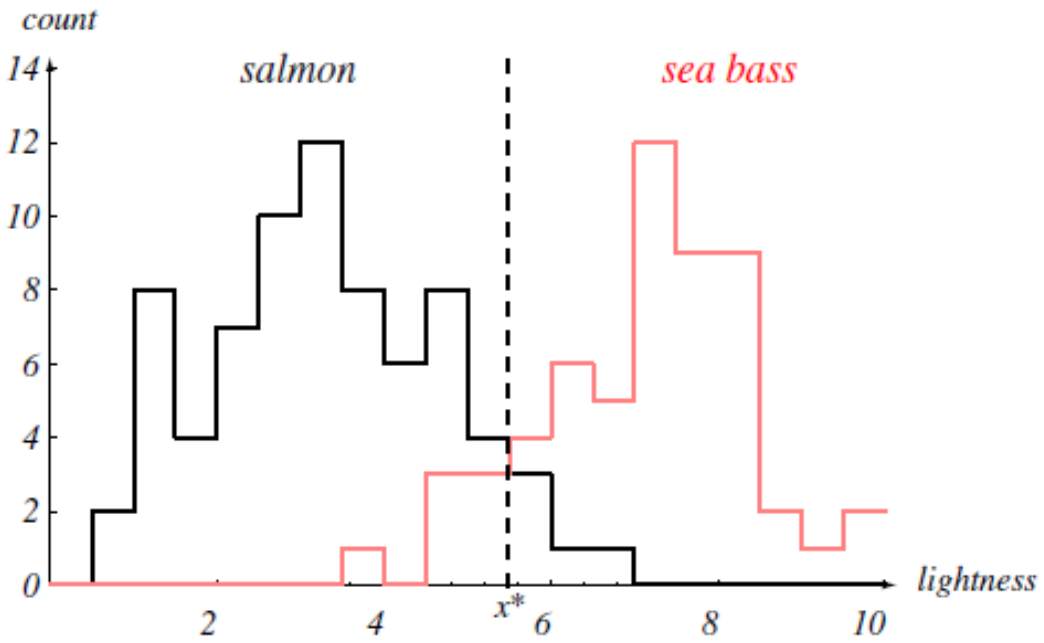
- A feature is an observable variable.
- A feature space is a set from which we can sample or observe values. Examples of features:
  - Length
  - Width
  - Lightness
  - Location of Dorsal Fin
- For simplicity, let's assume that our features are all continuous values.
- Denote a scalar feature as x and a vector feature as $\mathbf{x}$. For a d-dimensional feature space, $\mathbf{x} \in \mathbb{R}^d$.

Toy Examples Walkthrough



Length Histogram Feature

Lightness Histogram Feature



Toy Example Walkthrough

Toy Example Walkthrough

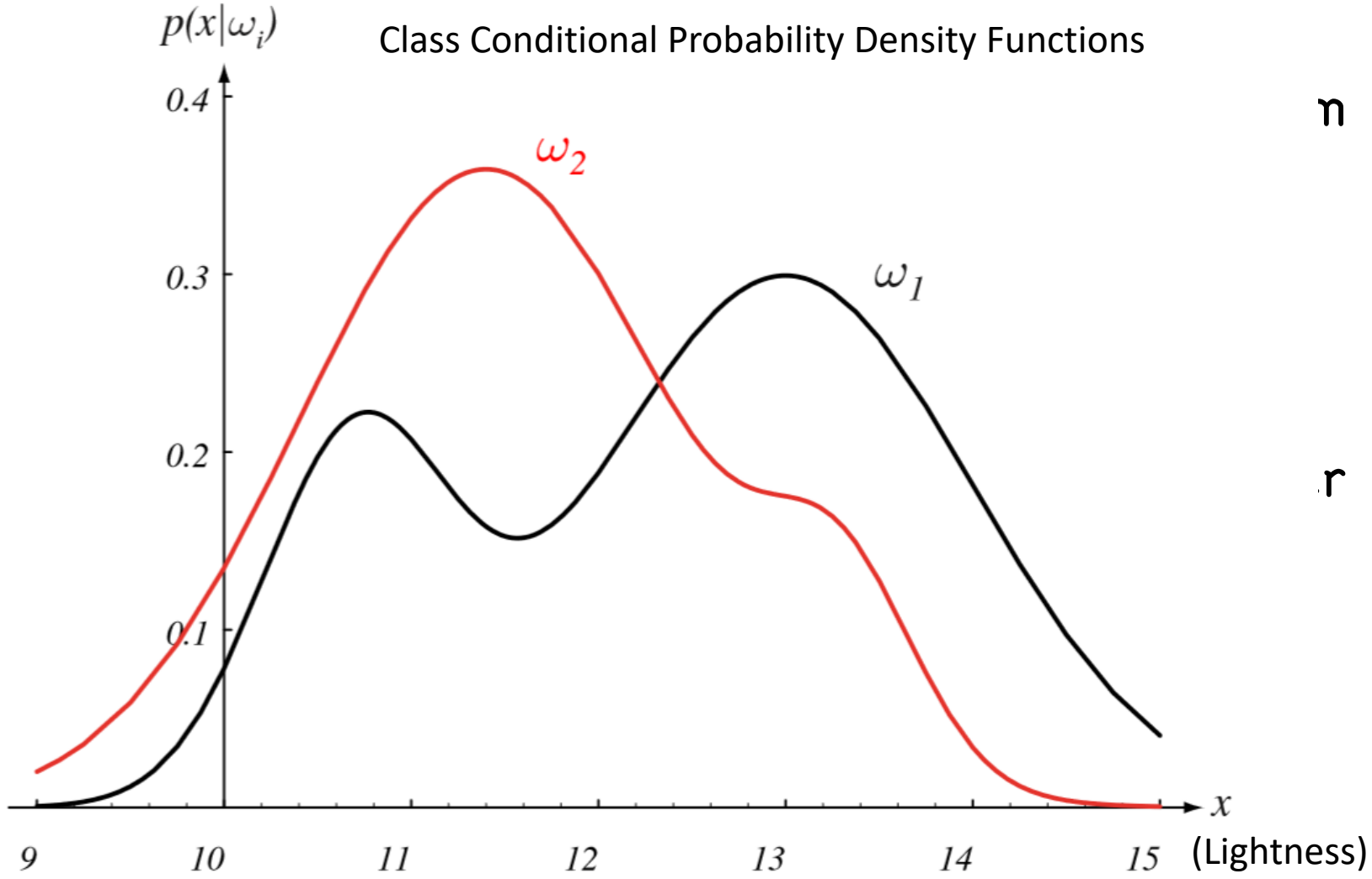
- Each feature can be represented as continuous random variable.
- Instead of histogram representation, we can visualize overlapped class conditional probability density functions
- These pdf curves are normalized such that area under each curve is strictly one.

Toy Example Walkthrough

- Each feature can be represented as continuous random variable.
- Instead of histogram representation, we can visualize overlapped class conditional probability density functions
- These pdf curves are normalized such that area under each curve is strictly one.
- These density plots can very well be interpreted as continuous counterpart of histogram representation.

Toy Example Walkthrough

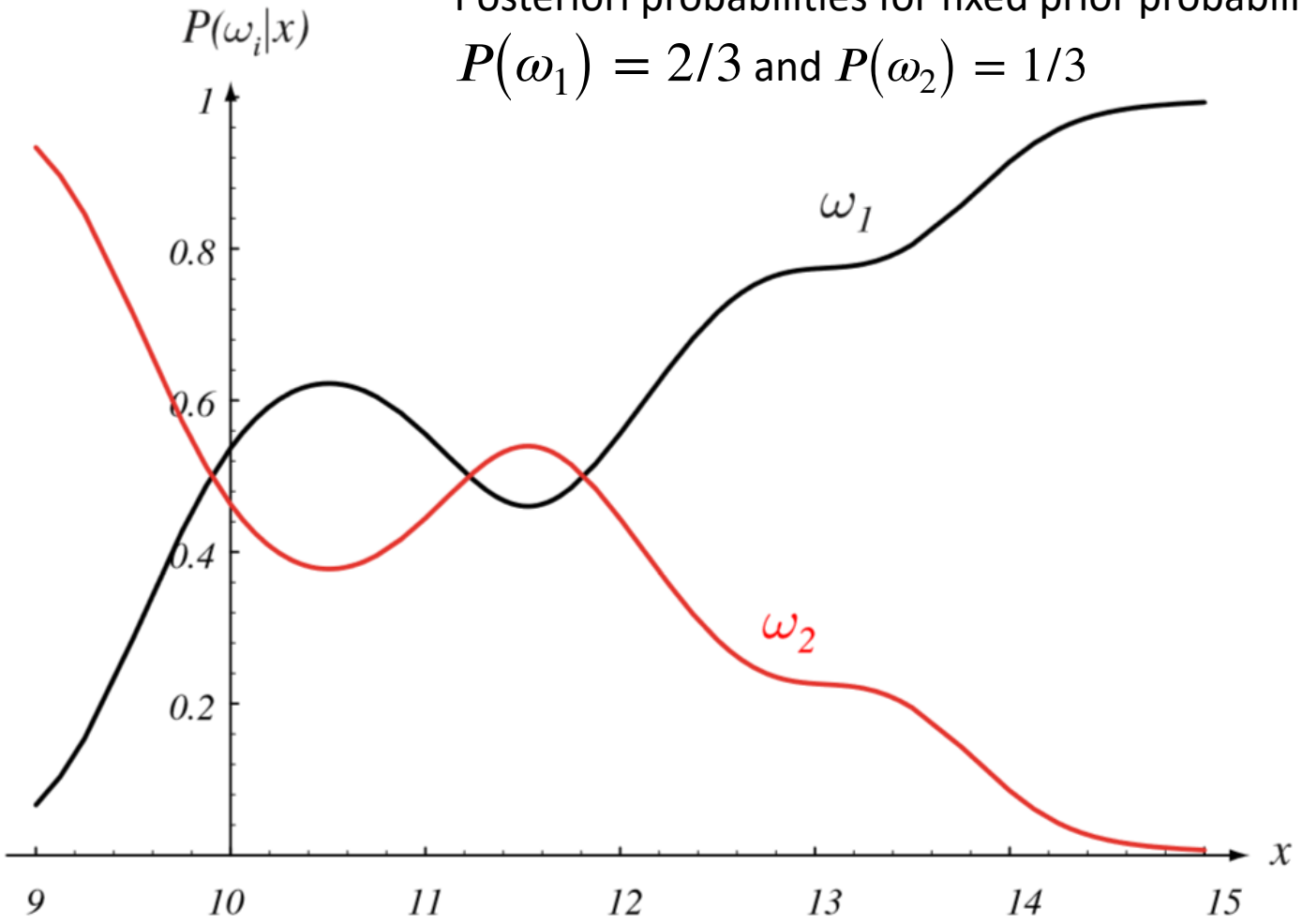
Class Conditional Probability Density Functions



Bayes Classifier

Posteriori probabilities for fixed prior probabilities

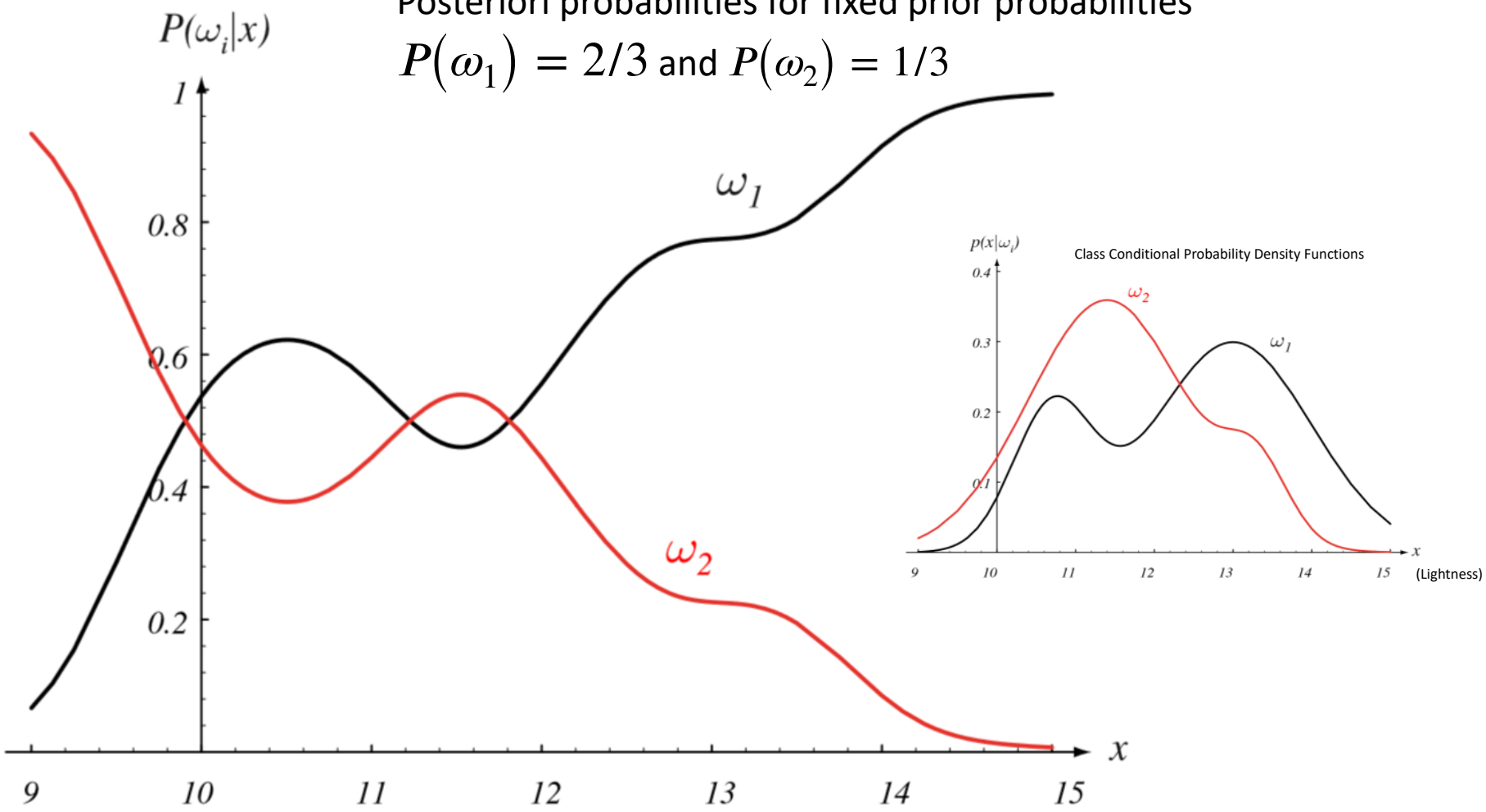
$P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$



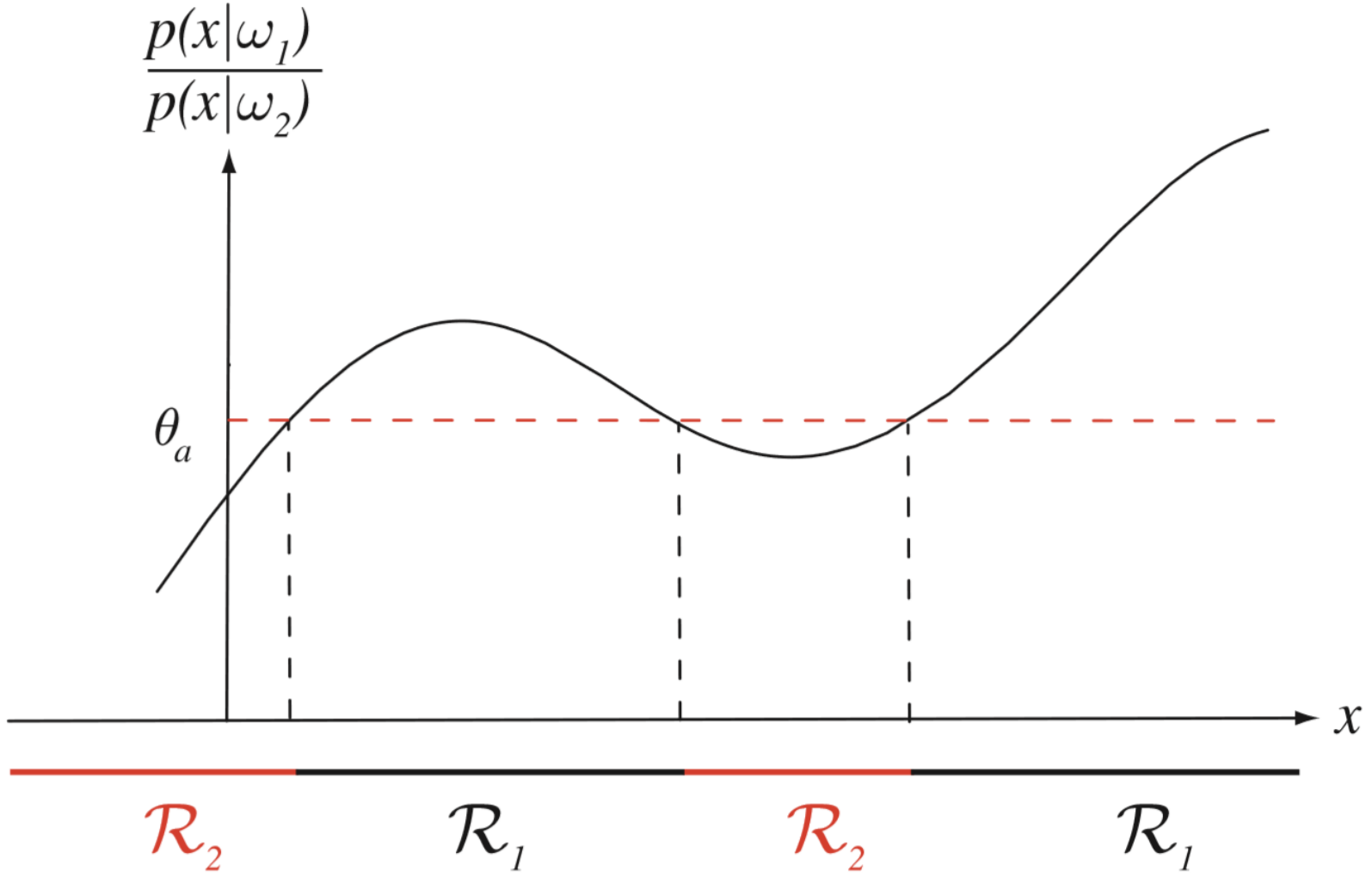
Bayes Classifier

Posteriori probabilities for fixed prior probabilities

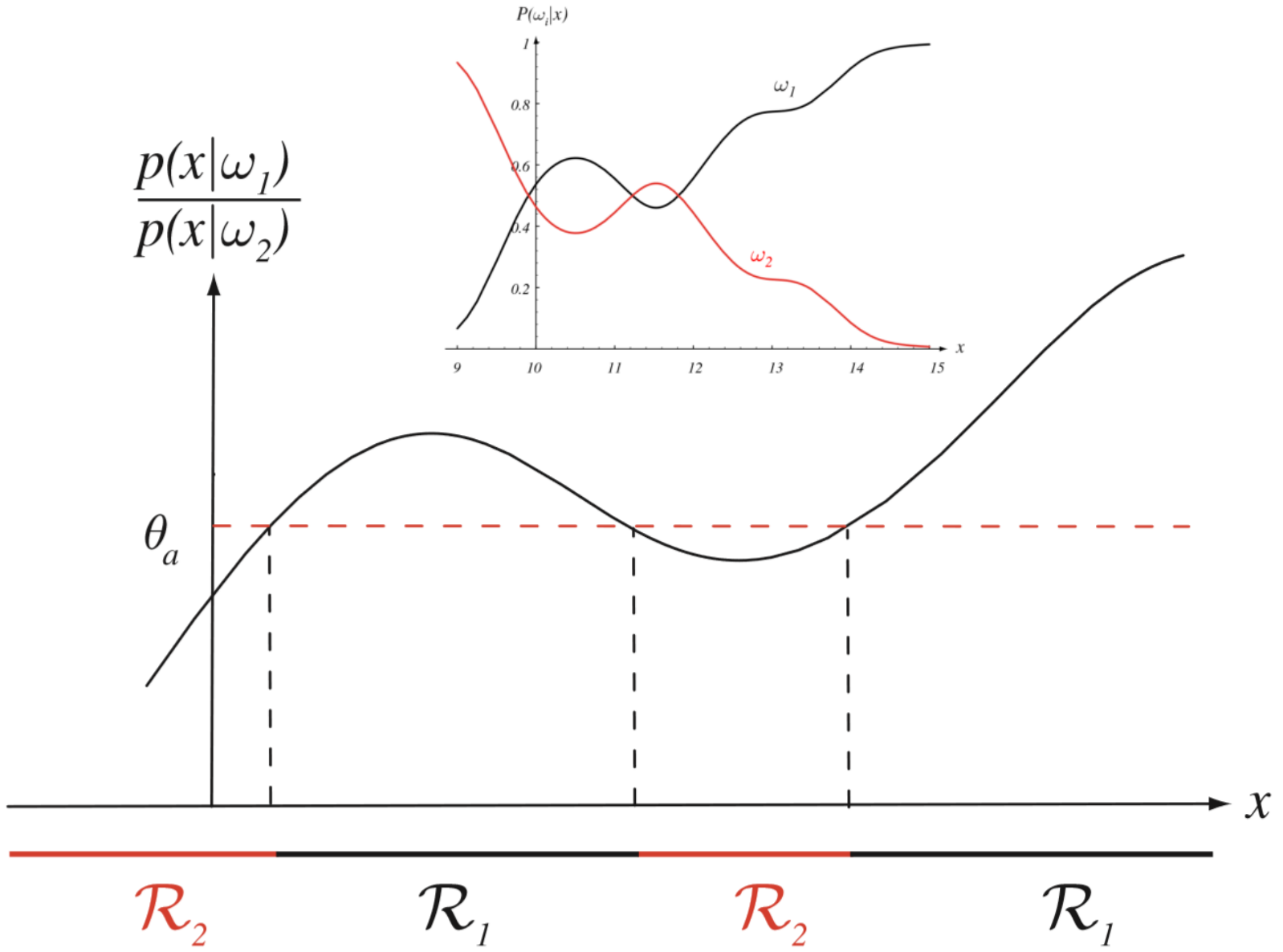
$P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$



Bayesian Decision Theory



Bayesian Decision Theory



Bayesian Decision Theory for Multiclass Classification

Bayesian Decision Theory for Multiclass Classification

- Bayes formula can be generalized to:
  - Multi-dimensional feature space i.e., $\mathbf{x} = [x_1, \quad \cdots, \quad x_d]^T$
  - Multiple classes i.e., $\{\omega_1, \ldots, \omega_c\}$

Bayesian Decision Theory for Multiclass Classification

- Bayes formula can be generalized to:
  - Multi-dimensional feature space i.e., $\mathbf{x} = [x_1, \quad \cdots, \quad x_d]^T$
  - Multiple classes i.e., $\{\omega_1, \ldots, \omega_c\}$

Bayesian Decision Theory for Multiclass Classification

- Bayes formula can be generalized to:
  - Multi-dimensional feature space i.e., $\mathbf{x} = [x_1, \quad \cdots, \quad x_d]^T$
  - Multiple classes i.e., $\{\omega_1, \ldots, \omega_c\}$

- $$P\left(\omega_j \middle| \mathbf{x}\right) = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{p(\mathbf{x})} = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{\sum_{i=1}^{c} p\left(\mathbf{x} \middle| \omega_i\right) P\left(\omega_i\right)}$$

Bayesian Decision Theory for Multiclass Classification

- Bayes formula can be generalized to:
  - Multi-dimensional feature space i.e., $\mathbf{x} = [x_1, \quad \cdots, \quad x_d]^T$
  - Multiple classes i.e., $\{\omega_1, \ldots, \omega_c\}$

- $$P\left(\omega_j \middle| \mathbf{x}\right) = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{p(\mathbf{x})} = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{\sum_{i=1}^{c} p\left(\mathbf{x} \middle| \omega_i\right) P\left(\omega_i\right)}$$

Bayesian Decision Theory for Multiclass Classification

- Bayes formula can be generalized to:
  - Multi-dimensional feature space i.e., $\mathbf{x} = [x_1, \quad \cdots, \quad x_d]^T$
  - Multiple classes i.e., $\{\omega_1, \ldots, \omega_c\}$

- $$P\left(\omega_j \middle| \mathbf{x}\right) = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{p(\mathbf{x})} = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{\sum_{i=1}^{c} p\left(\mathbf{x} \middle| \omega_i\right) P\left(\omega_i\right)}$$

- Decide class ω_k if $k = \arg max_{j \in \{1,\ldots,c\}} P(\omega_j | \mathbf{x})$

Bayesian Decision Theory for Multiclass Classification

- Bayes formula can be generalized to:
  - Multi-dimensional feature space i.e., $\mathbf{x} = [x_1, \quad \cdots, \quad x_d]^T$
  - Multiple classes i.e., $\{\omega_1, \ldots, \omega_c\}$

- $$P\left(\omega_j \middle| \mathbf{x}\right) = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{p(\mathbf{x})} = \frac{p\left(\mathbf{x} \middle| \omega_j\right) P\left(\omega_j\right)}{\sum_{i=1}^{c} p\left(\mathbf{x} \middle| \omega_i\right) P\left(\omega_i\right)}$$

- Decide class ω_k if $k = \arg max_{j \in \{1,\ldots,c\}} P(\omega_j | \mathbf{x})$

Questions

Optimality of Bayes Classifier

Probability of Error

Probability of Error

1. For a given observation x, we would be inclined to let the posterior govern our decision:
2. What is our probability of error?

Probability of Error

1. For a given observation x, we would be inclined to let the posterior govern our decision:
2. What is our probability of error?
3. For the two class situation, we have

Probability of Error

1. For a given observation x, we would be inclined to let the posterior govern our decision:
2. What is our probability of error?
3. For the two class situation, we have

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

Terminology

Terminology

- A set of c categories $\omega_1, \omega_2, \ldots, \omega_c$
- A set of c actions $\alpha_1, \alpha_2, \ldots, \alpha_c$

Terminology

- A set of c categories $\omega_1, \omega_2, \ldots, \omega_c$
- A set of c actions $\alpha_1, \alpha_2, \ldots, \alpha_c$
- A loss function $\lambda_{ij} = \lambda\left(\alpha_i \middle| \omega_j\right)$

Terminology

- A set of c categories $\omega_1, \omega_2, \ldots, \omega_c$
- A set of c actions $\alpha_1, \alpha_2, \ldots, \alpha_c$
- A loss function $\lambda_{ij} = \lambda\left(\alpha_i \middle| \omega_j\right)$
  - i.e., the cost associated with taking action α_i when the correct classification category is ω_j (user-defined)

Terminology

- A set of c categories $\omega_1, \omega_2, \ldots, \omega_c$
- A set of c actions $\alpha_1, \alpha_2, \ldots, \alpha_c$
- A loss function $\lambda_{ij} = \lambda\left(\alpha_i \middle| \omega_j\right)$
  - i.e., the cost associated with taking action α_i when the correct classification category is ω_j (user-defined)
- Conditional risk $R(\alpha_i \,|\, \mathbf{x})$ – expected loss of taking action α_i given **x**

Terminology

- A set of c categories $\omega_1, \omega_2, \ldots, \omega_c$
- A set of c actions $\alpha_1, \alpha_2, \ldots, \alpha_c$
- A loss function $\lambda_{ij} = \lambda\left(\alpha_i \middle| \omega_j\right)$
  - i.e., the cost associated with taking action α_i when the correct classification category is ω_j (user-defined)
- Conditional risk $R(\alpha_i \,|\, \mathbf{x})$ – expected loss of taking action α_i given **x**
- Classification is performed by minimizing $R(\alpha_i \,|\, \mathbf{x})$ instead of maximizing $P(\omega_i \,|\, \mathbf{x})$

Conditional Risk $R(\alpha_i \mid \mathbf{x})$

Conditional Risk $R(\alpha_i \mid \mathbf{x})$

- Suppose we take action α_i when $\mathbf{x}$ is observed.
- The conditional risk is defined as the expected loss of taking action α_i given $\mathbf{x}$:

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

Conditional Risk $R(\alpha_i \mid \mathbf{x})$

- Suppose we take action α_i when $\mathbf{x}$ is observed.
- The conditional risk is defined as the expected loss of taking action α_i given $\mathbf{x}$:

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

- where $P(\omega_j \mid \mathbf{x}) = \dfrac{P(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})}$

Overall Risk

Overall Risk

- The overall risk is defined as:

$$R = \int_{\mathbf{x}} R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- where $\alpha(\mathbf{x})$ is a general decision rule which determines which action $\alpha_1, \alpha_2, \ldots, \alpha_c$ to take for any $\mathbf{x}$.

Overall Risk

- The overall risk is defined as:

$$R = \int_{\mathbf{x}} R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- where $\alpha(\mathbf{x})$ is a general decision rule which determines which action $\alpha_1, \alpha_2, \ldots, \alpha_c$ to take for any **x**.
  - i.e., $\alpha(\mathbf{x})$ depends on the classifier being used.

Overall Risk

- The overall risk is defined as:

$$R = \int_{\mathbf{x}} R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- where $\alpha(\mathbf{x})$ is a general decision rule which determines which action $\alpha_1, \alpha_2, \ldots, \alpha_c$ to take for any **x**.
  - i.e., $\alpha(\mathbf{x})$ depends on the classifier being used.

Decision Rule Using Conditional Risk

Decision Rule Using Conditional Risk

- The Bayes rule minimizes R by:
  (i) Computing $R(\alpha_i \mid \mathbf{x})$ for every α_i given an **x**

Decision Rule Using Conditional Risk

- The Bayes rule minimizes R by:

  (i) Computing $R(\alpha_i \mid \mathbf{x})$ for every α_i given an **x**

  (ii) Choosing the action α_i with the minimum conditional risk

Decision Rule Using Conditional Risk

- The Bayes rule minimizes R by:

  (i) Computing $R(\alpha_i \mid \mathbf{x})$ for every α_i given an **x**

  (ii) Choosing the action α_i with the minimum conditional risk $R(\alpha_i \mid \mathbf{x})$

- The resulting minimum R* is called Bayes risk and is the best performance that can be achieved:

Example: Two Category Classification

Example: Two Category Classification

- Define
  - α_1: decide ω_1
  - $\alpha_{2:}$ decide ω_2

Example: Two Category Classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

Example: Two Category Classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- Risk of predicting class α_i for example $\mathbf{x}$.

Example: Two Category Classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- Risk of predicting class α_i for example $\mathbf{x}$.

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

Example: Two Category Classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- Risk of predicting class α_i for example $\mathbf{x}$.

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

- Two-Category Classification

Example: Two Category Classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- Risk of predicting class α_i for example $\mathbf{x}$.

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

- Two-Category Classification

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

Example: Two Category Classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- Risk of predicting class α_i for example $\mathbf{x}$.

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

- Two-Category Classification

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

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

Example: Two Category Classification (Minimum risk decision rule)

Example: Two Category Classification (Minimum risk decision rule)

- **Choose** ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

or

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

Example: Two Category Classification (Minimum risk decision rule)

- **Choose** ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

  or

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

  or

Example: Two Category Classification (Minimum risk decision rule)

– Choose ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

or

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

or

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

Example: Two Category Classification (Minimum risk decision rule)

– **Choose** ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

or

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

or

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

or

Example: Two Category Classification (Minimum risk decision rule)

– **Choose** ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

or

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

or

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

or

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

Example: Two Category Classification (Minimum risk decision rule)

– **Choose** ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

or

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

or

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

or

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

Likelihood Ratio

Threshold

Special Case: 0-1 Loss

Special Case: 0-1 Loss

- Let $\lambda_{ij} = \lambda\left(\alpha_i \middle| \omega_j\right) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

Special Case: 0-1 Loss

- Let $\lambda_{ij} = \lambda\left(\alpha_i \middle| \omega_j\right) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

Special Case: 0-1 Loss

- Let $\lambda_{ij} = \lambda\left(\alpha_i \middle| \omega_j\right) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

- $$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{2} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \\ &= \sum_{j=1}^{2} \mathbb{I}[j \neq i] P(\omega_j|\mathbf{x}) \\ &= 1 - P(\omega_j|\mathbf{x}) \end{aligned}$$

Special Case: Zero-One Loss Function (cont'd)

Special Case: Zero-One Loss Function (cont'd)

- If we choose ω_i corresponding to largest $P(\omega_i \mid \mathbf{x})$ then we minimize $R(\alpha_i \mid \mathbf{x})$
- Decision rule:

Special Case: Zero-One Loss Function (cont'd)

- If we choose ω_i corresponding to largest $P(\omega_i \mid \mathbf{x})$ then we minimize $R(\alpha_i \mid \mathbf{x})$

- Decision rule:
  Decide ω_1 if $R(\omega_1 \mid \mathbf{x}) < R(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

Special Case: Zero-One Loss Function (cont'd)

- If we choose ω_i corresponding to largest $P(\omega_i \mid \mathbf{x})$ then we minimize $R(\alpha_i \mid \mathbf{x})$

- Decision rule:
  Decide ω_1 if $R(\omega_1 \mid \mathbf{x}) < R(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2
  Or

Special Case: Zero-One Loss Function (cont'd)

- If we choose ω_i corresponding to largest $P(\omega_i \mid \mathbf{x})$ then we minimize $R(\alpha_i \mid \mathbf{x})$

- Decision rule:

  Decide ω_1 if $R(\omega_1 \mid \mathbf{x}) < R(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

  Or

  Decide ω_1 if $1 - P(\omega_1 \mid \mathbf{x}) < 1 - P(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

Special Case: Zero-One Loss Function (cont'd)

- If we choose ω_i corresponding to largest $P(\omega_i \mid \mathbf{x})$ then we minimize $R(\alpha_i \mid \mathbf{x})$

- Decision rule:

  Decide ω_1 if $R(\omega_1 \mid \mathbf{x}) < R(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

  Or

  Decide ω_1 if $1 - P(\omega_1 \mid \mathbf{x}) < 1 - P(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

  Or

Special Case: Zero-One Loss Function (cont'd)

- If we choose ω_i corresponding to largest $P(\omega_i \mid \mathbf{x})$ then we minimize $R(\alpha_i \mid \mathbf{x})$

- Decision rule:

  Decide ω_1 if $R(\omega_1 \mid \mathbf{x}) < R(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

  Or

  Decide ω_1 if $1 - P(\omega_1 \mid \mathbf{x}) < 1 - P(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

  Or

  Decide ω_1 if $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

Special Case: Zero-One Loss Function (cont'd)

- If we choose ω_i corresponding to largest $P(\omega_i \mid \mathbf{x})$ then we minimize $R(\alpha_i \mid \mathbf{x})$

- Decision rule:

  Decide ω_1 if $R(\omega_1 \mid \mathbf{x}) < R(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

  Or

  Decide ω_1 if $1 - P(\omega_1 \mid \mathbf{x}) < 1 - P(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

  Or

  Decide ω_1 if $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$; otherwise decide ω_2

- The overall risk in this case is the average probability error (which is minimized by the Bayes rule!).

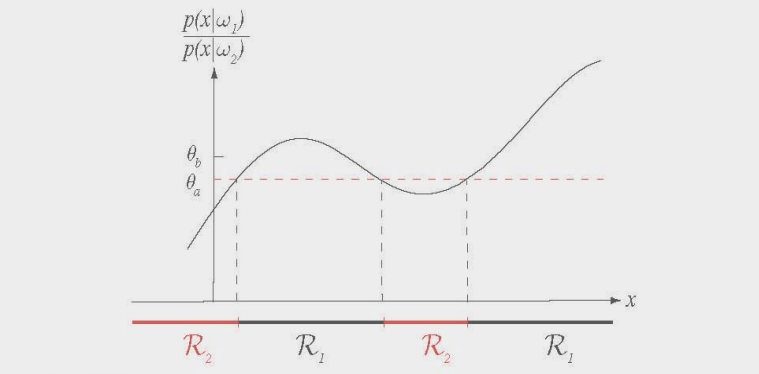
Example

Example

- Assuming a zero-one loss function:
  - **Decide** ω_1 if $\dfrac{p(x|\omega_1)}{p(x|\omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2
- Assuming a general loss function:
  - **Decide** ω_1 if $\dfrac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \dfrac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2

Example

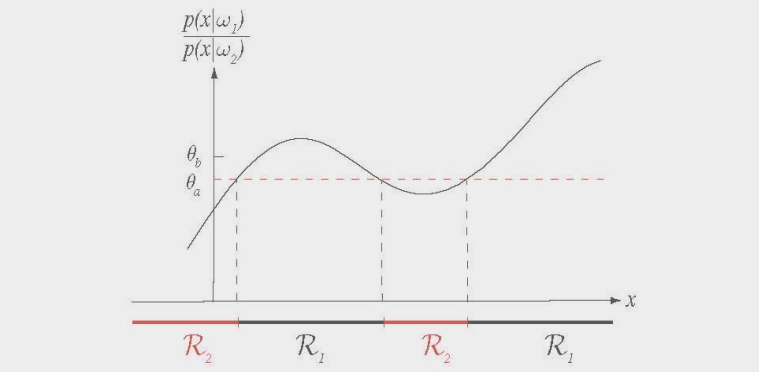
- Assuming a zero-one loss function:
  - **Decide** ω_1 if $\dfrac{p(x\,|\,\omega_1)}{p(x\,|\,\omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2
- Assuming a general loss function:
  - **Decide** ω_1 if $\dfrac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} > \dfrac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2



(decision regions)

Example

- Assuming a zero-one loss function:
  - **Decide** ω_1 if $\dfrac{p(x|\omega_1)}{p(x|\omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2
- Assuming a general loss function:
  - **Decide** ω_1 if $\dfrac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \dfrac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise **decide** ω_2



(decision regions)

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

$$\theta_b = \frac{P(\omega_2)(\lambda_{12}-\lambda_{22})}{P(\omega_1)(\lambda_{21}-\lambda_{11})}$$

Questions

Discriminants for the Bayes Classifier and Decision Surfaces

Discriminant Functions

Discriminant Functions

- Represent a classifier by a set of discriminant functions, one for each class:

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

- A feature vector **x** is assigned to class ω_i if:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i$$

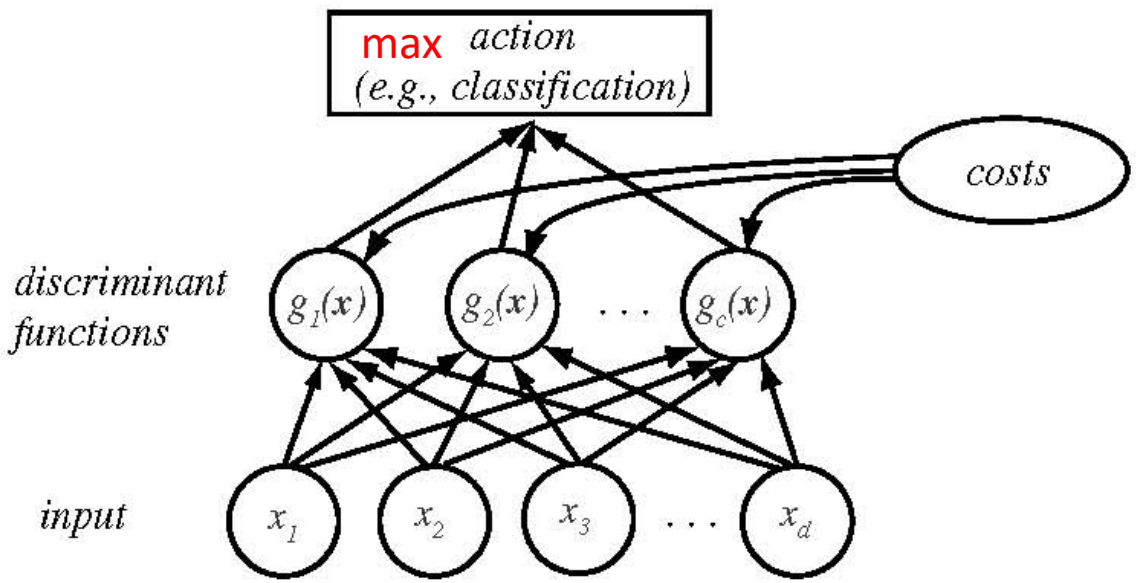
Discriminant Functions

- Represent a classifier by a set of discriminant functions, one for each class:

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

- A feature vector **x** is assigned to class ω_i if:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i$$



Discriminants for the Bayes Classifier

Discriminants for the Bayes Classifier

- Assuming a general loss function $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
- Assuming the zero-one loss function

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

Discriminants for the Bayes Classifier

- Assuming a general loss function $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
- Assuming the zero-one loss function

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

- Ignoring the denominator term, we get $g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$

Discriminants for the Bayes Classifier

- Assuming a general loss function $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
- Assuming the zero-one loss function

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

- Ignoring the denominator term, we get $g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$
- Replacing $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$, where $f()$ is monotonically increasing, the same classification results can be obtained.

Discriminants for the Bayes Classifier

- Assuming a general loss function $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
- Assuming the zero-one loss function

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

- Ignoring the denominator term, we get $g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$
- Replacing $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$, where $f()$ is monotonically increasing, the same classification results can be obtained.
- For example, take $f() = \ln()$, we get

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$$

Discriminants for the Bayes Classifier

- Assuming a general loss function $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
- Assuming the zero-one loss function

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

- Ignoring the denominator term, we get $g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$
- Replacing $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$, where $f()$ is monotonically increasing, the same classification results can be obtained.
- For example, take $f() = \ln()$, we get

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$$

- **We use this form extensively !!!**

Case of two Categories

Case of two Categories

- More common to use a single discriminant function (dichotomizer) instead of two:
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$

Case of two Categories

- More common to use a single discriminant function (dichotomizer) instead of two:
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$
- Decide ω_1 if $g(\mathbf{x}) > 0$; and ω_2 otherwise

Case of two Categories

- More common to use a single discriminant function (dichotomizer) instead of two:
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$
- Decide ω_1 if $g(\mathbf{x}) > 0$; and ω_2 otherwise
- Examples:

Case of two Categories

- More common to use a single discriminant function (dichotomizer) instead of two:
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$
- Decide ω_1 if $g(\mathbf{x}) > 0$; and ω_2 otherwise
- Examples:
  - $g(\mathbf{x}) = P(\omega_1 \mid \mathbf{x}) - P(\omega_2 \mid \mathbf{x})$

Case of two Categories

- More common to use a single discriminant function (dichotomizer) instead of two:
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$
- Decide ω_1 if $g(\mathbf{x}) > 0$; and ω_2 otherwise
- Examples:
  - $g(\mathbf{x}) = P(\omega_1 | \mathbf{x}) - P(\omega_2 | \mathbf{x})$
  - $g(\mathbf{x}) = \ln \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$

Decision Regions and Boundaries

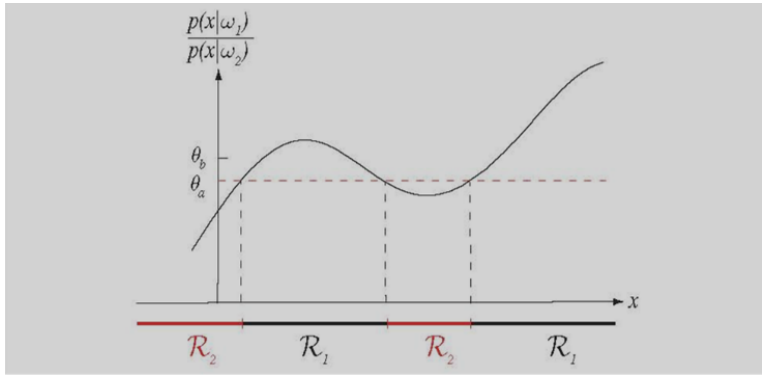
Decision Regions and Boundaries

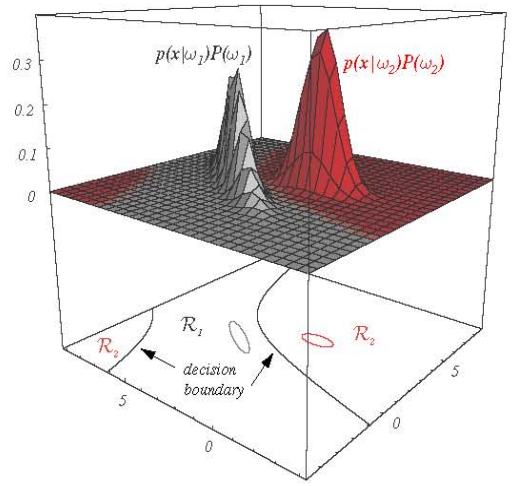
- Discriminants divide a feature space in decision regions R_1, R_2, …, R_c, separated by decision boundaries.

- Next, we will analyze the form of discriminants and decision boundaries (i.e., linear vs non-linear) when $p(\mathbf{x}/\omega_i)$ is modelled by a multivariate Gaussian density.

Decision Regions and Boundaries

- Discriminants divide a feature space in decision regions R_1, R_2, …, R_c, separated by decision boundaries.





How does is the decision boundary defined?

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

- Next, we will analyze the form of discriminants and decision boundaries (i.e., linear vs non-linear) when $p(\mathbf{x}/\omega_i)$ is modelled by a multivariate Gaussian density.

Discriminant Function assuming Multivariate Gaussian Density

Discriminant Function assuming Multivariate Gaussian Density

- Consider the following discriminant function:
  - $g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$
- assuming that $(\mathbf{x} \in \mathbb{R}^d)$

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

Discriminant Function assuming Multivariate Gaussian Density

- Consider the following discriminant function:
  - $g_i(\mathbf{x}) = \ln p(\mathbf{x}\,|\,\omega_i) + \ln P(\omega_i)$
- assuming that $(\mathbf{x} \in \mathbb{R}^d)$

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

- In this case, the discriminant can be written as:

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

Discriminant Function assuming Multivariate Gaussian Density

- Consider the following discriminant function:
  - $g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$
- assuming that $(\mathbf{x} \in \mathbb{R}^d)$

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

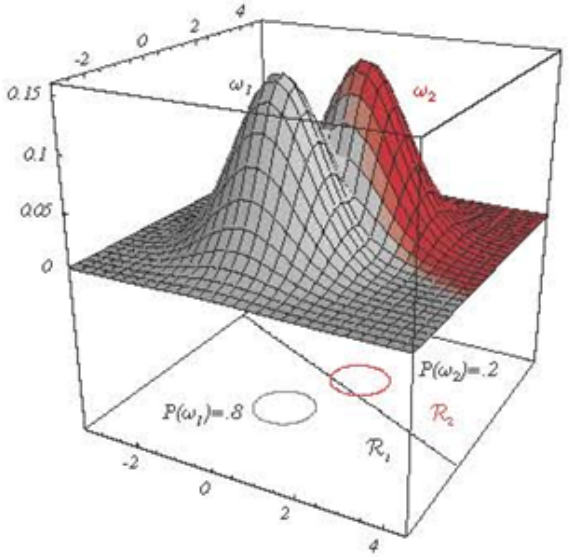
- In this case, the discriminant can be written as:

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

- Model complexity depends on the form of Σ_i; let's consider different cases to better understand this!

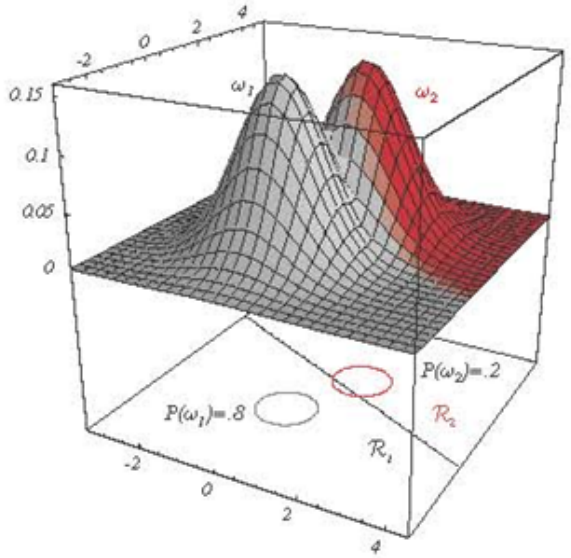
Case I: $\Sigma_i = \sigma^2 I, \forall i$

Case I: $\Sigma_i = \sigma^2 I, \forall i$



Case I: $\Sigma_i = \sigma^2 I, \forall i$

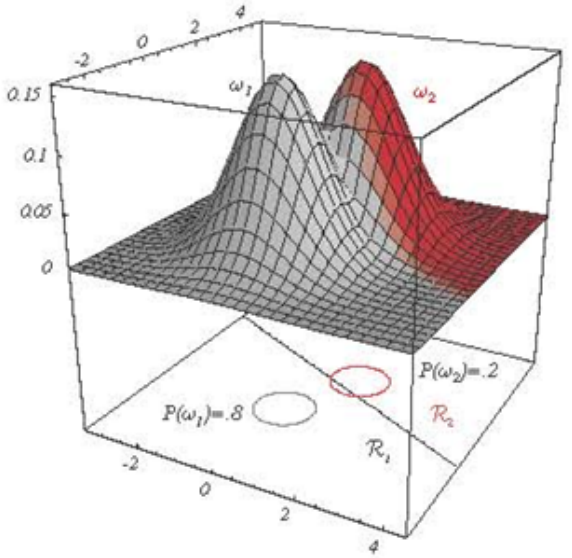
- Each class is modelled by the same diagonal matrix with equal diagonal elements
- This is true when features are uncorrelated (or statistically independent) with same variance



Case I: $\Sigma_i = \sigma^2 I, \forall i$

- Each class is modelled by the same diagonal matrix with equal diagonal elements
- This is true when features are uncorrelated (or statistically independent) with same variance
- How could the discriminant be simplified in this case?

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

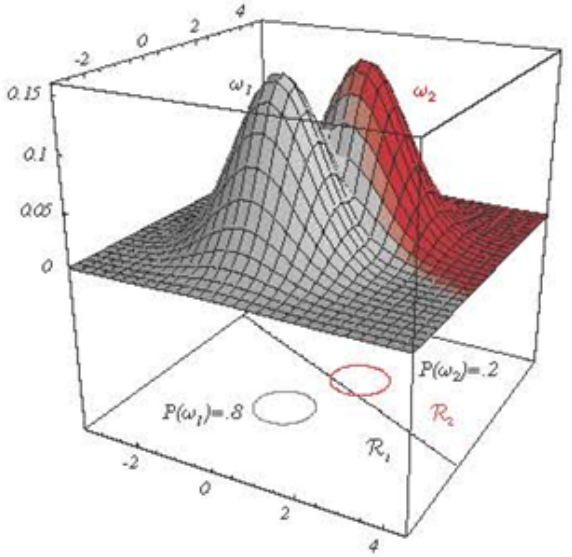


Case I: $\Sigma_i = \sigma^2 I, \forall i$

- Each class is modelled by the same diagonal matrix with equal diagonal elements
- This is true when features are uncorrelated (or statistically independent) with same variance
- How could the discriminant be simplified in this case?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then



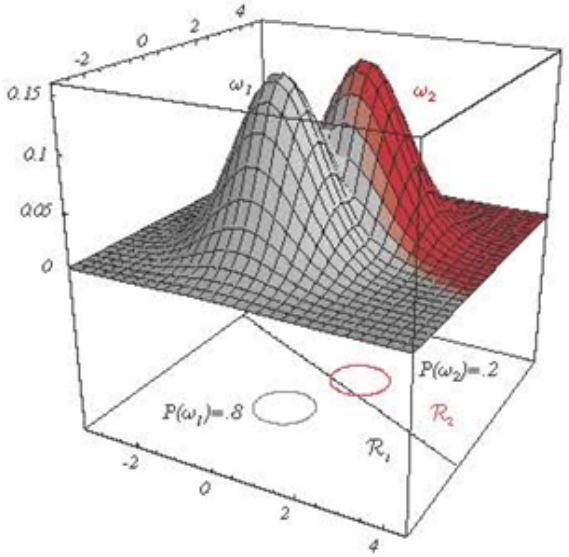
Case I: $\Sigma_i = \sigma^2 I, \forall i$

- Each class is modelled by the same diagonal matrix with equal diagonal elements
- This is true when features are uncorrelated (or statistically independent) with same variance
- How could the discriminant be simplified in this case?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then

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



Case I: $\Sigma_i = \sigma^2 I, \forall i$

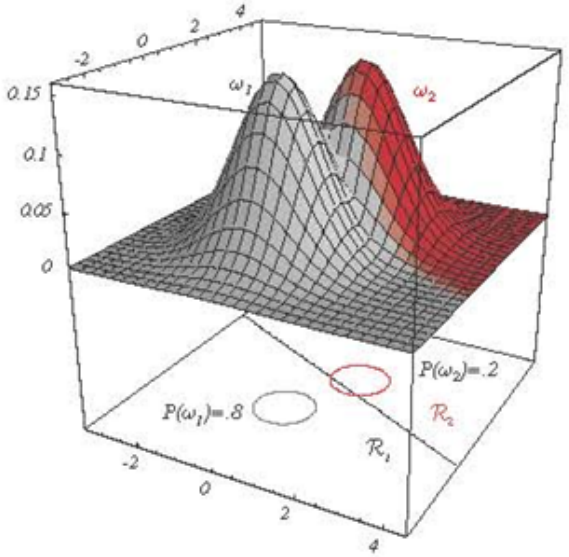
- Each class is modelled by the same diagonal matrix with equal diagonal elements
- This is true when features are uncorrelated (or statistically independent) with same variance
- How could the discriminant be simplified in this case?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then

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

- This is actually a linear discriminant, let's see why!



Case I: $\Sigma_i = \sigma^2 I, \forall i$ (discriminant function)

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (discriminant function)

- Expanding the above expression:

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

- Disregarding $\mathbf{x}^T\mathbf{x}$ (common across all $g_i(\mathbf{x})$), we get a linear discriminant: $g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (discriminant function)

- Expanding the above expression:

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

- Disregarding $\mathbf{x}^T\mathbf{x}$ (common across all $g_i(\mathbf{x})$), we get a linear discriminant: $g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$

- Where $\mathbf{w}_i = \frac{1}{\sigma^2}\mu_i$ and $w_{i0} = -\frac{1}{\sigma^2}\mu_i^T\mu_i + \ln P(\omega_i)$

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (discriminant function)

- Expanding the above expression:

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

- Disregarding $\mathbf{x}^T\mathbf{x}$ (common across all $g_i(\mathbf{x})$), we get a linear discriminant: $g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$

- Where $\mathbf{w}_i = \frac{1}{\sigma^2}\mu_i$ and $w_{i0} = -\frac{1}{\sigma^2}\mu_i^T\mu_i + \ln P(\omega_i)$

- What is the form of the **decision boundary** in this case?

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- Decision boundary is determined by hyperplanes; setting

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

$$\Rightarrow \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

- Where $\mathbf{w} = \mu_i - \mu_j$ and

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - (\mu_i - \mu_j)\frac{\sigma^2}{\| \mu_i - \mu_j \|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- Decision boundary is determined by hyperplanes; setting

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

$$\Rightarrow \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

- Where $\mathbf{w} = \mu_i - \mu_j$ and

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - (\mu_i - \mu_j)\frac{\sigma^2}{\| \mu_i - \mu_j \|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

- Properties of decision boundary:

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- Decision boundary is determined by hyperplanes; setting

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

$$\Rightarrow \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

- Where $\mathbf{w} = \mu_i - \mu_j$ and

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - (\mu_i - \mu_j)\frac{\sigma^2}{\| \mu_i - \mu_j \|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

- Properties of decision boundary:

1. It passes through $\mathbf{x}_0$

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- Decision boundary is determined by hyperplanes; setting

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

$$\Rightarrow \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

- Where $\mathbf{w} = \mu_i - \mu_j$ and

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - (\mu_i - \mu_j)\frac{\sigma^2}{\| \mu_i - \mu_j \|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

- Properties of decision boundary:

1. It passes through $\mathbf{x}_0$
2. It is orthogonal to the line linking the means.

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- Decision boundary is determined by hyperplanes; setting

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

$$\Rightarrow \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

- Where $\mathbf{w} = \mu_i - \mu_j$ and

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - (\mu_i - \mu_j)\frac{\sigma^2}{\| \mu_i - \mu_j \|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

- Properties of decision boundary:

1. It passes through $\mathbf{x}_0$
2. It is orthogonal to the line linking the means.
3. If σ is very small, x_0 is insensitive to $P(\omega_i)$ and $P(\omega_j)$

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- Decision boundary is determined by hyperplanes; setting

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

$$\Rightarrow \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

- Where $\mathbf{w} = \mu_i - \mu_j$ and

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - (\mu_i - \mu_j)\frac{\sigma^2}{\| \mu_i - \mu_j \|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

- Properties of decision boundary:

1. It passes through $\mathbf{x}_0$
2. It is orthogonal to the line linking the means.
3. If σ is very small, x_0 is insensitive to $P(\omega_i)$ and $P(\omega_j)$
4. If $P(\omega_i) = P(\omega_j)$, then $\mathbf{x}_0$ is the mid point of the two means.

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- Decision boundary is determined by hyperplanes; setting

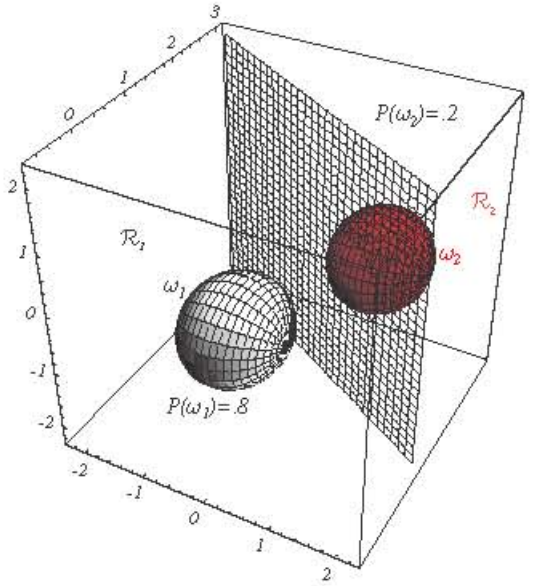
$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

$$\Rightarrow \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

- Where $\mathbf{w} = \mu_i - \mu_j$ and

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - (\mu_i - \mu_j)\frac{\sigma^2}{\| \mu_i - \mu_j \|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

- Properties of decision boundary:
  1. It passes through $\mathbf{x}_0$
  2. It is orthogonal to the line linking the means.
  3. If σ is very small, x_0 is insensitive to $P(\omega_i)$ and $P(\omega_j)$
  4. If $P(\omega_i) = P(\omega_j)$, then $\mathbf{x}_0$ is the mid point of the two means.



Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- When $P(\omega_i)$ are all *equal*, then the discriminant can be further simplified:

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

- This is known as the Euclidean distance classifier.

Case I: $\Sigma_i = \sigma^2 I, \forall i$ (decision surface)

- When $P(\omega_i)$ are all *equal*, then the discriminant can be further simplified:

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

- This is known as the Euclidean distance classifier.

- Very simple classifier, what does it do exactly?

Case II: $\Sigma_i = \Sigma$

Case II: $\Sigma_i = \Sigma$

- Each class is modelled by the same matrix, not necessarily diagonal
- Clusters are hyper-ellipsoidal with same size (centered at μ)

Case II: $\Sigma_i = \Sigma$

- Each class is modelled by the same matrix, not necessarily diagonal
- Clusters are hyper-ellipsoidal with same size (centered at μ)
- How could the discriminant be simplified?

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

Case II: $\Sigma_i = \Sigma$

- Each class is modelled by the same matrix, not necessarily diagonal
- Clusters are hyper-ellipsoidal with same size (centered at μ)
- How could the discriminant be simplified?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then

Case II: $\Sigma_i = \Sigma$

- Each class is modelled by the same matrix, not necessarily diagonal
- Clusters are hyper-ellipsoidal with same size (centered at μ)
- How could the discriminant be simplified?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then

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

- Expanding the above term and disregarding the quadratic term, we get $g_i(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_{i0}$ where

Case II: $\Sigma_i = \Sigma$

- Each class is modelled by the same matrix, not necessarily diagonal
- Clusters are hyper-ellipsoidal with same size (centered at μ)
- How could the discriminant be simplified?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then

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

- Expanding the above term and disregarding the quadratic term, we get

$g_i(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_{i0}$ where

$$\mathbf{w} = \Sigma^{-1}\mu_i \text{ and } w_{i0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(\omega_i)$$

Case II: $\Sigma_i = \Sigma$

- Each class is modelled by the same matrix, not necessarily diagonal
- Clusters are hyper-ellipsoidal with same size (centered at μ)
- How could the discriminant be simplified?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then

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

- Expanding the above term and disregarding the quadratic term, we get

$g_i(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_{i0}$ where

$$\mathbf{w} = \Sigma^{-1}\mu_i \text{ and } w_{i0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(\omega_i)$$

- This is also a linear discriminant

Case II: $\Sigma_i = \Sigma$

- Each class is modelled by the same matrix, not necessarily diagonal
- Clusters are hyper-ellipsoidal with same size (centered at μ)
- How could the discriminant be simplified?

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

- If we disregard $\frac{d}{2}\ln(2\pi)$ and $\frac{1}{2}\ln(|\Sigma_i|)$, then

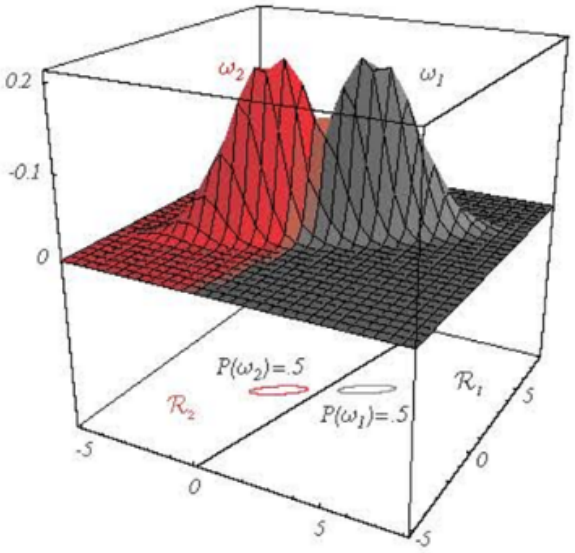
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln P(\omega_i)$$

- Expanding the above term and disregarding the quadratic term, we g

$g_i(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_{i0}$ where

$$\mathbf{w} = \Sigma^{-1}\mu_i \text{ and } w_{i0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(\omega_i)$$

- This is also a linear discriminant



Case II (cont'd)

Case II (cont'd)

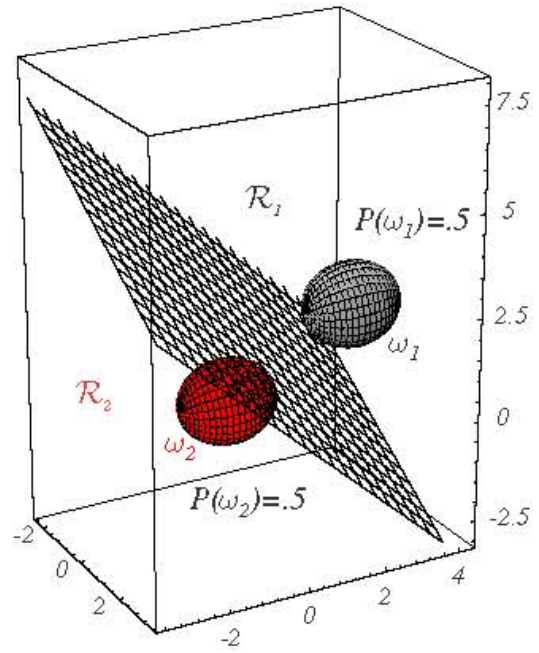
- *Decision boundary is determined by hyperplanes: setting* $g_i(\mathbf{x}) = g_j(\mathbf{x})$ *gives* $\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$
- *Where* $\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$ *and*

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)^T\Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

Case II (cont'd)

- *Decision boundary is determined by hyperplanes: setting* $g_i(\mathbf{x}) = g_j(\mathbf{x})$ *gives* $\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$
- *Where* $\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$ *and*

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)^T\Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$



Questions