



ARISTOTLE UNIVERSITY OF THESSALONIKI



FACULTY OF ENGINEERING

Pattern Recognition & Machine Learning

Bayes Decision Theory

Panagiotis C. Petrantonakis

Assistant Professor

Dept. of Electrical and Computer Engineering

ppetrant@ece.auth.gr

Fall Semester

Introduction

- The decision problem is posed in probabilistic terms!
- Assume ω_1 and ω_2 the states of nature: sea bass and salmon
- a priori probabilities: $P(\omega_1)$ and $P(\omega_2)$
- What is the decision rule that you would choose knowing only a priori probabilities?
- What is the probability of error?
- How can we improve our decision rule?



Introduction



- The decision problem is posed in probabilistic terms!
- Assume ω_1 and ω_2 the states of nature: sea bass and salmon
- a priori probabilities: $P(\omega_1)$ and $P(\omega_2)$
- What is the decision rule that you would choose knowing only a priori probabilities?
 - Choose ω_1 if $P(\omega_1) > P(\omega_2)$
- What is the probability of error?
- How can we improve our decision rule?

Introduction



- The decision problem is posed in probabilistic terms!
- Assume ω_1 and ω_2 the states of nature: sea bass and salmon
- a priori probabilities: $P(\omega_1)$ and $P(\omega_2)$
- What is the decision rule that you would choose knowing only a priori probabilities?
 - Choose ω_1 if $P(\omega_1) > P(\omega_2)$
- What is the probability of error?
 - the smaller of $P(\omega_1)$ and $P(\omega_2)$
- How can we improve our decision rule?

Introduction



- The decision problem is posed in probabilistic terms!
- Assume ω_1 and ω_2 the states of nature: sea bass and salmon
- a priori probabilities: $P(\omega_1)$ and $P(\omega_2)$
- What is the decision rule that you would choose knowing only a priori probabilities?
 - Choose ω_1 if $P(\omega_1) > P(\omega_2)$
- What is the probability of error?
 - the smaller of $P(\omega_1)$ and $P(\omega_2)$
- How can we improve our decision rule?
 - using feature vectors to detect pattern (Pattern Recognition)

Using a feature

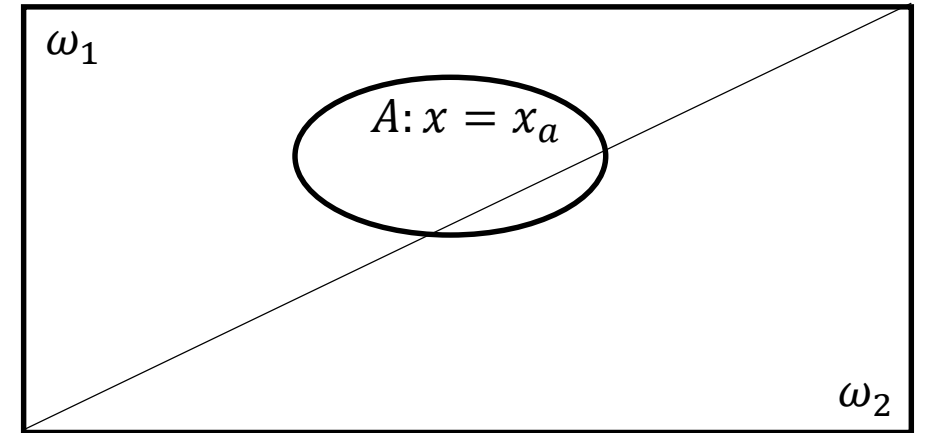


- Assume x to be a continuous random variable
 - Values of x are actually the feature values (e.g., fish length)
- x has a distribution that depends on the state of nature: $p(x|\omega)$
- Suppose we know both $P(\omega_1)$ and $P(\omega_2)$ as well as the **conditional density functions** $p(x|\omega_i)$, $i = 1, 2$
- If we get a measurement (feature value) $x = x_a$. How does this measurement influence our decision?

From the “Probabilities” course we know:

$$P(A) = P(A \cap \omega_1) + P(A \cap \omega_2)$$

Also:



$$P(A \cap \omega_i) = P(\omega_i \cap A) = P(A|\omega_i)P(\omega_i) = P(\omega_i|A)P(A) \text{ (product rule)}$$

From above I get:

$$P(A) = P(A|\omega_1)P(\omega_1) + P(A|\omega_2)P(\omega_2) \text{ and } P(\omega_i|A) = \frac{P(A|\omega_i)P(\omega_i)}{P(A)}$$

Bayes Theorem

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

Bayes Theorem

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

or

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

Bayes Theorem

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

or

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

or

$$P(\omega_i|x) = \frac{\textit{likelihood}}{\textit{evidence}} P(\omega_i)$$

Bayes Theorem

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

or

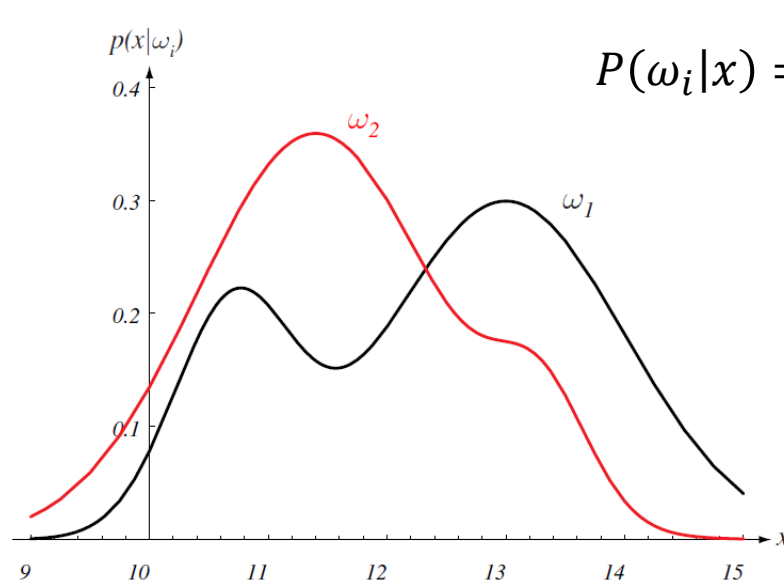
$$P(\omega_i|x) = \frac{p(x|\omega_i)}{p(x)} P(\omega_i)$$

or

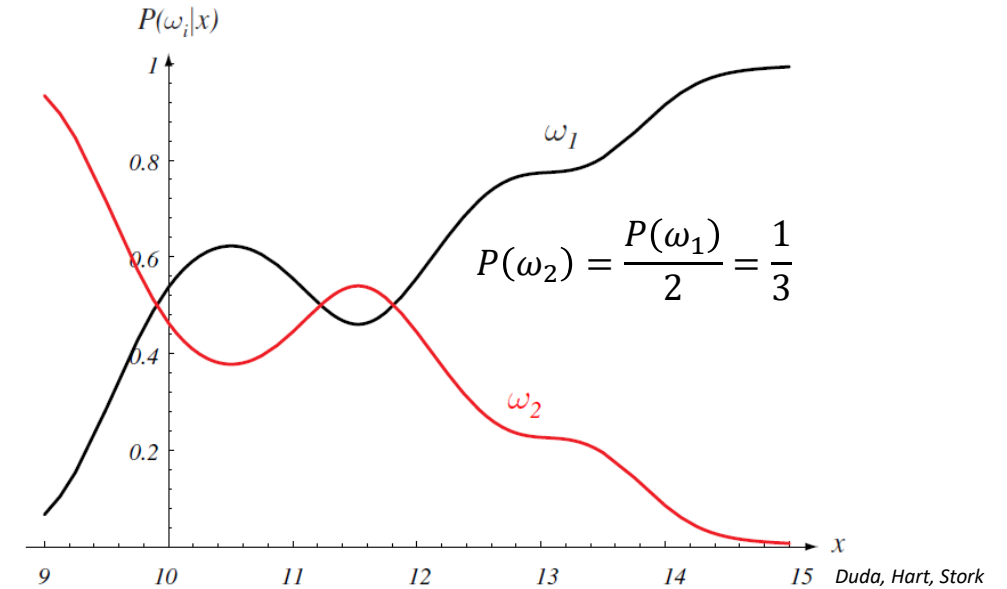
$$P(\omega_i|x) = \frac{\textit{likelihood}}{\textit{evidence}} P(\omega_i)$$

Bayes transforms prior to posterior once a feature has been observed!

Class Conditional pdf vs. posterior probability



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



- Class Conditional probability density functions

- Show the probability density of measuring a particular feature value x given that the observed pattern is in state of nature ω_i

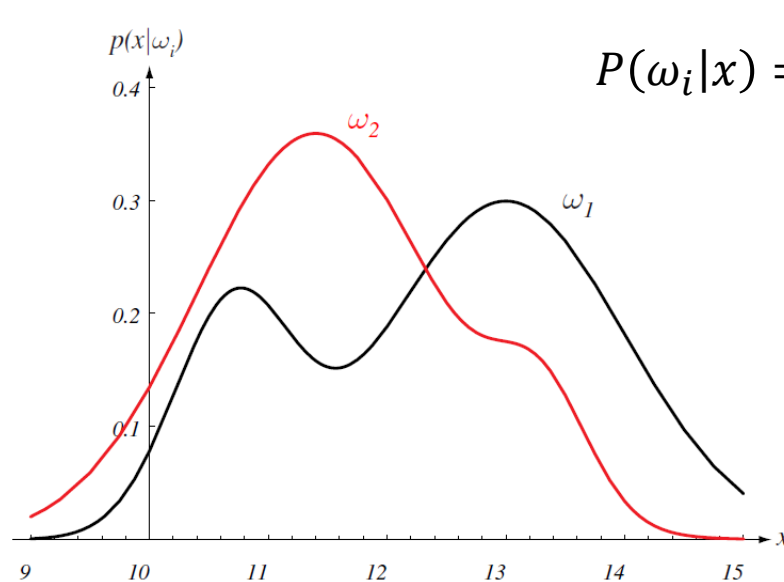
- Posterior probabilities

- at every x the posteriors **sum to 1**.

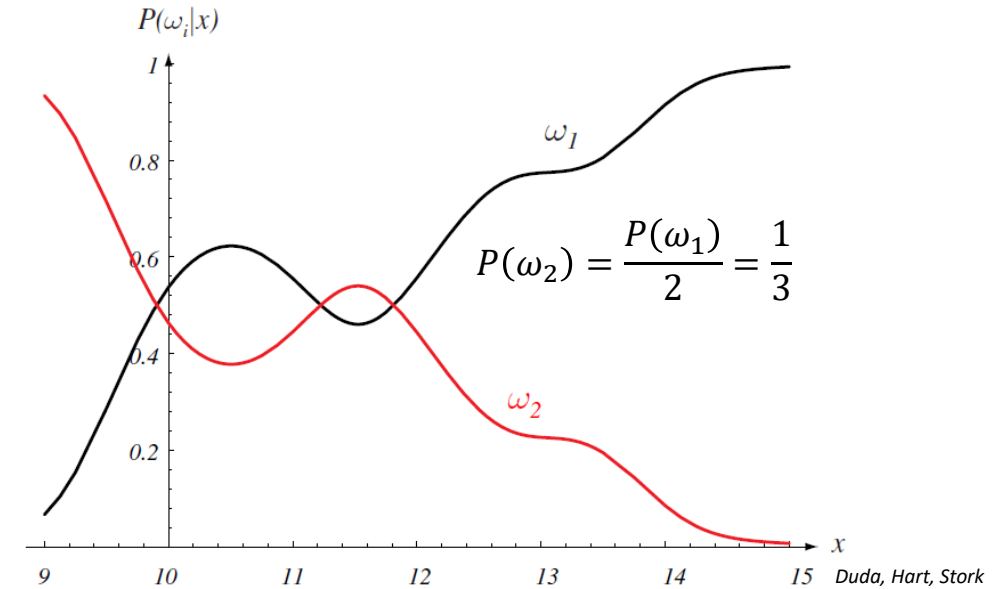
- What is the role of $p(x)$?

- scale factor: guarantees that the posterior probabilities sum to one!
- it “says” how frequently we measure a certain value of x (feature value)

Class Conditional pdf vs. posterior probability



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



- Class Conditional probability density functions

- Show the probability density of measuring a particular feature value x given that the observed pattern is in state of nature ω_i

- $p(x|\omega_1) + p(x|\omega_2) = 1$
 - T or F?

- Posterior probabilities

- at every x the posteriors sum to 1.

- What is the role of $p(x)$?

- scale factor: guarantees that the posterior probabilities sum to one!
- it “says” how frequently we measure a certain value of x (feature value)

Probability of error in BDT

- Whenever we observe a value of x the probability of error is

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

- So, how can we minimize the probability of error?

Probability of error in BDT

- Whenever we observe a value of x the probability of error is

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

- So, how can we minimize the probability of error?
 - Deciding ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ and ω_2 otherwise

- The above rule minimizes the average probability of error:

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x) dx = \int_{-\infty}^{\infty} P(\text{error}|x) p(x) dx$$

Probability of error in BDT

- Whenever we observe a value of x the probability of error is

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

- So, how can we minimize the probability of error?
 - Deciding ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ and ω_2 otherwise

- The above rule minimizes the average probability of error:

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x) dx = \int_{-\infty}^{\infty} P(\text{error}|x) p(x) dx$$

- This is the **minimum error** of any deterministic classifier!!
 - Assuming that we know the real distributions that generate the data!

Bayes Decision Rule

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

- BDR : ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise ω_2
- Special case: $P(\omega_1) = P(\omega_2)$
- special case: $p(x|\omega_1) = p(x|\omega_2)$

Bayes Decision Rule

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

- BDR : ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise ω_2
- Special case: $P(\omega_1) = P(\omega_2)$
 - the states of nature are equally probable; the decision is based entirely on the likelihoods
- special case: $p(x|\omega_1) = p(x|\omega_2)$

Bayes Decision Rule

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

- BDR : ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise ω_2
- Special case: $P(\omega_1) = P(\omega_2)$
 - the states of nature are equally probable; the decision is based entirely on the likelihoods
- special case: $p(x|\omega_1) = p(x|\omega_2)$
 - that particular observation gives us no information about the state of nature.

Bayes Decision Rule

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

- BDR : ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise ω_2
- Special case: $P(\omega_1) = P(\omega_2)$
 - the states of nature are equally probable; the decision is based entirely on the likelihoods
- special case: $p(x|\omega_1) = p(x|\omega_2)$
 - that particular observation gives us no information about the state of nature.
- In the general case, both factors are important in making decision. Bayes rule combines them to achieve the minimum probability of error.

Generalize the BDT

- Use more than one features
 - replace scalar value x with a feature vector $\mathbf{x} \in \mathbb{R}^d$ (d -dimensional feature space)
- More than two classes
 - $\{\omega_1, \dots, \omega_c\}$
- Allowing actions instead of simple classification; actions: $\{\alpha_1, \dots, \alpha_k\}$
 - this allows the adoption of more actions such as rejection!
 - I may reject the option of classification if I am not sure and it is not too costly...
- Loss function, $\lambda(a_i | \omega_j)$
 - states exactly how costly an action is.
 - Some kind of classification mistakes is more costly than others.

Conditional Risk

- Bayes formula: $P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x})} P(\omega_i), i = 1, \dots, c$
 - where $p(\mathbf{x}) = \sum_{j=1}^c p(\mathbf{x}|\omega_j) P(\omega_j)$
- Suppose that we observe a particular \mathbf{x} and we contemplate taking action a_i . The expected loss (conditional risk) of taking action a_i is:

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

- Whenever we encounter a particular observation we can minimize our conditional risk by selecting the appropriate action.
 - This will lead again to optimal performance (as in the previous simple example)

General Decision Rule in BDT

- Assume decision function $\gamma(\cdot)$ that tells us which action to take for every possible observation, e.g., $\gamma(\mathbf{x}) = \alpha_3$.

- Then the overall risk is:

$$\mathcal{R} = \int R(\gamma(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

- Thus if $\gamma(\mathbf{x})$ is delivering the action that minimizes the conditional error at every observation then the overall risk is minimized.
- **General Bayes Rule:** find $R(\alpha_i|\mathbf{x})$ for $i = 1, \dots, k$ and then select the action for which $R(\alpha_i|\mathbf{x})$ is minimum. The resulting risk is called, *Bayes Risk* and it is the best performance that can be achieved.

Two class classification (revisited)

- Let us consider the binary classification case based on the generalized formulation of the BDT.
- Actions:
 - α_1 : deciding that the true state of nature is ω_1
 - α_2 : deciding that the true state of nature is ω_2
- Loss:
 - $\lambda(a_i | \omega_j)$: the loss incurred for deciding ω_i when the true state of nature is ω_j
 - for simplicity we will use the notation λ_{ij} .
- Conditional risks:
 - $R(\alpha_1 | \mathbf{x}) = \lambda_{11}P(\omega_1 | \mathbf{x}) + \lambda_{12}P(\omega_2 | \mathbf{x})$
 - $R(\alpha_2 | \mathbf{x}) = \lambda_{21}P(\omega_1 | \mathbf{x}) + \lambda_{22}P(\omega_2 | \mathbf{x})$

Two class classification (revisited)

- GBR: ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and ω_2 otherwise. Thus:

Two class classification (revisited)

- GBR: ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and ω_2 otherwise. Thus:

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

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

Two class classification (revisited)

- GBR: ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and ω_2 otherwise. Thus:

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

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

by substituting $P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x})} P(\omega_i)$ we get:

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

Two class classification (revisited)

- GBR: ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and ω_2 otherwise. Thus:

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

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

by substituting $P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x})} P(\omega_i)$ we get:

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

- Thus, the GBR decides ω_1 if the **likelihood ratio** exceeds a threshold value that is **independent** of the observation \mathbf{x} !

Two class classification (revisited)

- GBR: ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and ω_2 otherwise. Thus:

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

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

- What if I determine: $\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$?

Two class classification (revisited)

- GBR: ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and ω_2 otherwise. Thus:

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

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

- What if I determine: $\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$?

- BDR : ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise ω_2

Discriminant Functions

- Previously:
 - General Bayes Rule: find $R(\alpha_i|\mathbf{x})$ for $i = 1, \dots, k$ and then select the action for which $R(\alpha_i|\mathbf{x})$ is **minimum**.
- Thus, feature \mathbf{x} is assigned to class ω_i (in case actions are about deciding classes) if:

$$R(\alpha_i|\mathbf{x}) < R(\alpha_j|\mathbf{x}) \quad \forall j \neq i$$

- If we define functions $g_i(\cdot)$ such that I decide class ω_i if:

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

then what is the $g(\cdot)$ function in generalized BDT?

Discriminant Functions

- Previously:
 - General Bayes Rule: find $R(\alpha_i|\mathbf{x})$ for $i = 1, \dots, k$ and then select the action for which $R(\alpha_i|\mathbf{x})$ is **minimum**.
- Thus, feature \mathbf{x} is assigned to class ω_i (in case actions are about deciding classes) if:

$$R(\alpha_i|\mathbf{x}) < R(\alpha_j|\mathbf{x}) \quad \forall j \neq i$$

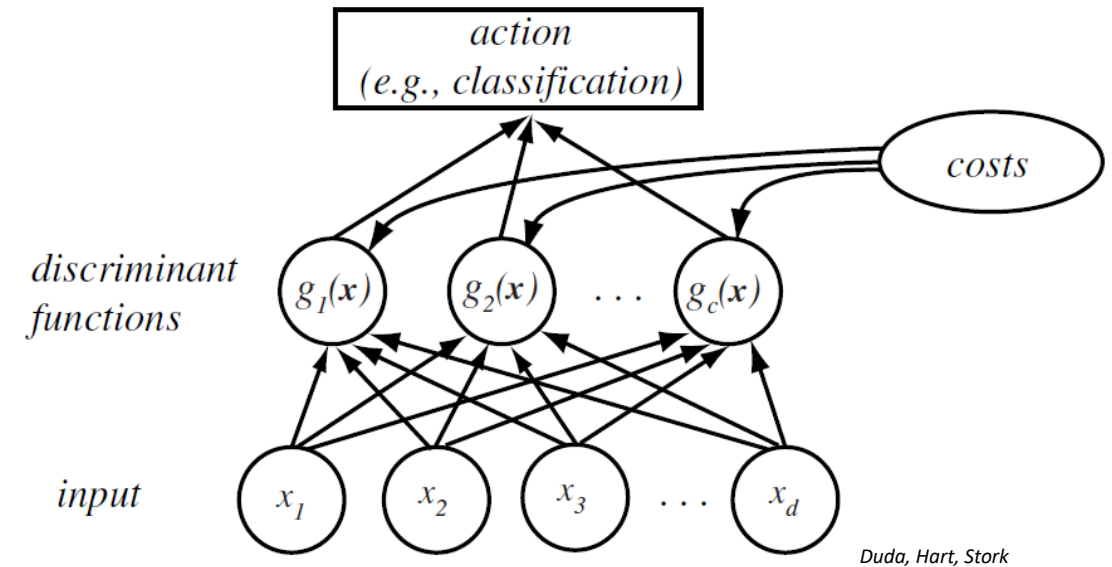
- If we define functions $g_i(\cdot)$ such that I decide class ω_i if:

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

then what is the $g(\cdot)$ function in generalized BDT? $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$

Discriminant Functions

- In general:
 - define functions $g_i(\cdot)$ such that:
$$\text{decide } \omega_1 \text{ if } g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall i \neq j$$
- Thus the classifier is viewed as a machine that computes c discriminant functions and selects the class that corresponds to the largest discriminant value.

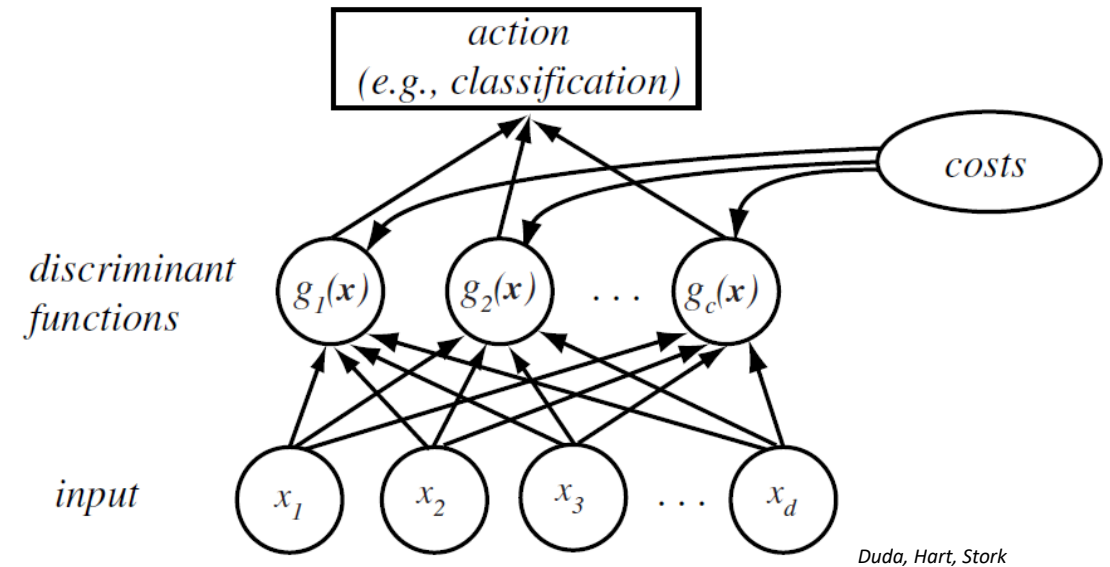


Discriminant Functions

- In general:
 - define functions $g_i(\cdot)$ such that:

decide ω_1 if $g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall i \neq j$

- Thus the classifier is viewed as a machine that computes c discriminant functions and selects the class that corresponds to the largest discriminant value.



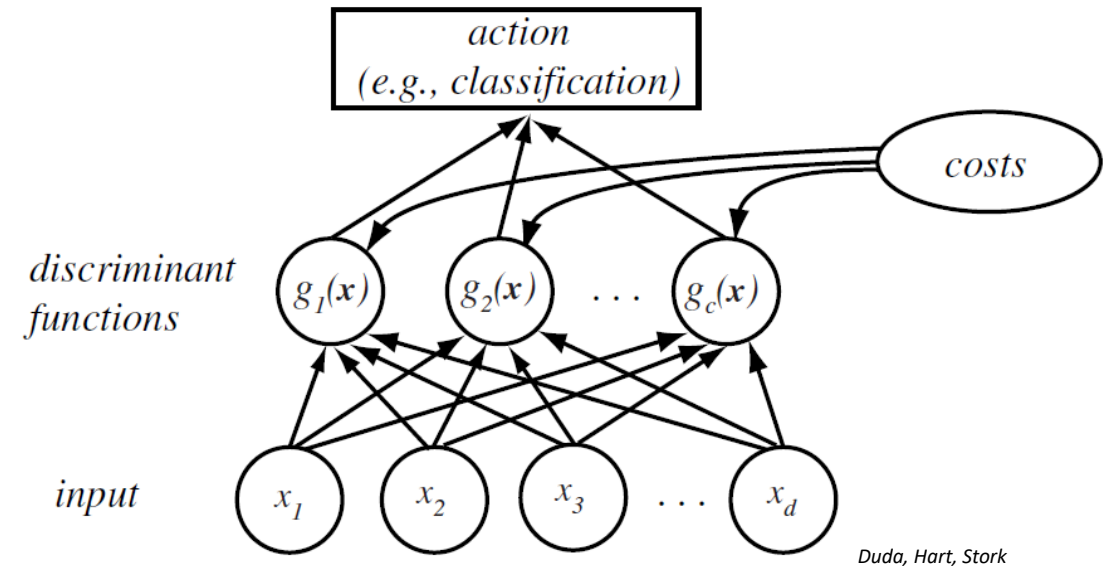
- The choice of discriminant functions is not unique! Why?

Discriminant Functions

- In general:
 - define functions $g_i(\cdot)$ such that:

decide ω_1 if $g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall i \neq j$

- Thus the classifier is viewed as a machine that computes c discriminant functions and selects the class that corresponds to the largest discriminant value.



- The choice of discriminant functions is not unique! Why?
 - if we replace every $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$ where f is a monotonically increasing function we will get identical results!
- This can lead to analytical and computational simplifications.

Discriminant Functions

- In the simplest case in BDT where $\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$ we can define the discriminant functions as:

$$g_i(\mathbf{x}) = P(\omega_i | \mathbf{x})$$

or

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

or

Discriminant Functions

- In the simplest case in BDT where $\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$ we can define the discriminant functions as:

$$g_i(\mathbf{x}) = P(\omega_i | \mathbf{x})$$

or

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

or

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

Discriminant Functions

- In the simplest case in BDT where $\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$ we can define the discriminant functions as:

$$g_i(\mathbf{x}) = P(\omega_i | \mathbf{x})$$

or

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

or

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

- The effect of the above formulation is that the feature space is divided in c distinct regions, $\mathfrak{R}_1, \dots, \mathfrak{R}_c$. If $g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall i \neq j$ then \mathbf{x} is in region \mathfrak{R}_i . The regions are separated by decision boundaries.

Discriminant Functions

- In the simplest case in BDT where $\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$ and we have only two classes instead of defining a discriminant function for each class we can instead define one discriminant function as:

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

or

$$g(\mathbf{x}) = P(\omega_1 | \mathbf{x}) - P(\omega_2 | \mathbf{x})$$

or

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

- Then the decision rule would be: decide ω_1 if $g(\mathbf{x}) > 0$ and ω_2 otherwise.

We need what we do not know!

- Thus, we have concluded to the discriminant function:

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

- We need to determine $p(\mathbf{x}|\omega_i)$ and $P(\omega_i)$
 - Maybe for the $P(\omega_i)$ case things are more easy as we can infer it from the training data.
- Of the various density functions that have been investigated for BDT the normal density function has received the most attention:
 - analytical tractability
 - models randomly corrupted versions of a single prototype vector
 - central limit theorem

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\}$$

Discriminant functions using normal density

- We have the discriminant functions:

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

and by substituting the normal density for $p(\mathbf{x}|\omega_i)$, i.e., $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ we get:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

- We will now examine this family of discriminant functions and resulting classification schemes for a number of special cases for the covariance matrix $\boldsymbol{\Sigma}$.

Case 1: $\Sigma_i = \sigma^2 I$

- The simplest case is when the features are statistically independent and each feature has the same variance σ^2
- The covariance matrix is diagonal, thus $\Sigma_i = \sigma^2 I$.
- **Geometrically speaking:** each sample fall in equal-size hyper-spherical clusters each one centered about the mean μ_i (depending on the class).

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

- Substituting $|\Sigma_i| = \sigma^{2d}$, $\Sigma_i^{-1} = \frac{1}{\sigma^2} I$, I get:

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

Case 1: $\Sigma_i = \sigma^2 I$ (cont'd)

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

- if the Euclidean distance of a sample from prototypes $\boldsymbol{\mu}_i$ is equal, e.g., from two prototypes, then the optimal decision will favor the a priori more probable class.
- Nevertheless, it is not necessary to compute distances from all prototypes as I can expand the above $g_i(\mathbf{x})$:

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x} - \boldsymbol{\mu}_i)^t (\mathbf{x} - \boldsymbol{\mu}_i)}{2\sigma^2} + \ln P(\omega_i) \Leftrightarrow$$

$$g_i(\mathbf{x}) = -\frac{\mathbf{x}^t \mathbf{x} - 2\boldsymbol{\mu}_i^t \mathbf{x} + \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i}{2\sigma^2} + \ln P(\omega_i) \Leftrightarrow$$

Case 1: $\Sigma_i = \sigma^2 I$ (cont'd)

$$g_i(\mathbf{x}) = -\frac{\mathbf{x}^t \mathbf{x} - 2\boldsymbol{\mu}_i^t \mathbf{x} + \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i}{2\sigma^2} + \ln P(\omega_i) \Leftrightarrow$$

- Even if I have a quadratic function on \mathbf{x} , the quadratic term $\mathbf{x}^T \mathbf{x}$ is the same for all i 's.
- Thus by omitting the quadratic term we can conclude with:

$$g_i(\mathbf{x}) = \frac{\boldsymbol{\mu}_i^t}{\sigma^2} \mathbf{x} - \frac{1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i) \Leftrightarrow$$

which is a **linear discriminant function** of the form:

$$g_i(\mathbf{x}) = \mathbf{w}_i \mathbf{x} + b_i$$

Case 1: $\Sigma_i = \sigma^2 I$ - Hyperplanes

$$g_i(\mathbf{x}) = \mathbf{w}_i \mathbf{x} + b_i$$

- The decision boundaries of a linear discriminant functions classifier is **linear hyperplanes** defined by the two discriminant functions of the classes with the highest posterior probabilities as:

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \rightarrow g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0 \rightarrow \mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \text{ (decision boundary)}$$

where $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and $\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

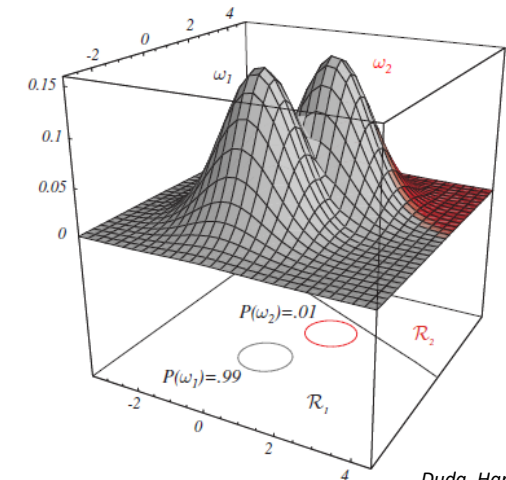
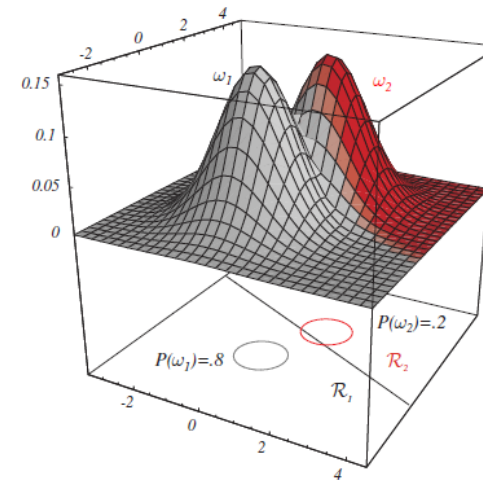
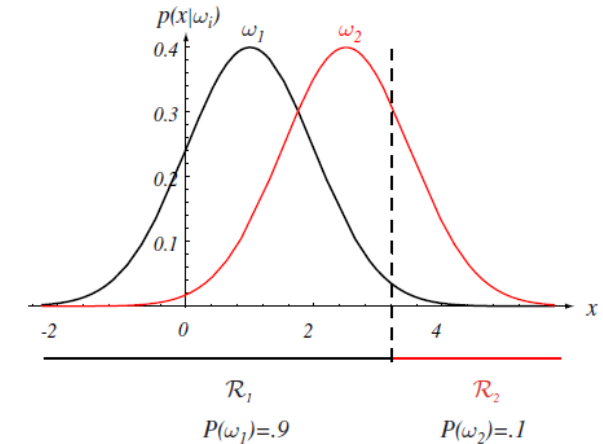
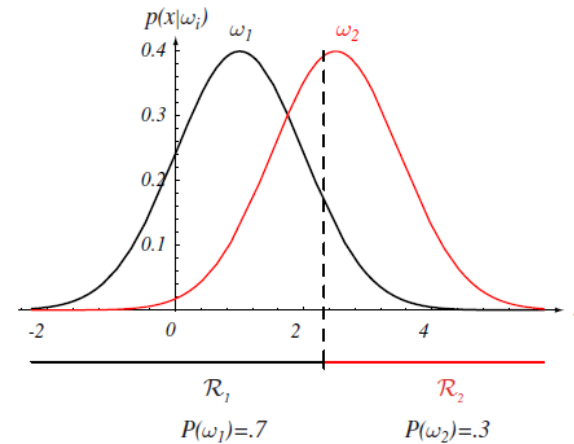
- The above equations define a hyperplane which passes through the point \mathbf{x}_0 and is orthogonal to the vector \mathbf{w} .
- $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$: the plane is orthogonal to the line linking the means
- If $P(\omega_i) = P(\omega_j)$ the hyperplane is the perpendicular bisector of that line.

Case 1: $\Sigma_i = \sigma^2 I$ - Hyperplanes

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$



Duda, Hart, Stork

Case 2: $\Sigma_i = \Sigma$

- Another simple case arises when covariance matrices are not diagonal, they are of arbitrary form but the same for all classes.
- Geometrically speaking:

Case 2: $\Sigma_i = \Sigma$

- Another simple case arises when covariance matrices are not diagonal, they are of arbitrary form but the same for all classes.
- **Geometrically speaking:** this corresponds to the situation where the samples fall in a hyper-ellipsoidal clusters of equal size and same shape centered about the mean vector μ_i .
- Now the discriminant function

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

becomes:

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

$$\text{Before we had: } \Sigma_i = \sigma^2 \mathbf{I}: g_i(\mathbf{x}) = -\frac{(\mathbf{x} - \mu_i)^t (\mathbf{x} - \mu_i)}{2\sigma^2} + \ln P(\omega_i)$$

Case 2: $\Sigma_i = \Sigma$ (cont'd)

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i)$$

- If the **Mahalanobis** distance of a sample from prototypes $\boldsymbol{\mu}_i$ is equal, e.g., from two prototypes, then the optimal decision will favor the a priori more probable class.
- On the other hand, if the a priori probabilities are the same the sample will be assigned to the class of the closest (in terms of the Mahalanobis distance) prototype.
- Again, expansion of the above form involves the quadratic term $\mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x}$ which can be omitted as it is the same across all classes.
- If we drop the quadratic form of the above discriminant functions we will end up (again) with linear discriminant functions...

Case 2: $\Sigma_i = \Sigma$ (cont'd)

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + b_i$$

where $\mathbf{w}_i = \Sigma^{-1} \boldsymbol{\mu}_i$ and $b_i = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i)$

- The decision boundaries are:

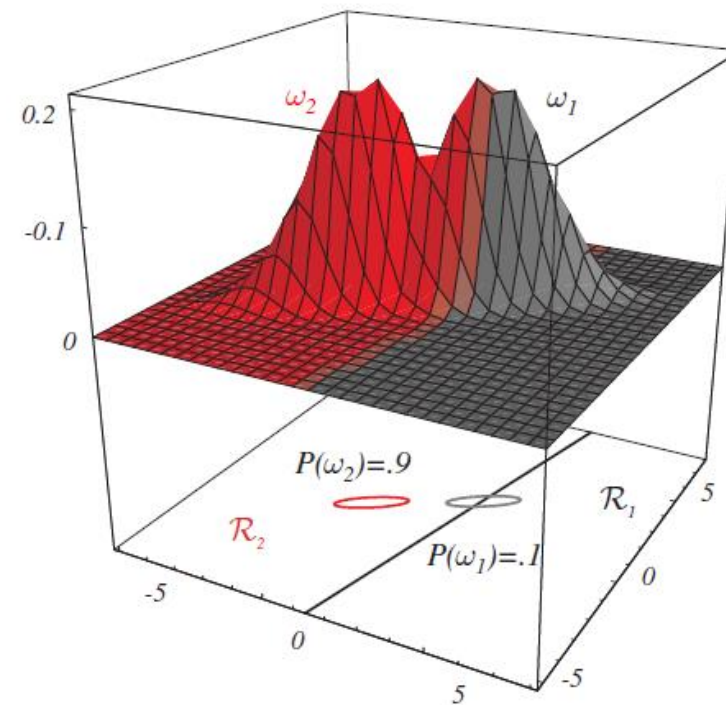
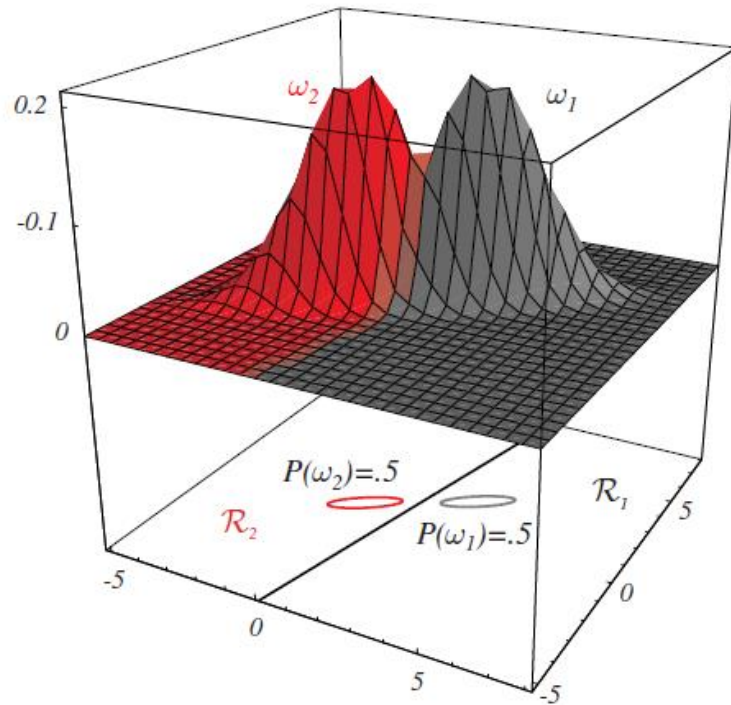
$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \rightarrow g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0 \rightarrow \mathbf{w}^t (\mathbf{x} - \mathbf{x}_0) = 0 \text{ (decision boundary)}$$

where $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ and $\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_{\Sigma}^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

- As $\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ is not generally in the direction of $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ the hyperplane is not generally orthogonal to the line connecting the means. However it intersects that line at point \mathbf{x}_0 .
- If $P(\omega_i) = P(\omega_j)$, \mathbf{x}_0 is halfway between the means. Otherwise, the hyperplane is shifted.

Case 2: $\Sigma_i = \Sigma$ - Hyperplanes

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0, \quad \text{where } \mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \text{ and } \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_{\Sigma}^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$



Duda, Hart, Stork

Case 3: Σ_i

- We will see now the general multivariate normal case where the covariance matrices are different for each class.
- Thus, from the general discriminant function:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

the only term that we can drop is the $-\frac{d}{2} \ln 2\pi$. So, we get:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + -\frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

and after certain expansion:

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i \mathbf{x} + b_i$$

where $\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1}$, $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i$ and $b_i = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$

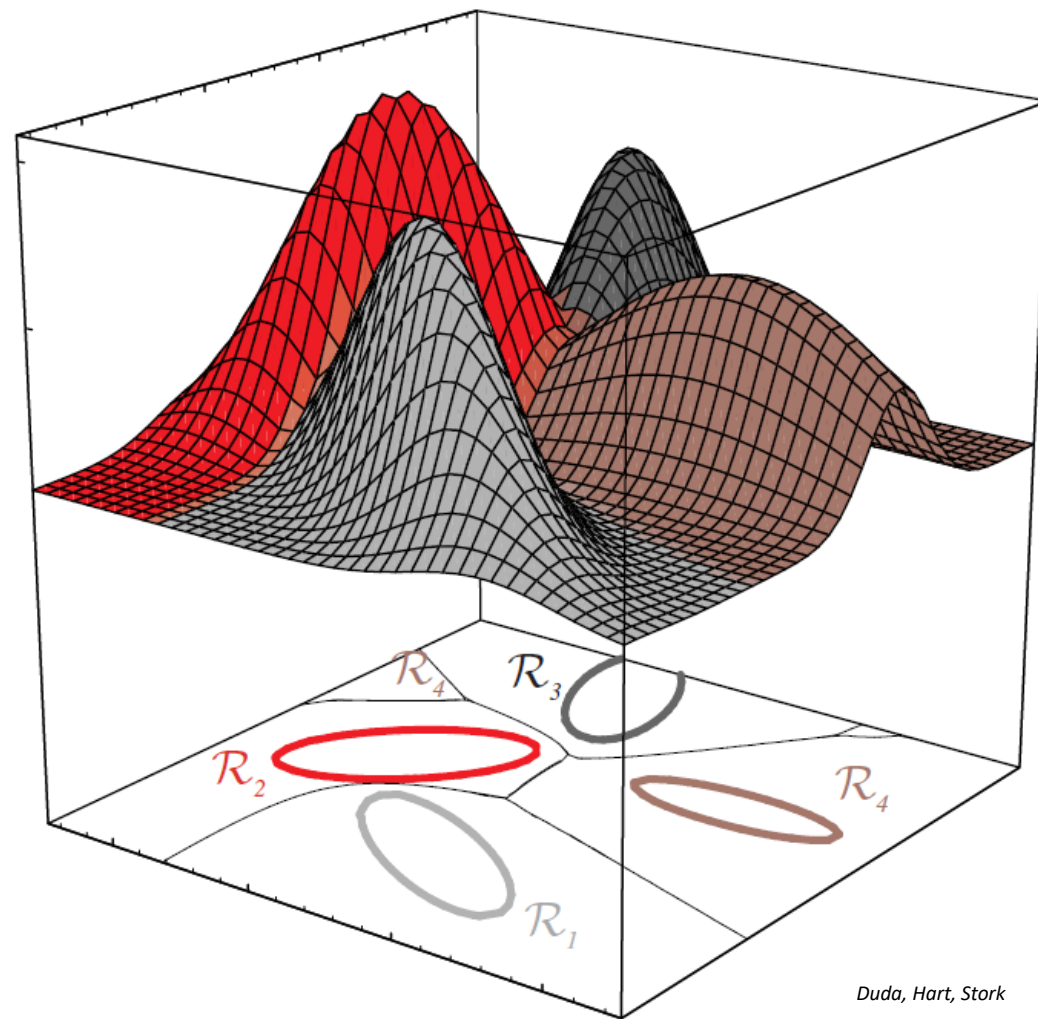
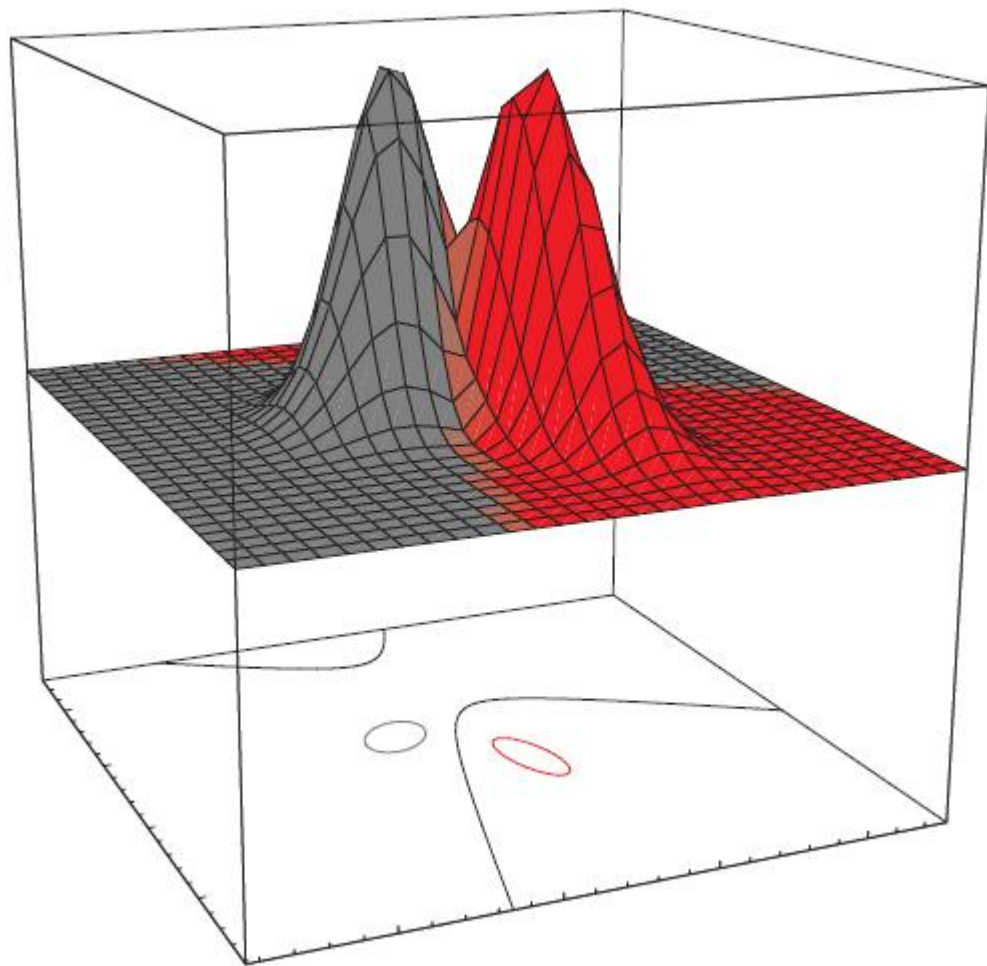
Case 3: Σ_i (cont'd)

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i \mathbf{x} + b_i$$

where $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$, $\mathbf{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i$ and $b_i = -\frac{1}{2}\boldsymbol{\mu}_i^t \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$

- The decision surfaces are hyperquadrics and get any of the general forms, e.g. , hyperplanes, hyperspheres, hyperellipsoids etc.
- The complexity of the surfaces depend on the distributions that are used for the $p(\mathbf{x}|\omega_i)$.
- More complex distributions, more complex decision surfaces. In any case, the same underlying theory holds there too.

Case 3: Σ_i - Surfaces



Duda, Hart, Stork

Sum up

- Bayes Theorem and minimum error rate
- Generalized Bayes decision Rule and minimum conditional risk
 - minimum over all risk
- Losses of actions (instead of mere classification)
- Discriminant functions
- Normal density functions
- Different cases, depending on the form of the covariance matrices
 - different decision boundaries



ARISTOTLE UNIVERSITY OF THESSALONIKI



FACULTY OF ENGINEERING

Questions?

Pattern Recognition & Machine Learning

Bayes Decision Theory