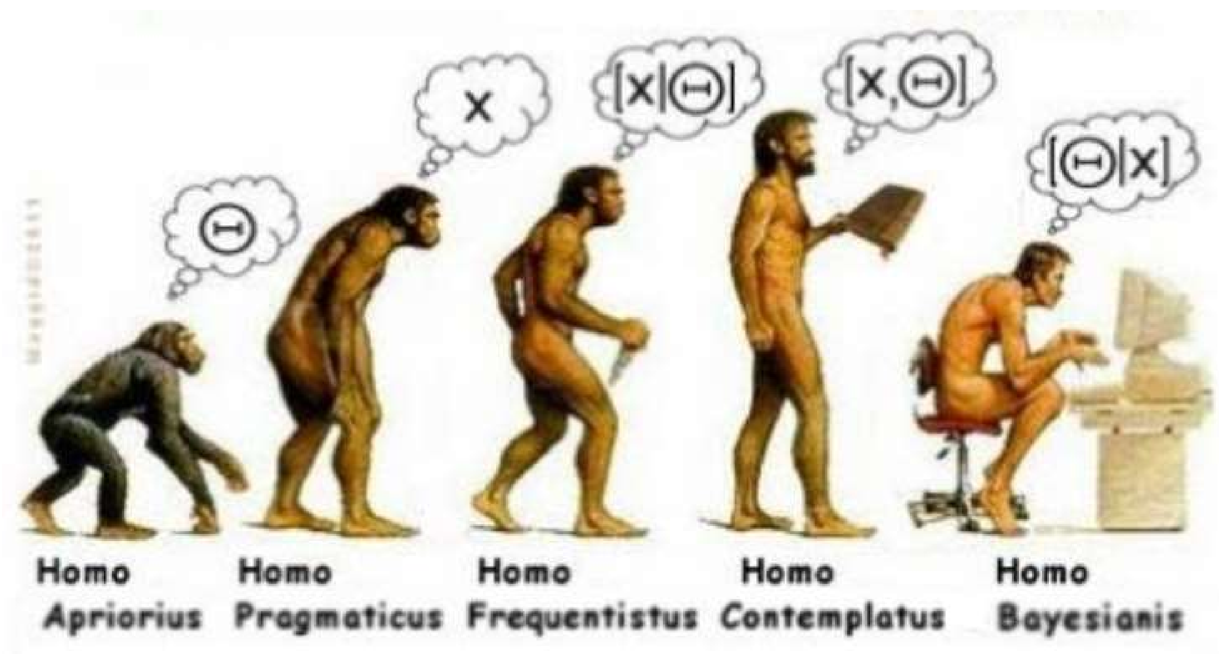
Chapter 2

Decision theory



Recommended bibliography: *Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*, Chapters 2 & 3

Credits: Some figures are taken from *Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000* with the permission of the authors

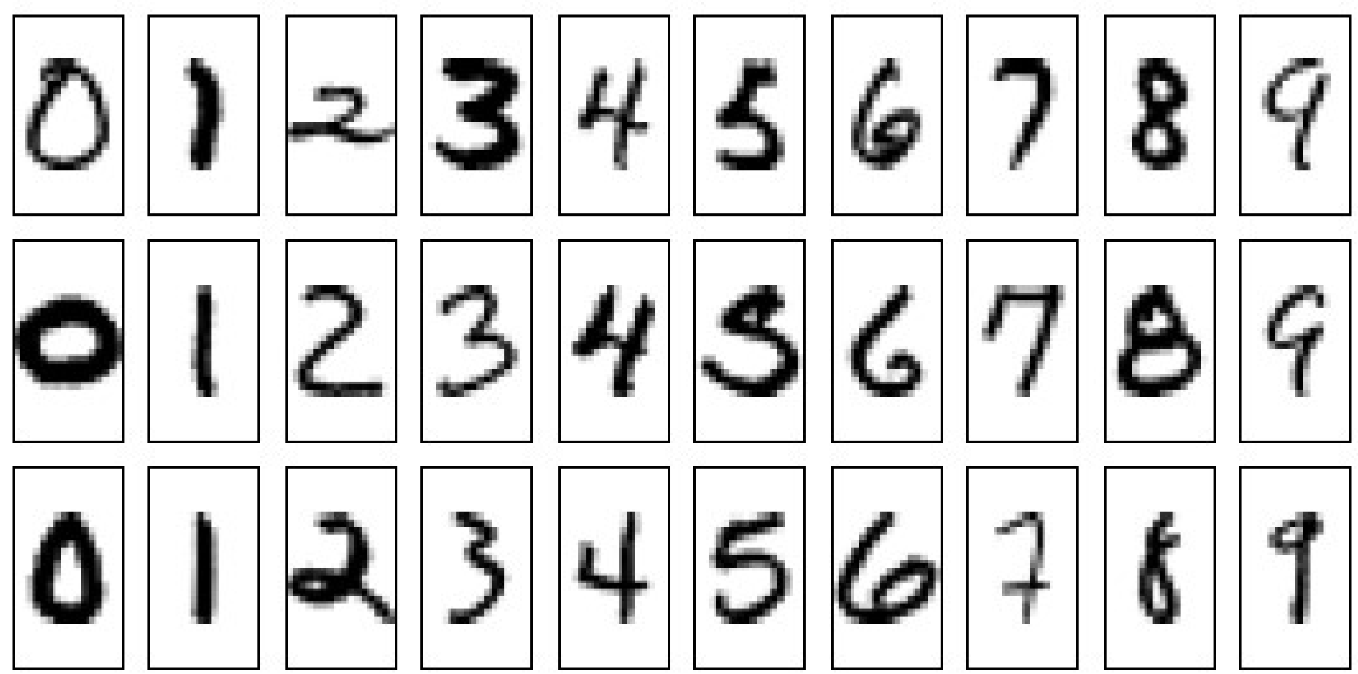
CONTENTS

1.1 INTRODUCTION

Why a probabilistic approach to decision taking?

1. We might have incomplete representations of reality (e.g., we do not have the DNA of the catched fishes)

2. We face problems intrinsically random (e.g., the identification of handwritten characters)



- Measurements: random vectors of size d

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$$

- State of nature: a random variable taking c posible values

$$\omega_1, \ldots, \omega_c$$

- Prior probabilities

$$\Pr(\omega_1), \ldots, \Pr(\omega_c) \qquad \sum_{i=1}^{c} \Pr(\omega_i) = 1$$

- Conditioned density functions

$$f_{\mathbf{x}|\omega_1}(\mathbf{x}|\omega_1), \ldots, f_{\mathbf{x}|\omega_c}(\mathbf{x}|\omega_c)$$

- Posterior probabilities

$$\Pr(\omega_j|\mathbf{x}) = \frac{f_{\mathbf{x}|\omega_j}(\mathbf{x}|\omega_j)\Pr(\omega_j)}{f_{\mathbf{x}}(\mathbf{x})} \qquad f_{\mathbf{x}}(\mathbf{x}) = \sum_{i=1}^{c} f_{\mathbf{x}|\omega_j}(\mathbf{x}|\omega_j)\Pr(\omega_j)$$

$$POSTERIOR = \frac{LIKELIHOOD \times PRIOR}{EVIDENCE}$$

- POSTERIOR: Probability of a certain nature state given the observed feature vector.
- LIKELIHOOD: Contains the characterization of data for a given class.
- PRIOR: Prior knowledge of the class.
- EVIDENCE: Scaling factor independent of the class.

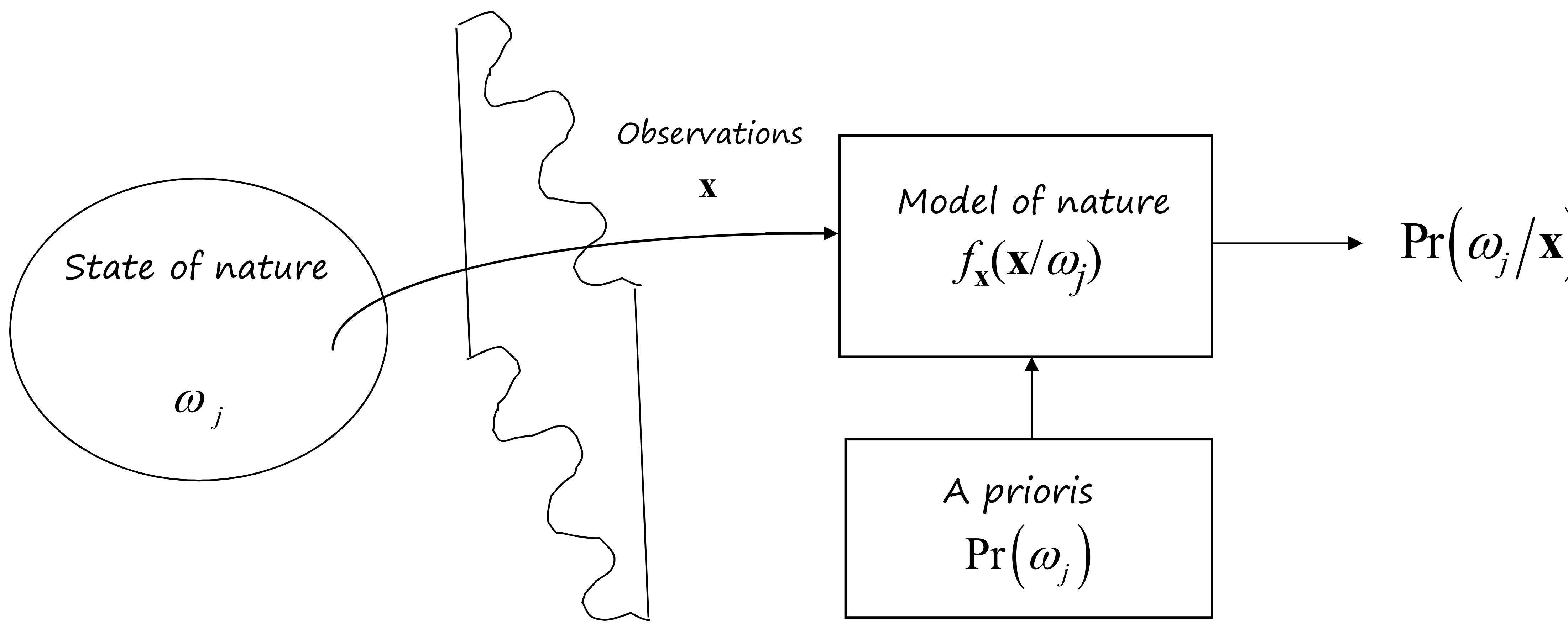
Two interpretations of probability appearing in our formulation...

- Frequentist

- Reliability (o belief)

What are we computing?

$$\Pr\left(\omega_j \middle| \mathbf{x}\right) = \frac{f_{\mathbf{x}}\left(\mathbf{x} \middle| \omega_j\right) \Pr\left(\omega_j\right)}{f_{\mathbf{x}}\left(\mathbf{x}\right)}$$

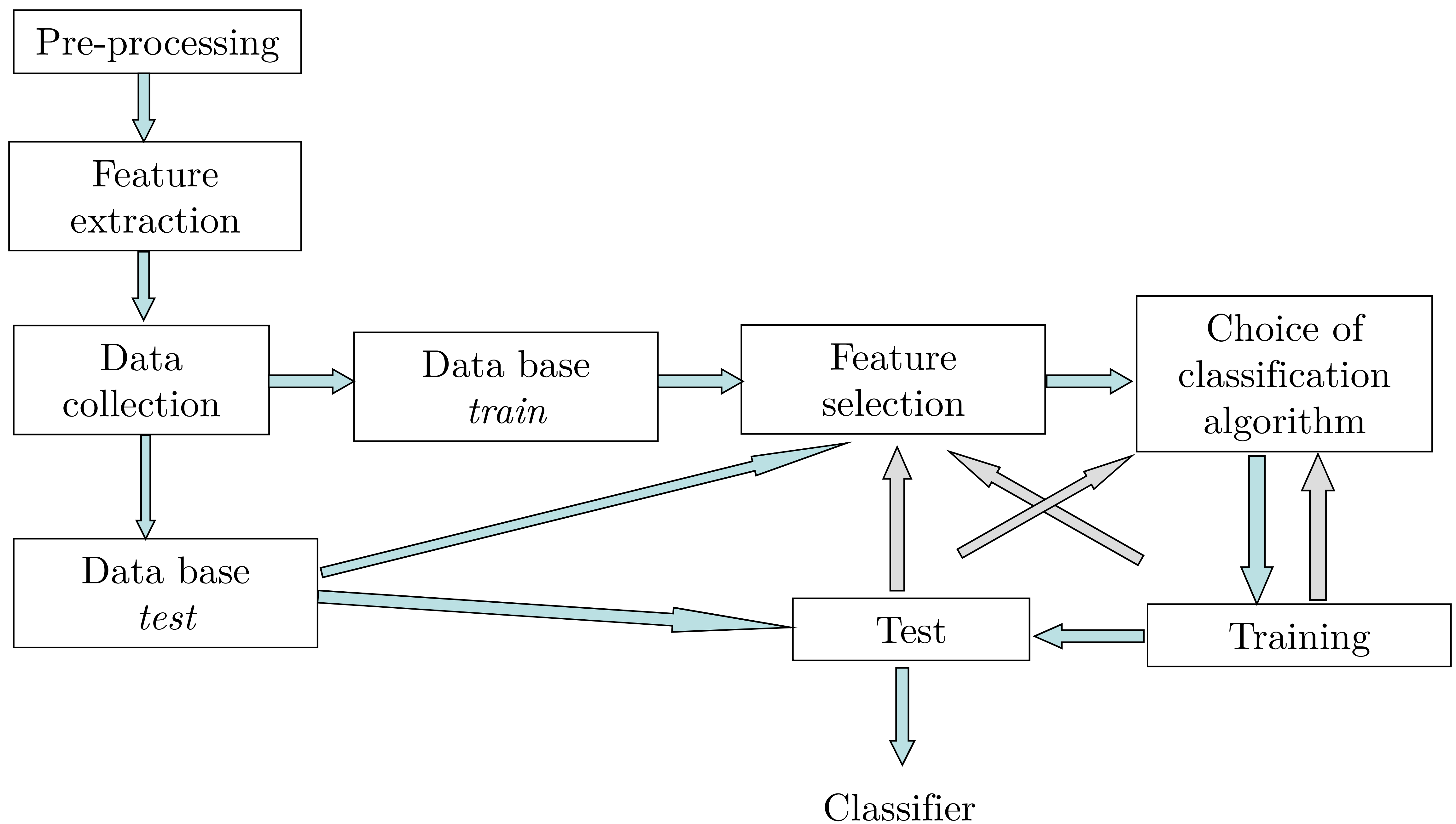


Thomas Bayes (1702-1761)
In 1763, two years after his death, *Essay Towards Solving a Problem in the Doctrine of Chances* is published. It contains a theory about the causes infered through the observed effects. More precisely formulated later by Pierre-Simon Laplace.

Immanuel Kant (1724-1804)
Ellaborates a synthesis between Descartes' rationalism and Hume's empirism, and publishes it in its treaty *Critique of Pure Reason*, in 1781. There he discuss about the *a priori,* the knowledge not based on experience and applies it to metaphysics.

Critik
der
reinen Vernunft
von
Immanuel Kant
Professor in Königsberg.
Riga,
verlegts Johann Friedrich Hartknoch
1781.

Stages in the design of a supervised classifier



- The classifier assigns a **cathegory (or class)** to the observed **feature vector**.

- **Training** entails adjusting a number of parameters using training feature vectors.

- The evaluation of the classifier has to be done as a function of the **chosen criterion**: mínimum classification error, minimum risk, etc.

- The **difficulty in the design** of the classifier depends on the variability of observations among classes:
  - Low inter-class variability
  - High intra-class variability

- **Computational efficiency** has different impact in training and in testing.

1.2 MAP DECISION RULE

- Decision rule, given vector $\mathbf{x}$

$$\omega_{MAP} = \arg\max_{\omega_i} \Pr\left(\omega_i \middle| \mathbf{x}\right)$$

- Probability of error, conditioned to $\mathbf{x}$

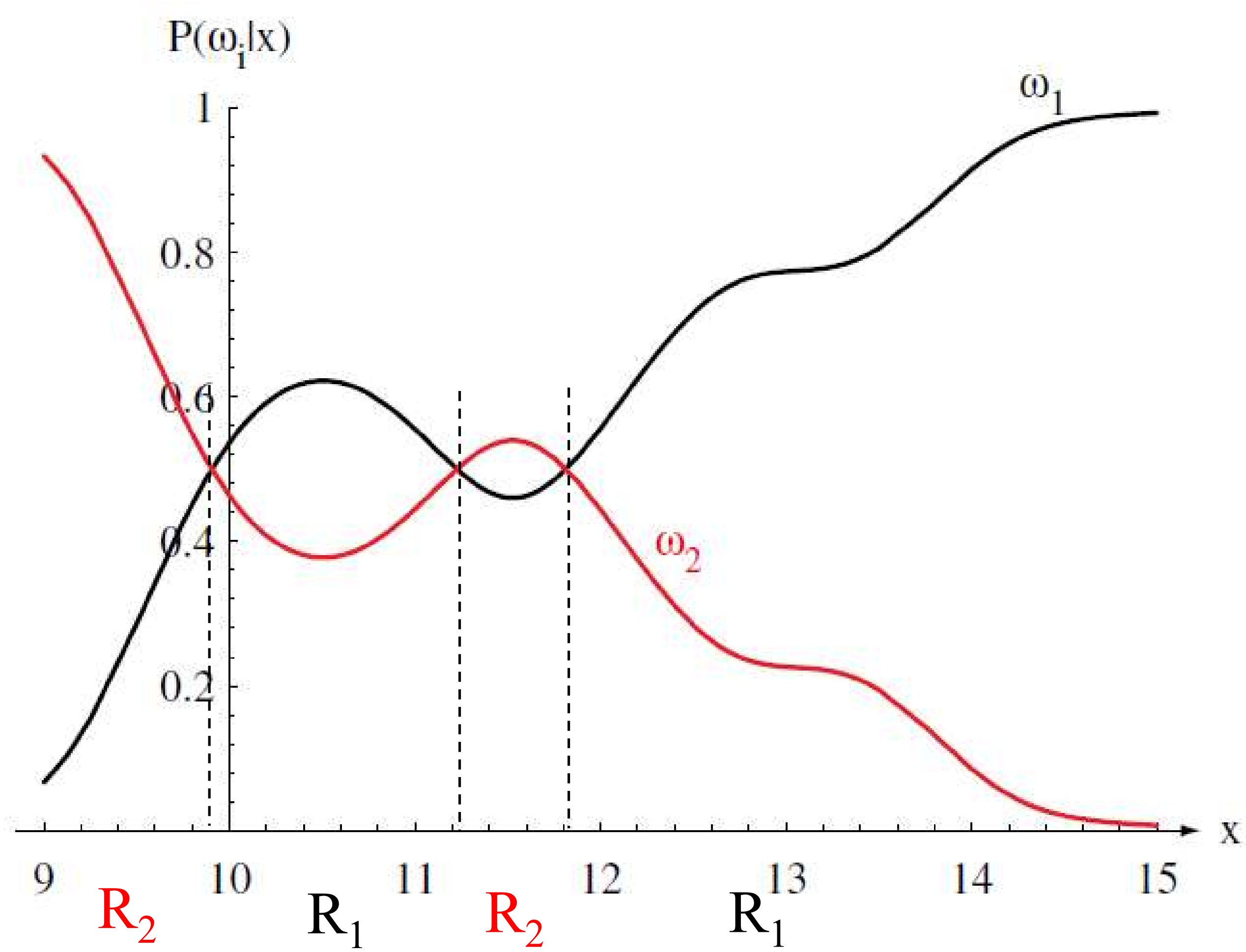
for two classes...

$$\Pr(e|\mathbf{x}) = \begin{cases} \Pr\left(\omega_1 \middle| \mathbf{x}\right) & \text{if } \omega_2 \text{ is decided} \\ \Pr\left(\omega_2 \middle| \mathbf{x}\right) & \text{if } \omega_1 \text{ is decided} \end{cases}$$

$$\Pr(e) = \Pr(\omega_2|\omega_1)\Pr(\omega_1) + \Pr(\omega_1|\omega_2)\Pr(\omega_2)$$

these are different, the second is the average of the first, over the possible values of $\mathbf{x}$.

- MAP minimizes the misclassification error rate.



$$\Pr\left(\omega_1 \middle| \mathbf{x}\right) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \Pr\left(\omega_2 \middle| \mathbf{x}\right)$$

$$\omega = \arg\max_{\omega_i} \Pr\left(\omega_i \middle| \mathbf{x}\right) = \arg\max_{\omega_i} \Pr\left(\omega_i\right) f_{\mathbf{x}}\left(\mathbf{x} \middle| \omega_i\right)$$

Particular cases:

- If one observation $\mathbf{x}_0$ does not bring information about the state of nature (class)

$$f\left(\mathbf{x}_0 \middle| \omega_1\right) = f\left(\mathbf{x}_0 \middle| \omega_2\right) \quad \Rightarrow \quad \omega = \arg\max_{\omega_i} \Pr\left(\omega_i \middle| \mathbf{x}\right) = \arg\max_{\omega_i} \Pr\left(\omega_i\right)$$

- If priors have the same value, the decision is uniquely based on the likelihood

$$\Pr(\omega_1) = \Pr(\omega_2) \quad \Rightarrow \quad \omega = \arg\max_{\omega_i} \Pr\left(\omega_i \middle| \mathbf{x}\right) = \arg\max_{\omega_i} f\left(\mathbf{x} \middle| \omega_i\right)$$

Is mimimum error rate a valid criterion?

- In **biomedical diagnose**, should I equally penalize the errors healthy/ill and ill/healthy?
- **Spam** email classification
- **Optical character recognition**, is equally important the error incurred in a vowel or in a consonant?
- **RADAR**, big difference in priors if targets are present or not in an scenario

1.3 MINIMUM RISK CLASSIFIERS

For the two classes case, the error rate probability is

$$P(e) = \Pr(\omega_1 \text{ happens and } \omega_2 \text{ is decided}) + \Pr(\omega_2 \text{ happens and } \omega_1 \text{ is decided}) =$$
$$= \Pr(\omega_2|\omega_1)\Pr(\omega_1) + \Pr(\omega_1|\omega_2)\Pr(\omega_2)$$

whereas for c classes: $P(e) = \sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} \Pr(\omega_i|\omega_j)\Pr(\omega_j)$

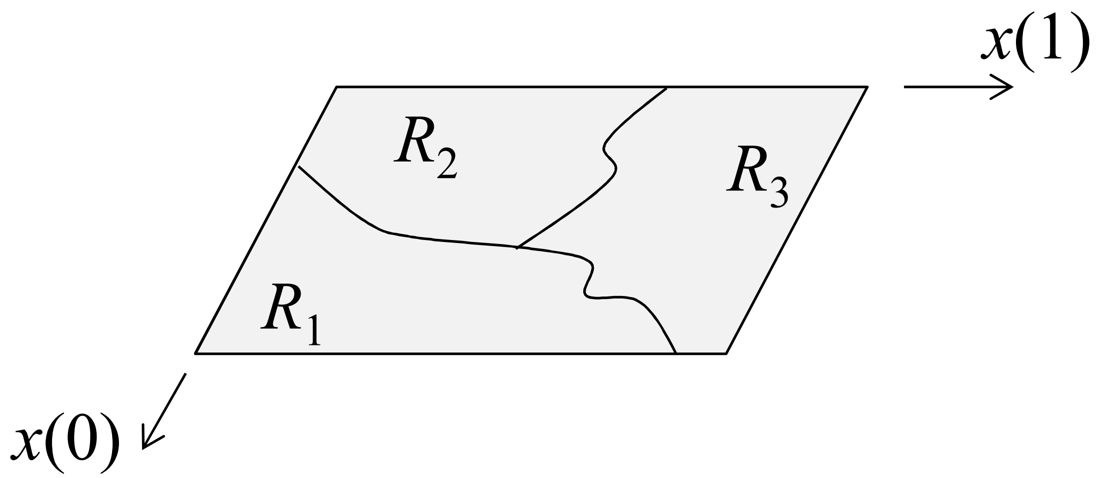
Decisions adopted may entail certains costs, and not all decisions be equally expensive. We can define the Bayes' risk function as:

$$\mathcal{R} = \sum_{i=1}^{c} \sum_{j=1}^{c} \pi_{ij} \Pr(\omega_i|\omega_j)\Pr(\omega_j)$$

where:

- π_{ij} is the cost associated to decide ω_i when ω_j happens.
- π_{ii} is the cost associated to correctly decide class ω_i

Objective: Design decision regions $R_1, \ldots, R_c$ so that the Bayes' risk is minimized.



$$\mathcal{R} = \sum_{i=1}^{c} \sum_{j=1}^{c} \pi_{ij} \Pr(\omega_i | \omega_j) \Pr(\omega_j)$$

$$R^d = R_1 \cup R_2 \ldots \cup R_C = \bigcup_{i=1}^{c} R_i$$

$$R_i \cap R_j = \varnothing$$

Theorem: For each $\mathbf{x}$, the class ω_i minimizing the risk $\mathcal{R}$ is that associated to the least conditional risk:

$$C(\omega_i|\mathbf{x}) \triangleq \sum_{j=1}^{c} \pi_{ij} \Pr(\omega_j \big| \mathbf{x})$$

Proof. The space of feature vectors $\mathbf{x}$ is split in M disjoint decision regions defined as $R_i = \{\mathbf{x} \mid \text{decide } \omega_i\}$ for $i = 0,\ldots,c$ - 1.

$$\mathcal{R} = \sum_{i=1}^{c}\sum_{j=1}^{c}\pi_{ij}\Pr(\omega_i|\omega_j)\Pr(\omega_j) = \sum_{i=1}^{c}\sum_{j=1}^{c}\pi_{ij}\left(\int_{\mathbf{x}\in R_i} f(\mathbf{x} \mid \omega_j)\,d\mathbf{x}\right)\Pr(\omega_j) =$$

$$= \sum_{i=1}^{c}\int_{\mathbf{x}\in R_i}\sum_{j=1}^{c}\pi_{ij} f(\mathbf{x} \mid \omega_j)\Pr(\omega_j)\,d\mathbf{x} =$$

Bayes theorem

$$= \sum_{i=1}^{c}\int_{\mathbf{x}\in R_i}\underbrace{\sum_{j=1}^{c}\pi_{ij}\Pr(\omega_j \mid \mathbf{x})}_{C_i(\mathbf{x}) \triangleq \text{Risk associated class } \omega_i} f(\mathbf{x})\,d\mathbf{x}$$

Therefore, to minimize $\mathcal{R}$: R_i is the domain of values of $\mathbf{x}$ where $C_i(\mathbf{x})$ is smaller than $C_k(\mathbf{x})$ for all $k \neq i$:

$$R_i = \left\{\mathbf{x} \in \mathbb{C}^{N\times 1} \;\middle|\; \sum_{j=1}^{c}\pi_{ij}\Pr(\omega_j \mid \mathbf{x}) < \sum_{j=1}^{c}\pi_{kj}\Pr(\omega_j \mid \mathbf{x}), \quad \forall k \neq i\right\}$$

♦

Minimizing Bayesian risk implies minimizing... $\omega_{MRB} = \arg\min_j C\left(\omega_j \middle| \mathbf{x}\right)$

- Binary case

$c = 2$ categories

Conditional risk

$$C\left(\omega_1 \middle| \mathbf{x}\right) = \pi_{11} \Pr\left(\omega_1 \middle| \mathbf{x}\right) + \pi_{12} \Pr\left(\omega_2 \middle| \mathbf{x}\right)$$
$$C\left(\omega_2 \middle| \mathbf{x}\right) = \pi_{21} \Pr\left(\omega_1 \middle| \mathbf{x}\right) + \pi_{22} \Pr\left(\omega_2 \middle| \mathbf{x}\right)$$

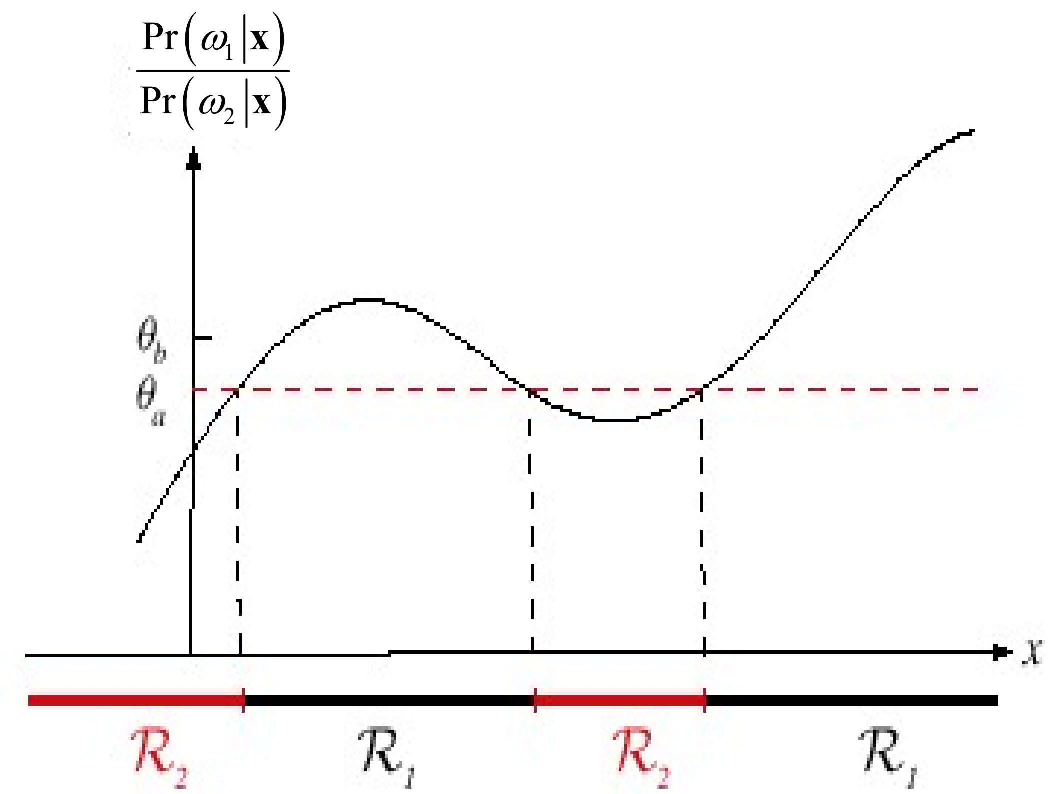
Likelihood ratio...

$$\frac{\Pr\left(\omega_1 \middle| \mathbf{x}\right)}{\Pr\left(\omega_2 \middle| \mathbf{x}\right)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \left(\frac{\pi_{12} - \pi_{22}}{\pi_{21} - \pi_{11}}\right) = \gamma$$

Minimum risk criterion biases the likelihood ratio with respect to MAP

Threshold independent of $\mathbf{x}$

- Increasing the threshold, more decisions on R_2 are taken (and viceversa)



- Minimum risk = Minimum Pr(e) if $\pi_{ij} = \begin{cases} \alpha & i = j \\ \pi & i \neq j \end{cases}$

Proof:

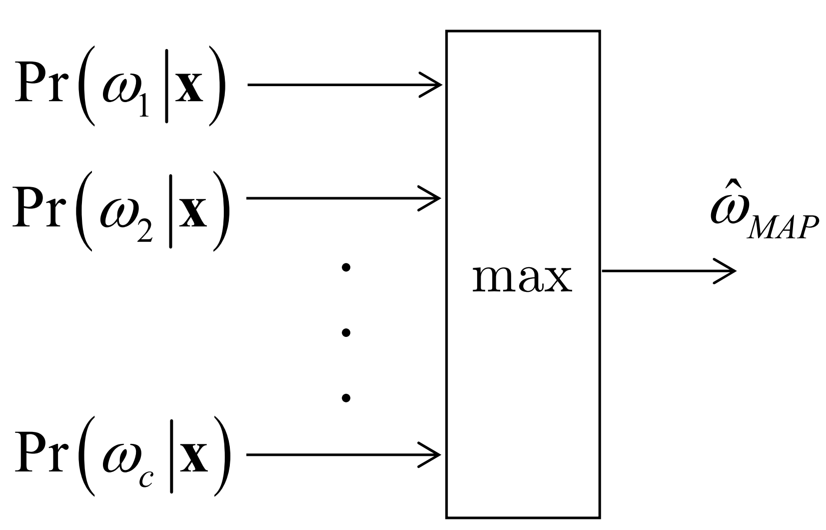
$$C(\omega_i|\mathbf{x}) = \sum_{j=1}^{c} \pi_{ij} \Pr(\omega_j|\mathbf{x}) = \alpha \Pr(\omega_i|\mathbf{x}) + \sum_{\substack{j=1 \\ j\neq i}}^{c} \pi \Pr(\omega_j|\mathbf{x})$$

MBR criterion becomes MAP

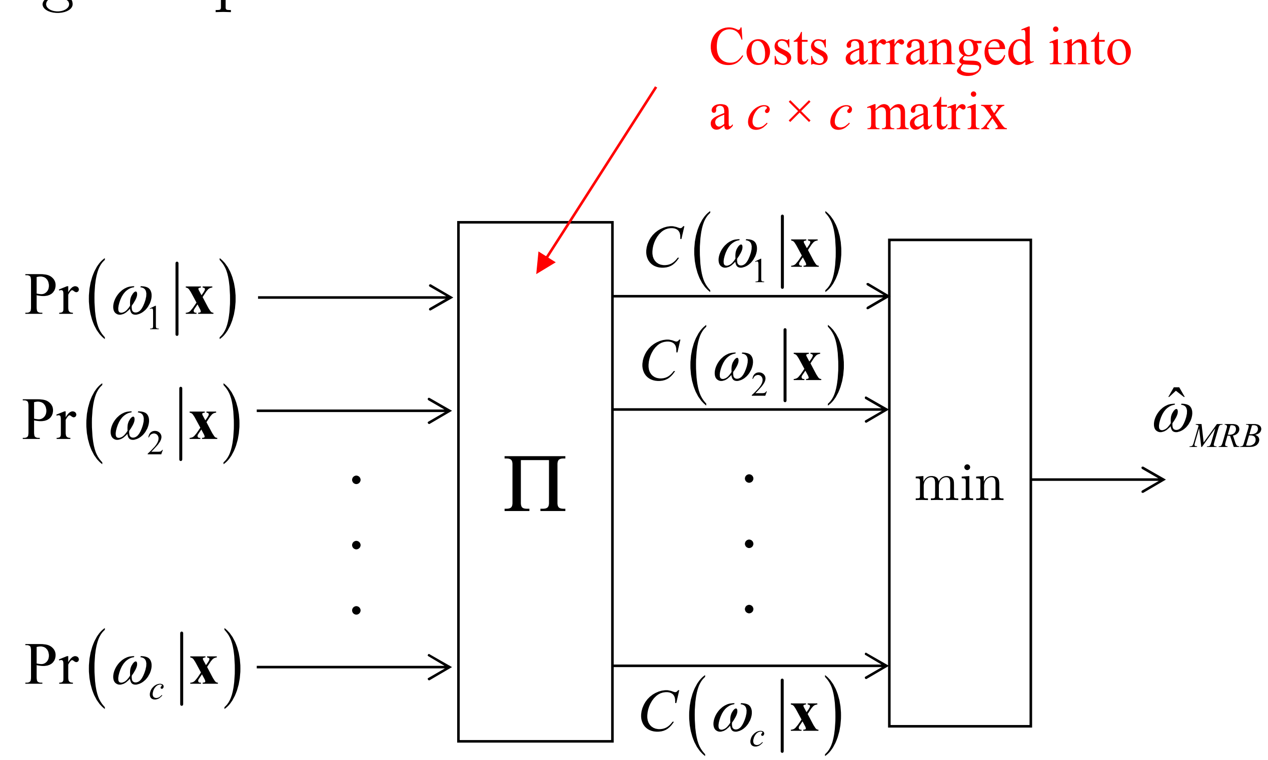
$$= \alpha \Pr(\omega_i|\mathbf{x}) + \pi\left(1 - \Pr(\omega_i|\mathbf{x})\right) = \pi - (\pi - \alpha)\Pr(\omega_i|\mathbf{x})$$

- In brief...

- MAP is a particular case of the minimum bayesian risk.
- Both are implemented using the posteriors:



MAP



Minimum Bayesian risk

Other criteria for decision can be defined…

- NEYMAN-PEARSON
  - Total risk is minimized subject to a restriction, e.g. upper bounding the classification error for class i, priors do not intervene:

$$\int_{\mathbf{x}\in R_i} C\left(\omega_i \middle| \mathbf{x}\right) f(\mathbf{x}) d\mathbf{x} < \text{constant}$$

- MINMAX
  - Used when a priori probabilities are not known, maybe because they are changing over time in an unknown way.
  - Minimizes the **worst total risk**, choosing the decision regions in such a way that the risk function does not depend on the a priori probabilities.
  - Example for $c = 2$ categories…

R_1 : region where ω_1 is decided

$$\mathcal{R} = \int_{R_1} C\left(\omega_1 \middle| \mathbf{x}\right) f\left(\mathbf{x}\right) d\mathbf{x} + \int_{R_2} C\left(\omega_2 \middle| \mathbf{x}\right) f\left(\mathbf{x}\right) d\mathbf{x}$$

$$= \int_{R_1} \left(\pi_{11} \Pr\left(\omega_1\right) f_{\mathbf{x}}(\mathbf{x}|\omega_1) + \pi_{12} \Pr\left(\omega_2\right) f_{\mathbf{x}}(\mathbf{x}|\omega_2)\right) d\mathbf{x} +$$

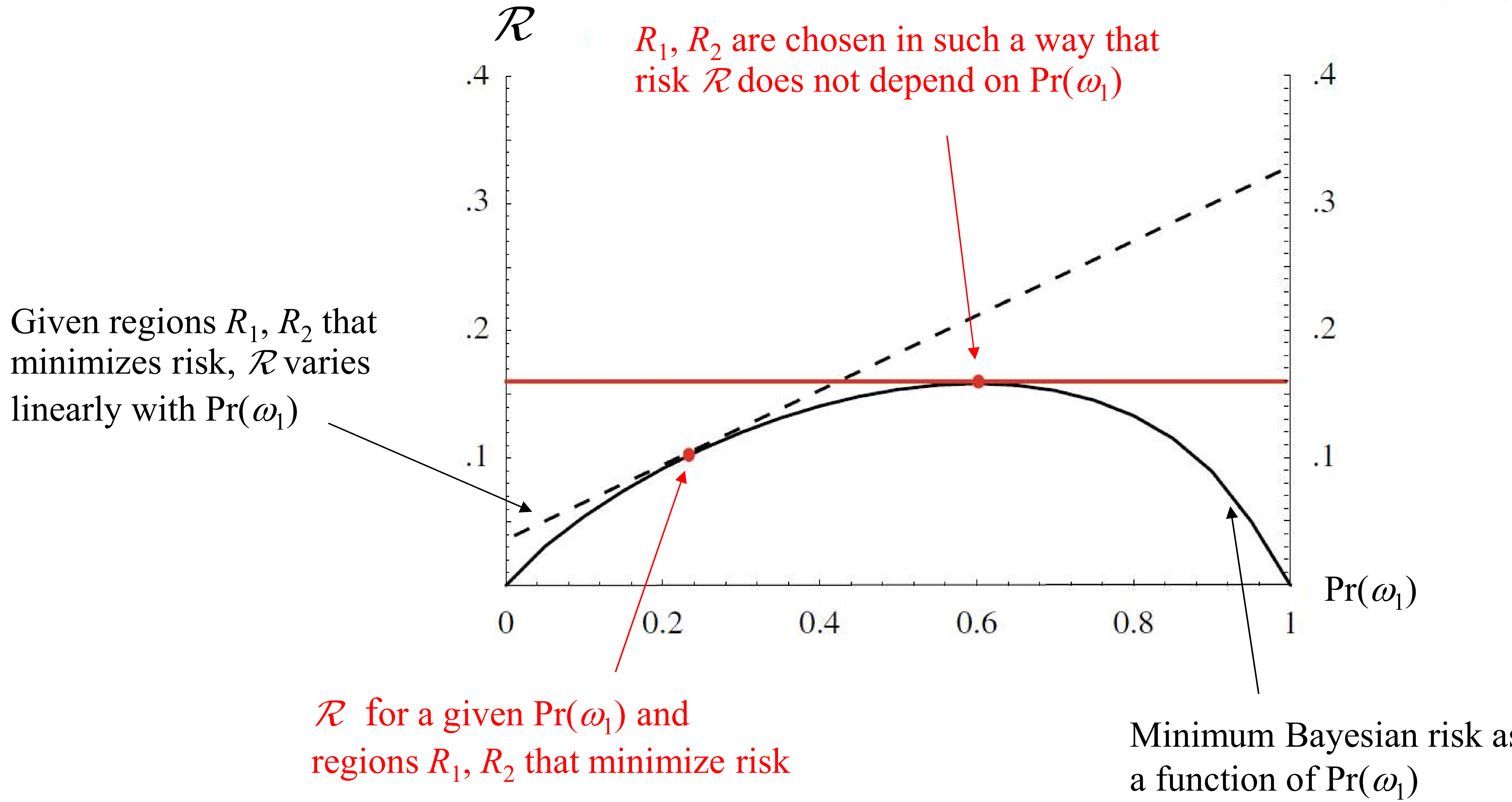
$$+ \int_{R_2} \left(\pi_{21} \Pr\left(\omega_1\right) f_{\mathbf{x}}(\mathbf{x}|\omega_1) + \pi_{22} \Pr\left(\omega_2\right) f_{\mathbf{x}}(\mathbf{x}|\omega_2)\right) d\mathbf{x} =$$

$$= \left\{ \begin{array}{l} \Pr\left(\omega_1\right) = 1 - \Pr\left(\omega_2\right) \\ \int_{R_1} f_{\mathbf{x}}(\mathbf{x}|\omega_1) d\mathbf{x} = 1 - \int_{R_2} f_{\mathbf{x}}(\mathbf{x}|\omega_1) d\mathbf{x} \end{array} \right\} = \pi_{22} + (\pi_{12} - \pi_{22}) \int_{R_1} f_{\mathbf{x}}(\mathbf{x}|\omega_2) d\mathbf{x} +$$

$$+ \Pr\left(\omega_1\right) \left(\pi_{11} - \pi_{22} + \left(\pi_{21} - \pi_{11}\right) \int_{R_2} f_{\mathbf{x}}(\mathbf{x}|\omega_1) d\mathbf{x} - \left(\pi_{12} - \pi_{22}\right) \int_{R_1} f_{\mathbf{x}}(\mathbf{x}|\omega_2) d\mathbf{x}\right)$$

If choosing regions R_1 and R_2 in such a way that this term is cancelled, $\mathcal{R}$ does not depend on $\Pr(\omega_1)$

$$\mathcal{R}_{\text{mini-max}} = \pi_{22} + (\pi_{12} - \pi_{22}) \int_{R_1} f_{\mathbf{x}}(\mathbf{x}|\omega_2) d\mathbf{x}$$


$\mathcal{R}$
R_1, R_2 are chosen in such a way that risk $\mathcal{R}$ does not depend on $\Pr(\omega_1)$
.4
.3
.2
.1
Given regions R_1, R_2 that minimizes risk, $\mathcal{R}$ varies linearly with $\Pr(\omega_1)$
$\Pr(\omega_1)$
0
0.2
0.4
0.6
0.8
1
$\mathcal{R}$ for a given $\Pr(\omega_1)$ and regions R_1, R_2 that minimize risk
Minimum Bayesian risk as a function of $\Pr(\omega_1)$

Course overview

Pre-processing → Feature selection → Classification

Techniques
PCA (lecture 3) MDA (lecture 3)

	Density function for classes	Classifier
Supervised	Parametrized Gaussian	Linear discriminant Quadratic discriminant[1] (lecture 2)
	Non-parametric	K-nearest neighbours[2] Fuzzy logic (lecture 4)
	Unknown	SVM[3] Neural networks[3] Decision trees[4] (lecture 4)
Unsupervised	Parametrized Gaussian	K-means (lecture 6)
	Unknown	Clustering (lecture 6)

1. Useful only if covariance matrices are not rank deficient.
2. Useful with the number of features is very large, even larger that the number of training vectors.
3. Imposes a structure to the classifier irrespective of the training data base.
4. Useful when non-numeric features are present.

1.4 DISCRIMINANTS AND DECISION REGIONS

- Definition of a discriminant function g_i for class i:
  - The classifier uses it to assign a class ω_i to a feature vector $\mathbf{x}$.
  - Classification criterion: decide class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x}) \qquad \forall j \neq i$

Minimum risk $\quad g_i(\mathbf{x}) = -C\left(\omega_i \middle| \mathbf{x}\right)$

MAP $\quad g_i(\mathbf{x}) = \Pr\left(\omega_i \middle| \mathbf{x}\right)$

A non-decreasing function can be applied and the criterion does not change

$$g_i(\mathbf{x}) = \Pr\left(\omega_i \middle| \mathbf{x}\right)$$

ln(.) is an increasing function

$$\Downarrow$$

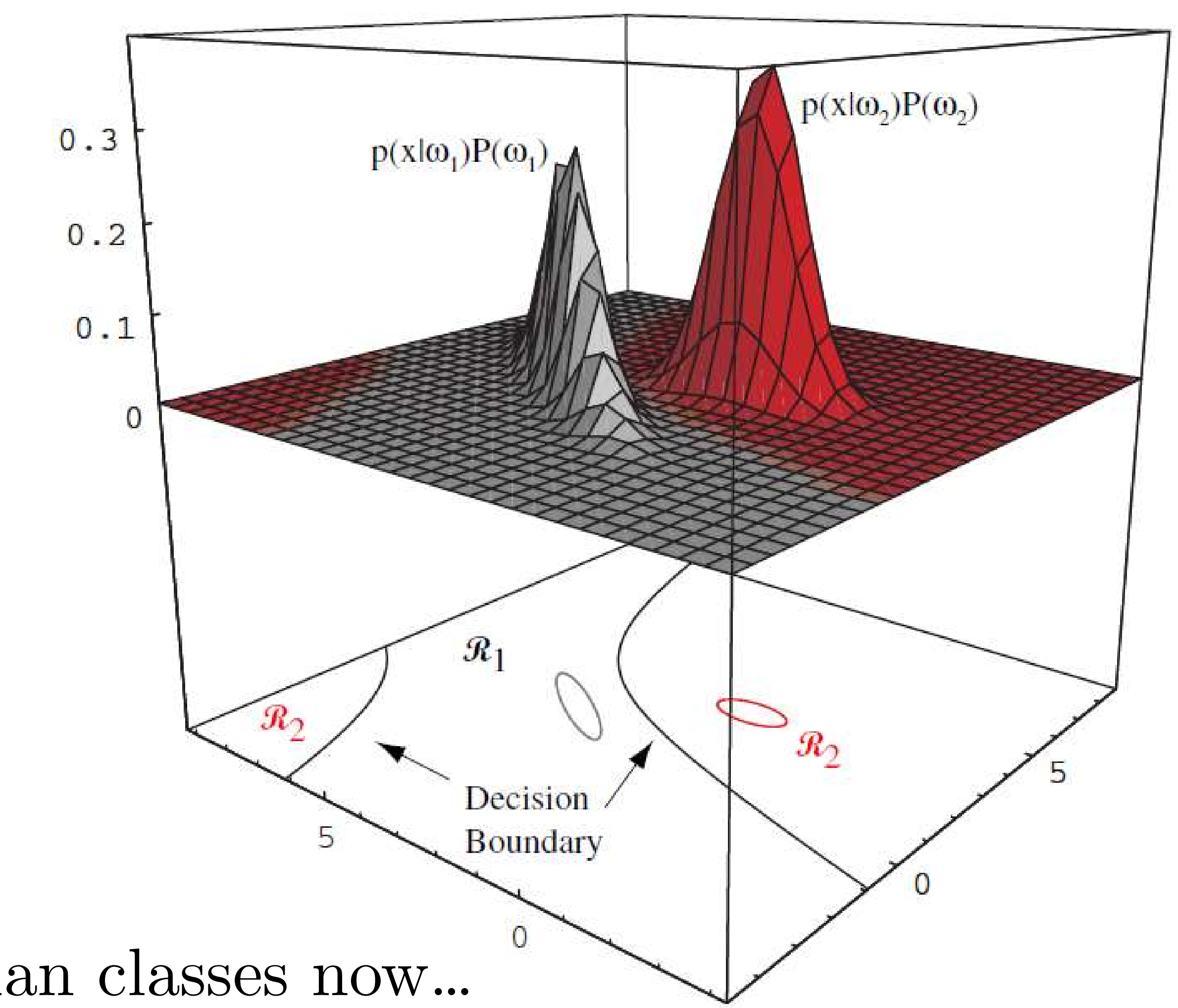
$$h_i(\mathbf{x}) = \ln \Pr\left(\omega_i \middle| \mathbf{x}\right) = \ln f_{\mathbf{x}}(\mathbf{x}|\omega_i) + \ln \Pr\left(\omega_i\right)$$

- For $c = 2$ categories: the decision regions are given by...

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

- Decision boundary: $g(\mathbf{x}) = 0$

- Examples
  - Binary communications BPSK, FSK -2
  - Detection of illness: Y/N



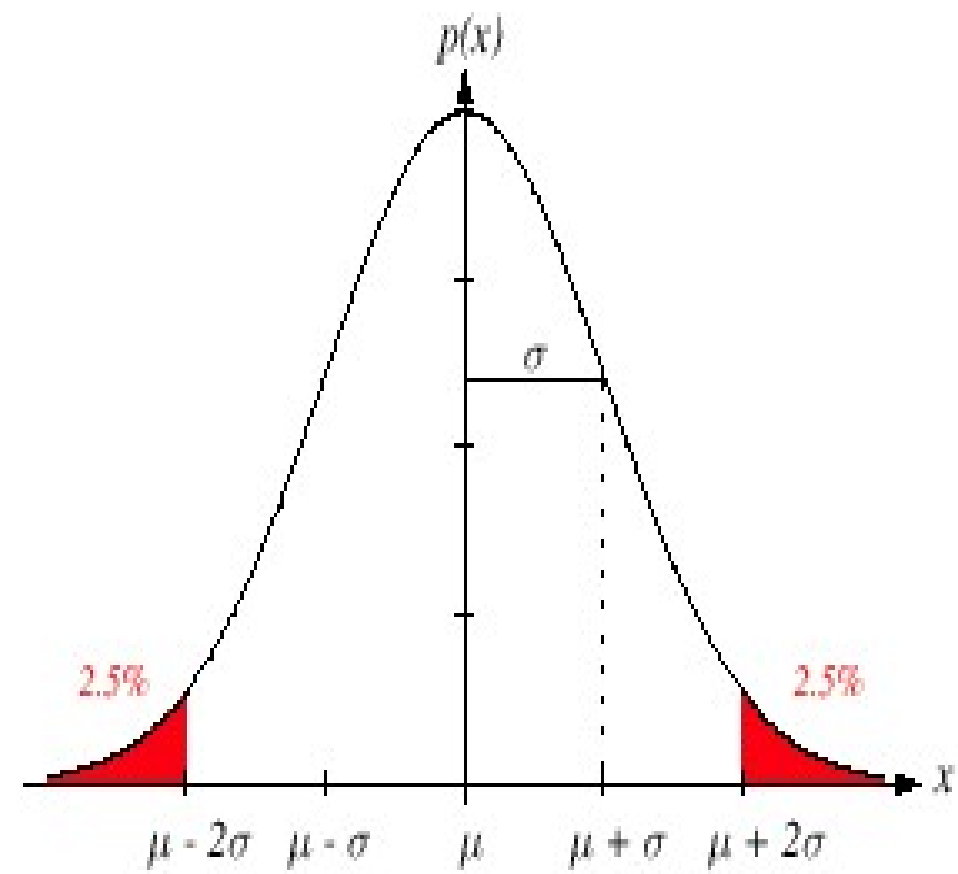
Let us consider the case of Gaussian classes now...

1.5 GAUSSIAN DENSITY FUNCTION

- **Univariate case...**

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

$$\mu = E\{x\} \qquad \sigma^2 = E\left\{(x-\mu)^2\right\}$$



- **Multivariate case...**

  - Statistical moments

    $$\mathbf{x} \in \mathbb{R}^d \qquad \boldsymbol{\mu} = E\{\mathbf{x}\} \in \mathbb{R}^d \qquad \mathbf{C} = E\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right\} \in \mathbb{R}^{d\times d}$$

  - Covariance matrix is positive semi-definite (real non-negative eigenvalues)
  - Density function of vector $\mathbf{x}$:

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\left|\mathbf{C}\right|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

– Linear transformations of Gaussian random variables are also Gaussian:

$$\mathbf{A} \in \mathbb{R}^{d \times k} \qquad \mathbf{y} = \mathbf{A}^T \mathbf{x} \in \mathbb{R}^k$$

$$\boldsymbol{\mu}_\mathbf{y} = E\{\mathbf{y}\} = E\{\mathbf{A}^T \mathbf{x}\} = \mathbf{A}^T E\{\mathbf{x}\} = \mathbf{A}^T \boldsymbol{\mu}_\mathbf{x}$$

$$\begin{aligned} \mathbf{C}_\mathbf{y} &= E\left\{(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})^T\right\} = \\ &= E\left\{(\mathbf{A}^T \mathbf{x} - \mathbf{A}^T \boldsymbol{\mu}_\mathbf{x})(\mathbf{A}^T \mathbf{x} - \mathbf{A}^T \boldsymbol{\mu}_\mathbf{x})^T\right\} = \\ &= E\left\{\mathbf{A}^T (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^T \mathbf{A}\right\} = \mathbf{A}^T \mathbf{C}_\mathbf{x} \mathbf{A} \end{aligned}$$

Whitening of a feature vector

– Spectral decomposition of $\mathbf{C}$

$$\mathbf{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

$$\mathbf{C} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

– Orthonormal eigenvectors $\mathbf{U} = \left[\mathbf{u}_1, \mathbf{u}_2, .., \mathbf{u}_d\right]$ $\quad \mathbf{U}\mathbf{U}^T = \mathbf{I}$

– Eigenvalues $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_d)$

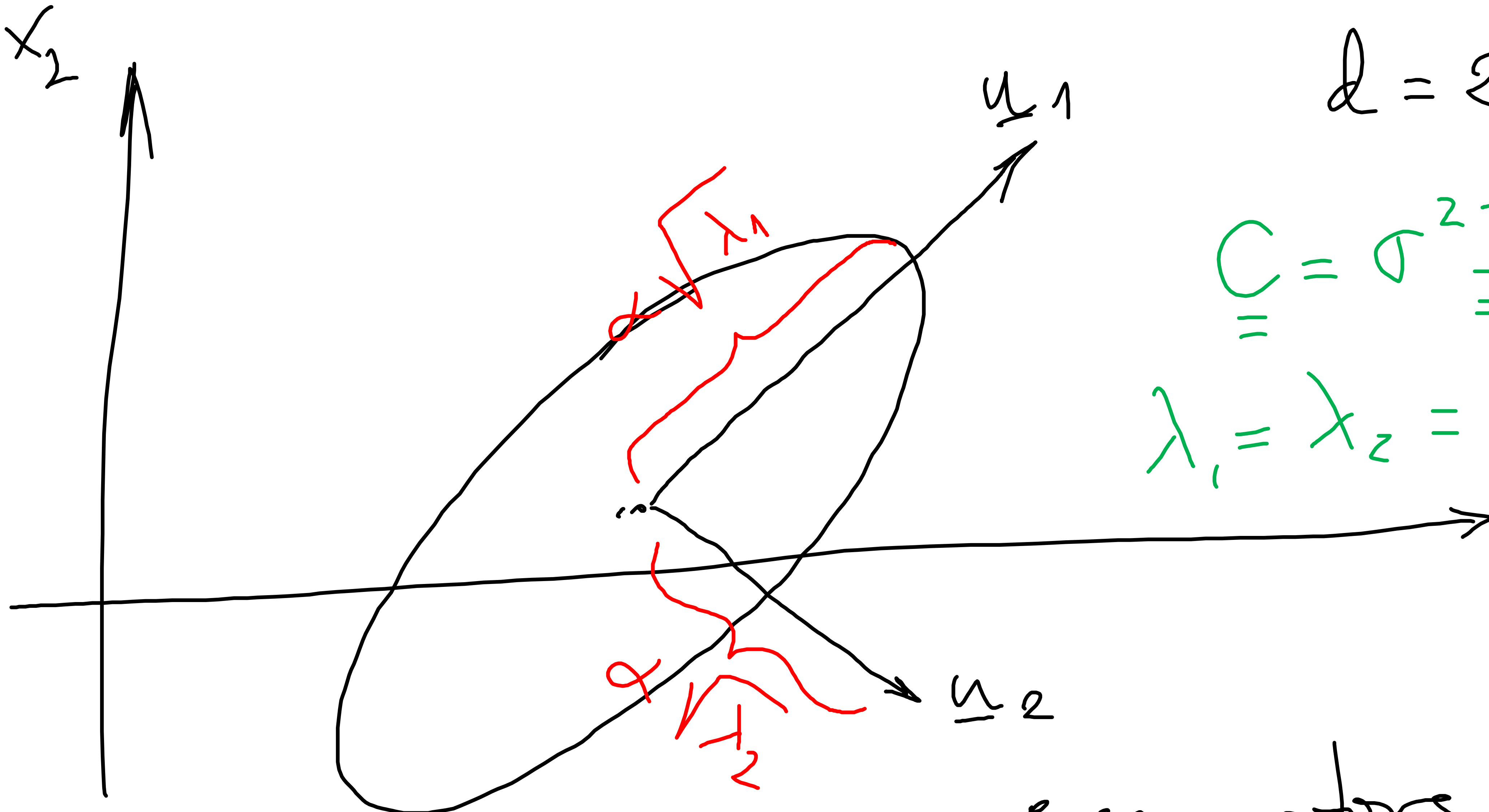
– A linear transform “whitens” the elements of vector $\mathbf{x}$

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}^{-1/2} \qquad \boldsymbol{\Lambda}^{-1/2} = diag\left(\sqrt{\lambda}_1, \sqrt{\lambda}_2, ..., \sqrt{\lambda}_d\right)$$

also

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T$$

- The sections of the gaussians are hiper-ellipsoids in a space of dimension d.
- Whitening turns hiper-ellipsoids into hiper-spheres.
- The samples of a Gaussian cluster are grouped around $\boldsymbol{\mu}$
- The main axis of the hiper-ellipsoids follow the direction of the eigenvectors of $\mathbf{C}$.
- The length of the principal axis of the hiper-ellipsoids are proportional to the square root of the eigenvalues.



$d = 2$

$\underline{\underline{C}} = \sigma^2 \underline{\underline{I}}$

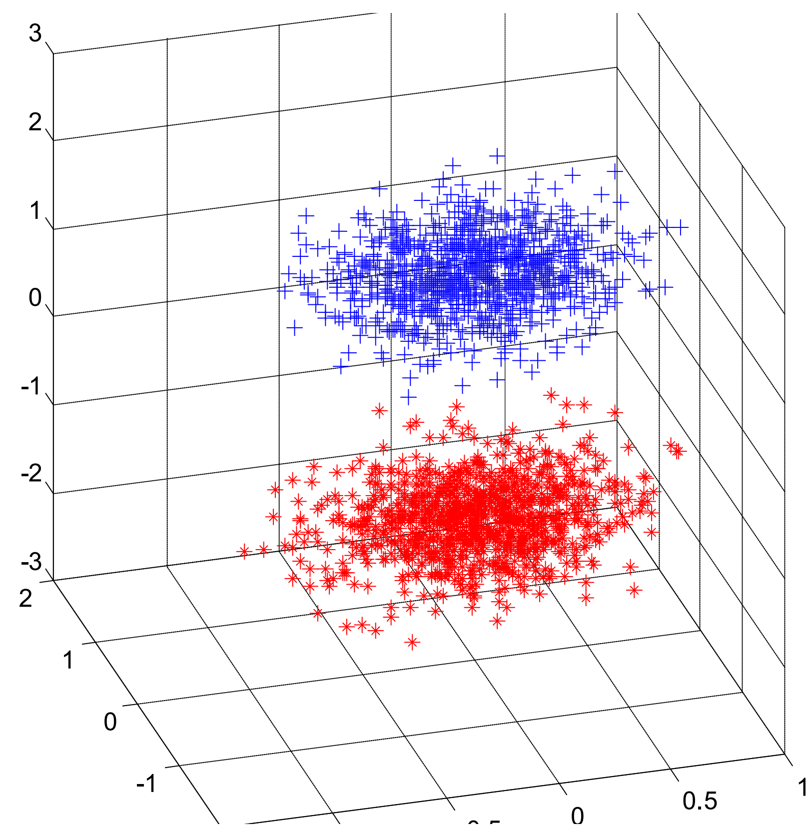
$\lambda_1 = \lambda_2 = \sigma^2$

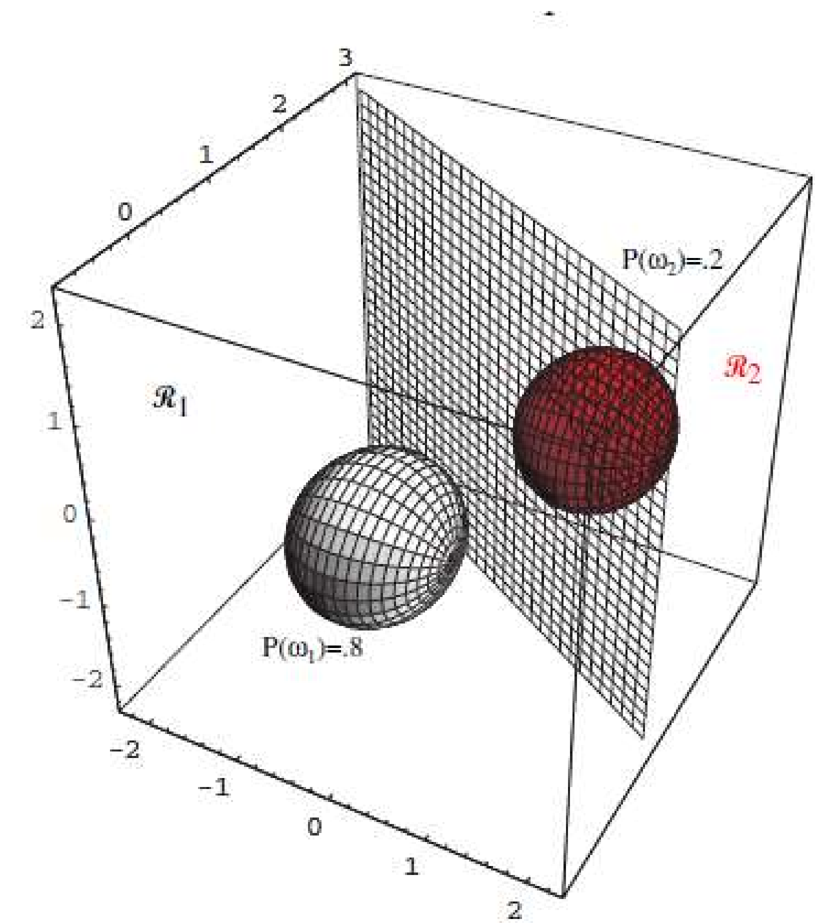
eigenvectors of $\underline{\underline{C}}_i$ of ω_i

$\underline{\underline{C}}_i \in \mathbb{R}^{2\times 2}$

$\underline{\underline{C}}\,\underline{u}_k = \lambda_k \underline{u}_k \quad k = 1, 2$

- Gaussian clusters...





1.6 DISCRIMINANTS FOR GAUSSIAN CLASSES

- Density function for class i: $f_{\mathbf{x}}(\mathbf{x}|\omega_i) \sim N\left(\boldsymbol{\mu}_i, \mathbf{C}_i\right)$

- A priori probability: $\Pr\left(\omega_i\right)$

- Discriminant function for MAP

$$h_i(\mathbf{x}) = \ln f_{\mathbf{x}}(\mathbf{x}|\omega_i) + \ln \Pr\left(\omega_i\right) =$$

$$= -\tfrac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T \mathbf{C}_i^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right) - \tfrac{d}{2}\ln\left(2\pi\right) - \tfrac{1}{2}\ln\left|\mathbf{C}_i\right| + \ln \Pr\left(\omega_i\right)$$

- Three cases for the covariance matrix...
  - Case 1 $\mathbf{C}_i = \sigma^2 \mathbf{I}$
  - Case 2 $\mathbf{C}_i = \mathbf{C}$
  - Case 3 $\mathbf{C}_i$ arbitrary

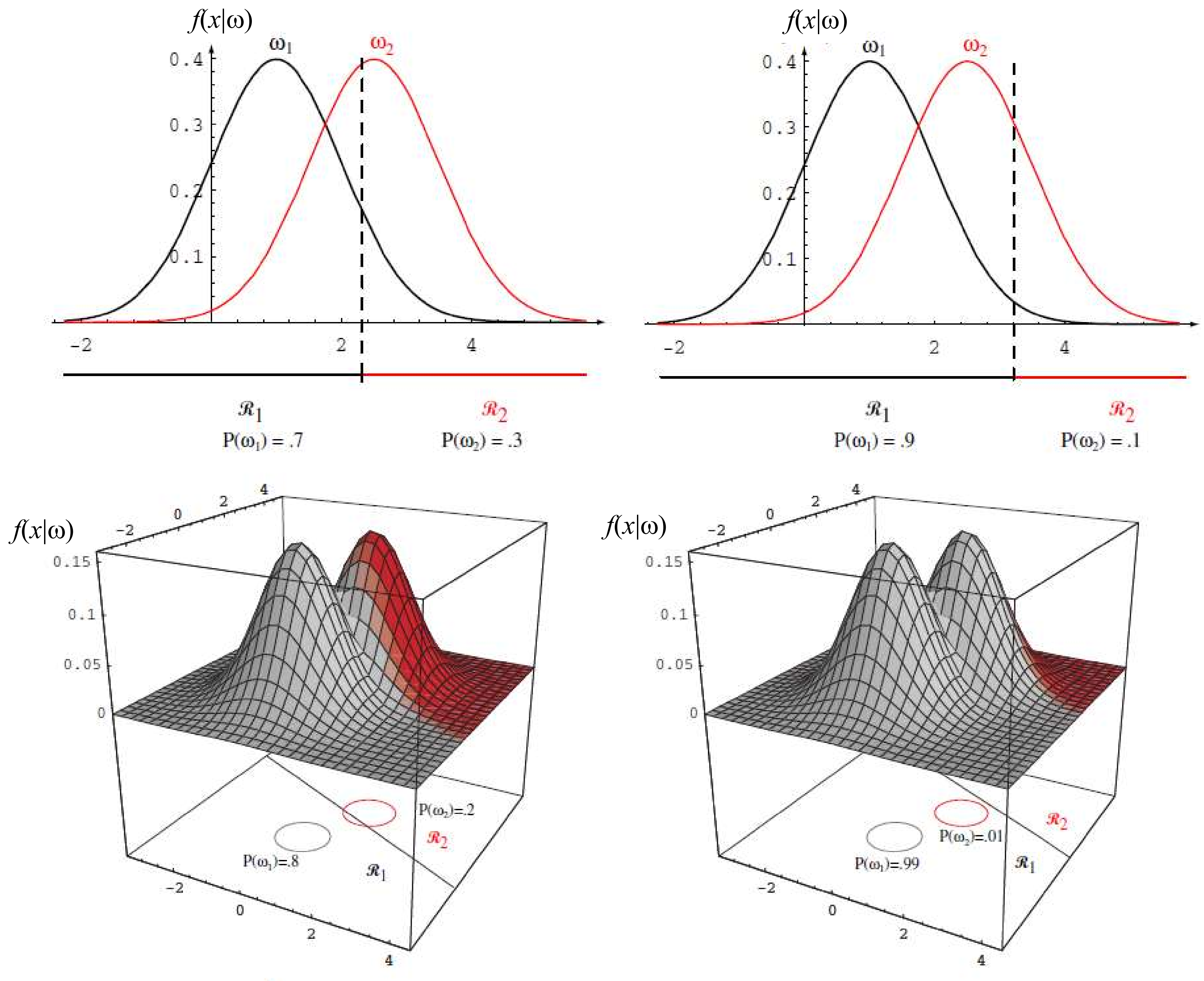
Case 1 $\mathbf{C}_i = \sigma^2 \mathbf{I}$

- The discriminant depends on the euclidean distance

$$h_i(\mathbf{x}) = -\tfrac{1}{2\sigma^2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T\left(\mathbf{x}-\boldsymbol{\mu}_i\right) + \ln \Pr\left(\omega_i\right)$$

- Decision boundaries are **hyperplanes**

$$h(\mathbf{x}) = h_i(\mathbf{x}) - h_j(\mathbf{x}) = 0 \quad \Rightarrow \quad \mathbf{w}^T\left(\mathbf{x}-\mathbf{x}_0\right) = 0$$


f(x|ω)
ω1
ω2
0.4
0.3
0.2
0.1
-2
2
4
ℛ1
P(ω1) = .7
ℛ2
P(ω2) = .3
f(x|ω)
ω1
ω2
0.4
0.3
0.2
0.1
-2
2
4
ℛ1
P(ω1) = .9
ℛ2
P(ω2) = .1
f(x|ω)
0.15
0.1
0.05
0
-2
0
2
4
P(ω2)=.2
ℛ2
P(ω1)=.8
ℛ1
f(x|ω)
0.15
0.1
0.05
0
-2
0
2
4
P(ω2)=.01
ℛ2
P(ω1)=.99
ℛ1

Exercise. Prove that:

$$h(\mathbf{x}) = h_i(\mathbf{x}) - h_j(\mathbf{x}) = 0 \quad \Rightarrow \quad \mathbf{w}^T\left(\mathbf{x} - \mathbf{x}_0\right) = 0$$

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

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

It is also called minimum distance classifier.

Case 2 $\mathbf{C}_i = \mathbf{C}$

- The discriminant depends on the Mahalanobis distance:

$$h_i(\mathbf{x}) = -\tfrac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T \mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right) + \ln \Pr\left(\omega_i\right)$$

- Decision boundaries are **hyperplanes**

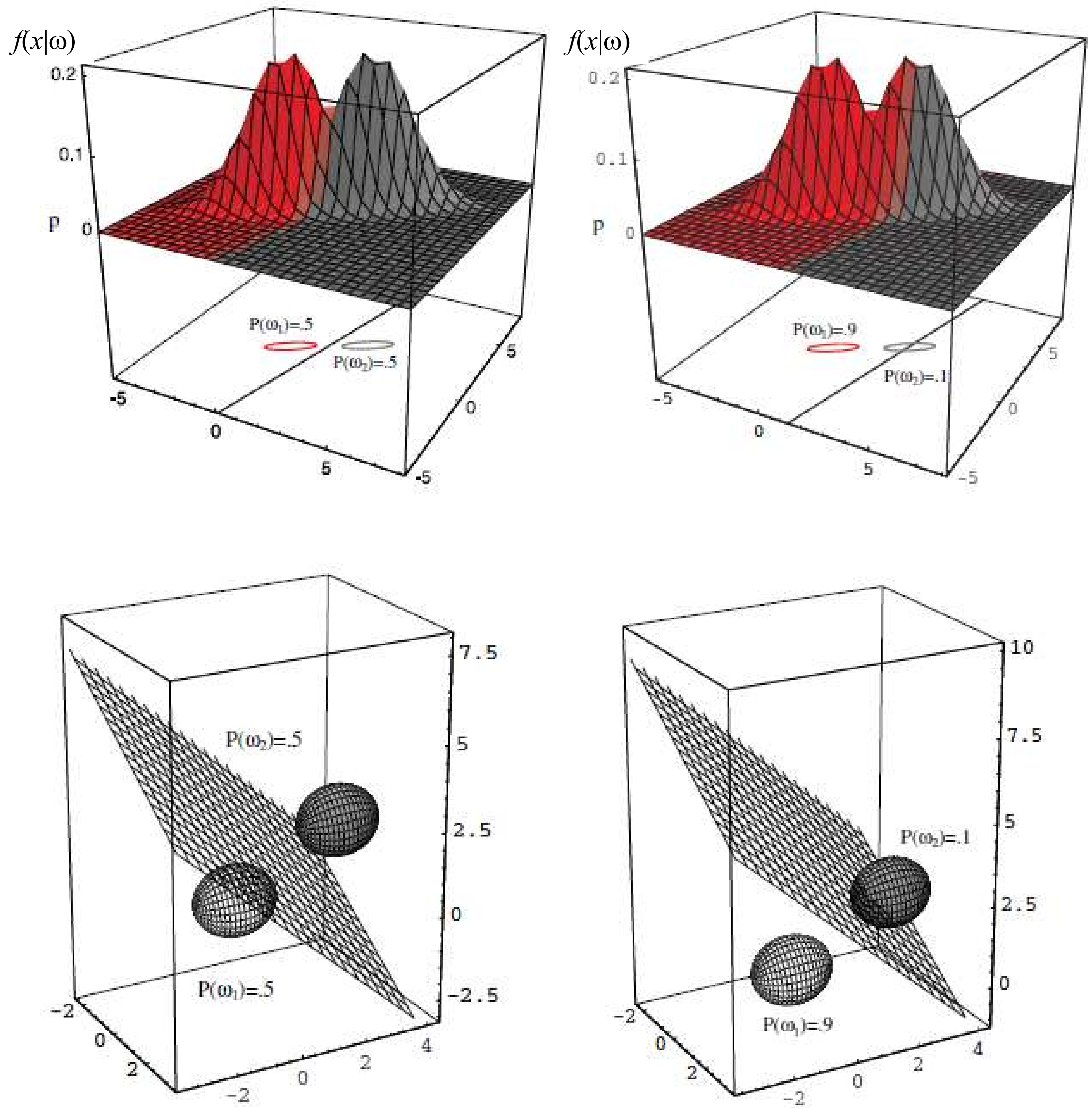
$$h(\mathbf{x}) = h_i(\mathbf{x}) - h_j(\mathbf{x}) = 0 \quad \Rightarrow \quad \mathbf{w}^T\left(\mathbf{x}-\mathbf{x}_0\right) = 0$$

Exercise. Prove that:

$$h(\mathbf{x}) = h_i(\mathbf{x}) - h_j(\mathbf{x}) = 0 \quad \Rightarrow \quad \mathbf{w}^T\left(\mathbf{x} - \mathbf{x}_0\right) = 0$$

$$\mathbf{w} = \mathbf{C}^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)$$

$$\mathbf{x}_0 = \tfrac{1}{2}\left(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j\right) - \frac{\ln\left(\Pr\left(\omega_i\right)\right) - \ln\left(\Pr\left(\omega_j\right)\right)}{\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \mathbf{C}^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)$$

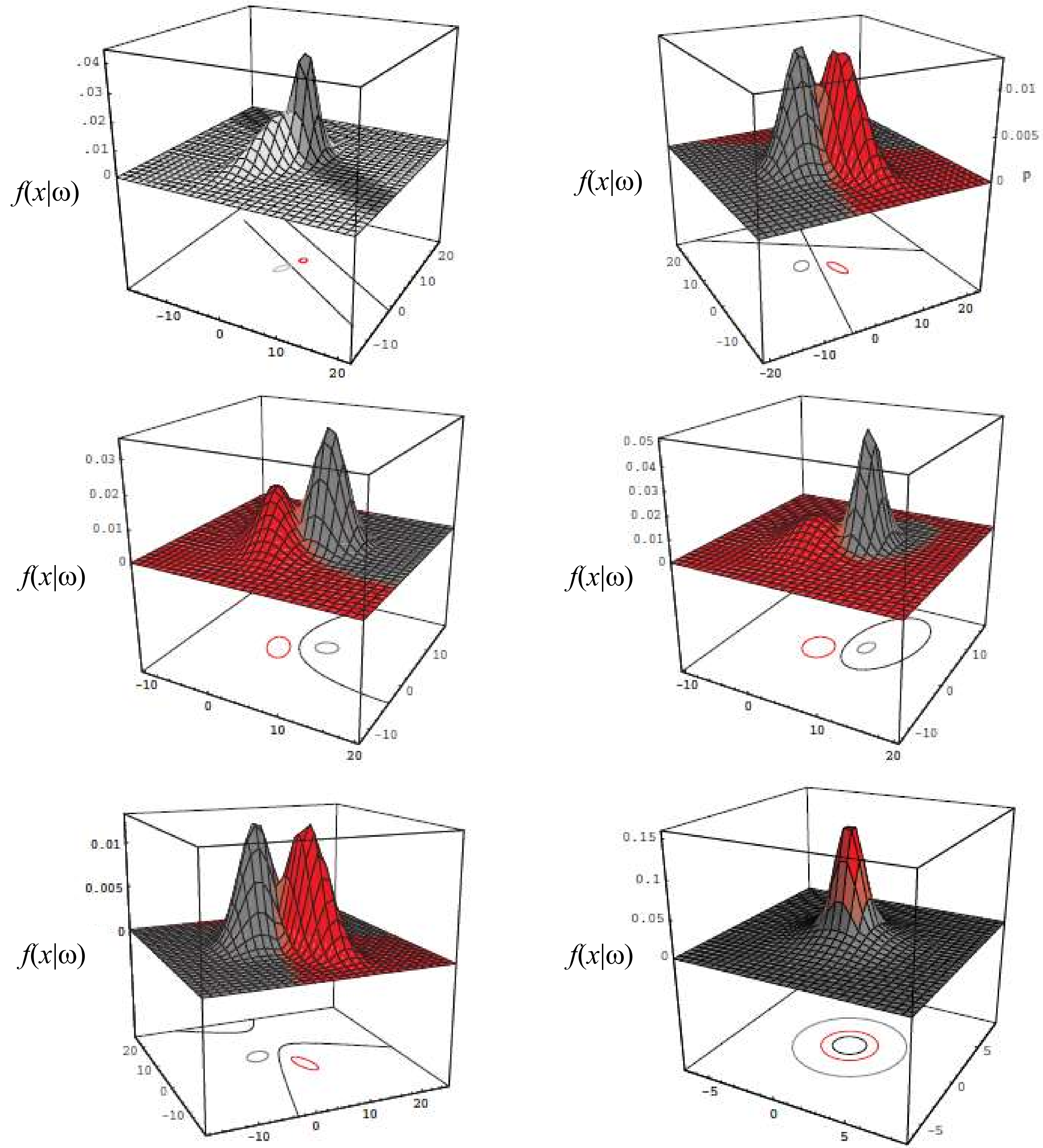


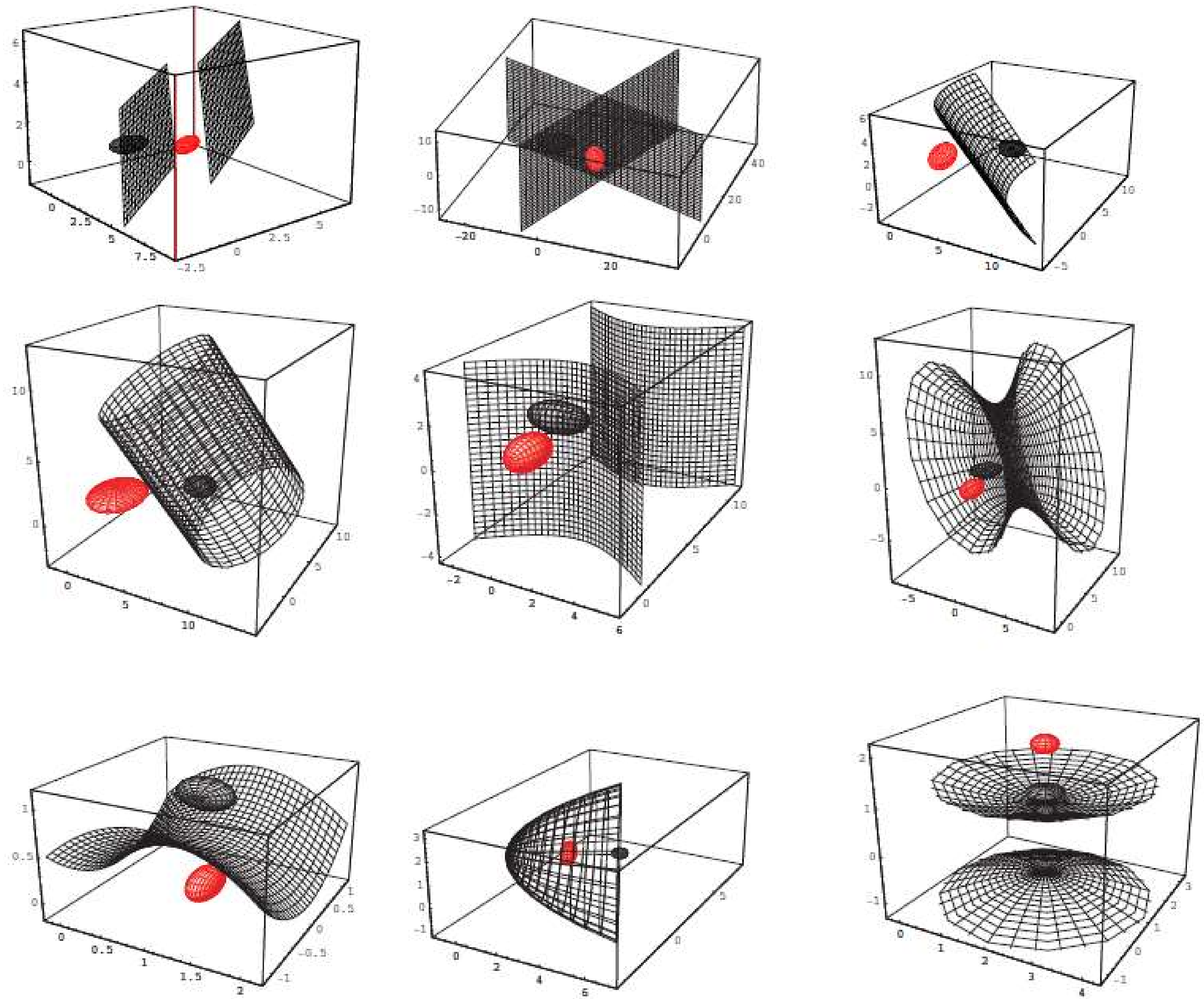
f(x|ω)
0.2
0.1
P
0
P(ω1)=.5
P(ω2)=.5
-5
0
5
f(x|ω)
0.2
0.1
P
0
P(ω1)=.9
P(ω2)=.1
-5
0
5
7.5
5
2.5
0
-2.5
P(ω2)=.5
P(ω1)=.5
-2
0
2
4
10
7.5
5
2.5
0
P(ω2)=.1
P(ω1)=.9
-2
0
2
4

Case 3 $\mathbf{C}_i$ arbitrary

$$h_i(\mathbf{x}) = -\tfrac{1}{2}\mathbf{x}^T\mathbf{C}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\mathbf{C}_i^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i^T\mathbf{C}_i^{-1}\boldsymbol{\mu}_i - \tfrac{1}{2}\ln\left|\mathbf{C}_i\right| + \ln \Pr\left(\omega_i\right)$$

- Decision boundaries separating two classes are hiper-quadratic:
  - **Hyper-planes**
  - **Hyper-spheres**
  - **Hyper-ellipsoids**
  - **Hyper-paraboloids**
  - **Hyper-hyperboloids**


f(x|ω)
f(x|ω)
f(x|ω)
f(x|ω)
f(x|ω)
f(x|ω)



Exercise. Determine the surfaces separating 2 regions

$$h_i(\mathbf{x}) = h_j(\mathbf{x}) \Rightarrow$$

$$-\tfrac{1}{2}\mathbf{x}^T\mathbf{C}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\mathbf{C}_i^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i^T\mathbf{C}_i^{-1}\boldsymbol{\mu}_i - \tfrac{1}{2}\ln\left|\mathbf{C}_i\right| + \ln\Pr\left(\omega_i\right)$$

$$+\tfrac{1}{2}\mathbf{x}^T\mathbf{C}_j^{-1}\mathbf{x} - \boldsymbol{\mu}_j^T\mathbf{C}_j^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}_j^T\mathbf{C}_j^{-1}\boldsymbol{\mu}_j + \tfrac{1}{2}\ln\left|\mathbf{C}_j\right| - \ln\Pr\left(\omega_j\right) = 0$$

$$\Rightarrow$$

$$\mathbf{x}^T\left(\tfrac{1}{2}\mathbf{C}_j^{-1} - \tfrac{1}{2}\mathbf{C}_i^{-1}\right)\mathbf{x} + \left(\boldsymbol{\mu}_i^T\mathbf{C}_i^{-1} - \boldsymbol{\mu}_j^T\mathbf{C}_j^{-1}\right)\mathbf{x}$$

$$-\tfrac{1}{2}\boldsymbol{\mu}_i^T\mathbf{C}_i^{-1}\boldsymbol{\mu}_i + \tfrac{1}{2}\boldsymbol{\mu}_j^T\mathbf{C}_j^{-1}\boldsymbol{\mu}_j - \tfrac{1}{2}\ln\frac{\left|\mathbf{C}_i\right|}{\left|\mathbf{C}_j\right|} + \ln\frac{\Pr\left(\omega_i\right)}{\Pr\left(\omega_j\right)} = 0$$

$$\boxed{h(\mathbf{x}) = h_i(\mathbf{x}) - h_j(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{v}^T\mathbf{x} + e = 0}$$

Gaussian classes are not the only ones for which linear discriminants are optimal. Let us take the case of discrete random variables...

- Components of $\mathbf{x}$, are discrete

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

- For c=2 categories and vectors of dimension d

$$p_i = \Pr\left(x_i = 1 \middle| \omega_1\right) = 1 - \Pr\left(x_i = 0 \middle| \omega_1\right)$$

$$q_i = \Pr\left(x_i = 1 \middle| \omega_2\right) = 1 - \Pr\left(x_i = 0 \middle| \omega_2\right)$$

- If features are statistically independent

$$f\left(\mathbf{x} \middle| \omega_1\right) = \prod_{i=1}^{d} \Pr\left(x_i \middle| \omega_1\right) = \prod_{i=1}^{d} \left(p_i\right)^{x_i} \left(1 - p_i\right)^{1-x_i}$$

$$f\left(\mathbf{x} \middle| \omega_2\right) = \prod_{i=1}^{d} \Pr\left(x_i \middle| \omega_2\right) = \prod_{i=1}^{d} \left(q_i\right)^{x_i} \left(1 - q_i\right)^{1-x_i}$$

Exercise. Evaluate the discriminant and prove linearity in x_i

$$h(\mathbf{x}) \equiv \ln \Pr(\omega_1 | \mathbf{x}) - \ln \Pr(\omega_2 | \mathbf{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

$$\Downarrow$$

$$h(\mathbf{x}) = \sum_{i=1}^{d} \left(x_i \ln \tfrac{p_i}{q_i} + (1 - x_i) \ln \tfrac{1-p_i}{1-q_i} \right) + \ln \frac{\Pr(\omega_1)}{\Pr(\omega_2)}$$

1.7 PERFORMANCE INDICATORS

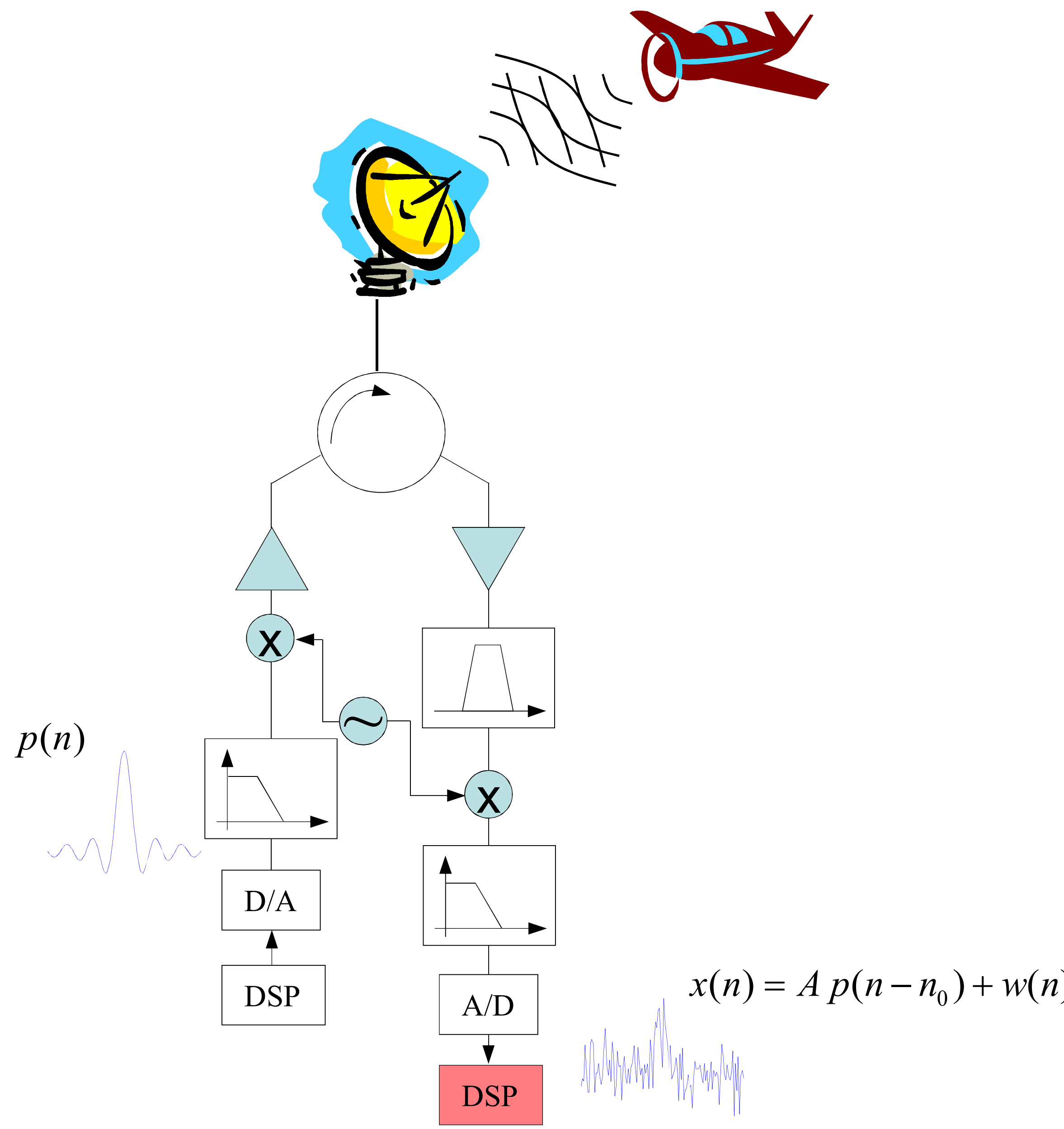
- Assume scalar values for $\mathbf{x}$ and two classes ($d = 1,\ c = 2$)

$$f_x(x|\omega_1) \sim N(\mu_1, \sigma^2)$$
$$f_x(x|\omega_2) \sim N(\mu_2, \sigma^2)$$

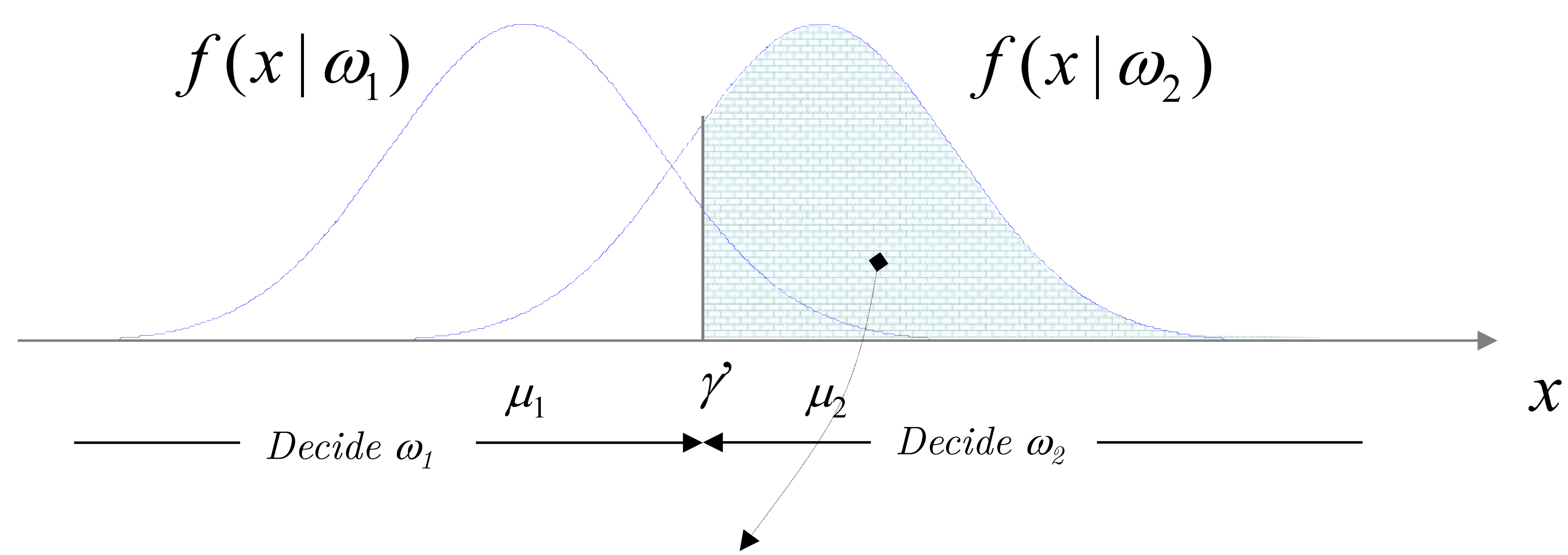
- The classifier uses a threshold γ' that defines decision regions

- The performance of the classifier can be measured in terms of 4 probabilities (in radar wording):

  - Detection $\Pr(x \geq \gamma' | \omega_2)$ Sensitivity of the detector
  - False alarm $\Pr(x \geq \gamma' | \omega_1)$
  - Miss $\Pr(x \leq \gamma' | \omega_2)$
  - Correct reject $\Pr(x \leq \gamma' | \omega_1)$ Specificity (or power) of the classifier

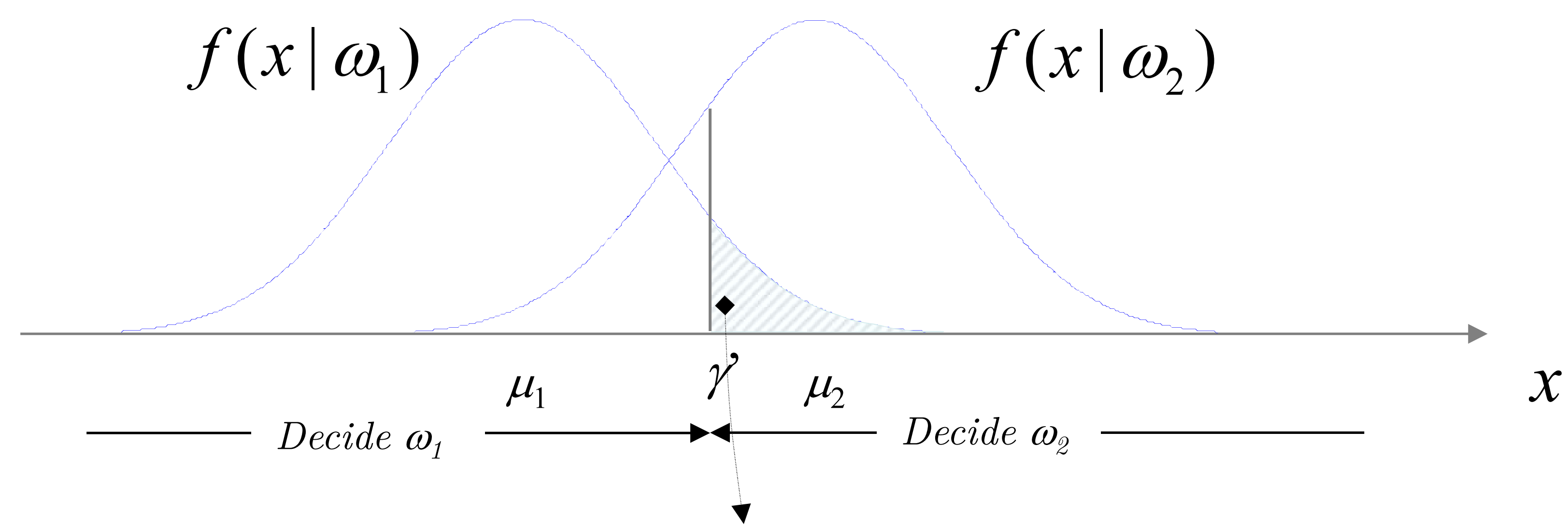

$p(n)$
D/A
DSP
A/D
DSP
$x(n) = A\,p(n-n_0) + w(n)$

Probability of detection (sensitivity), for Gaussian model:



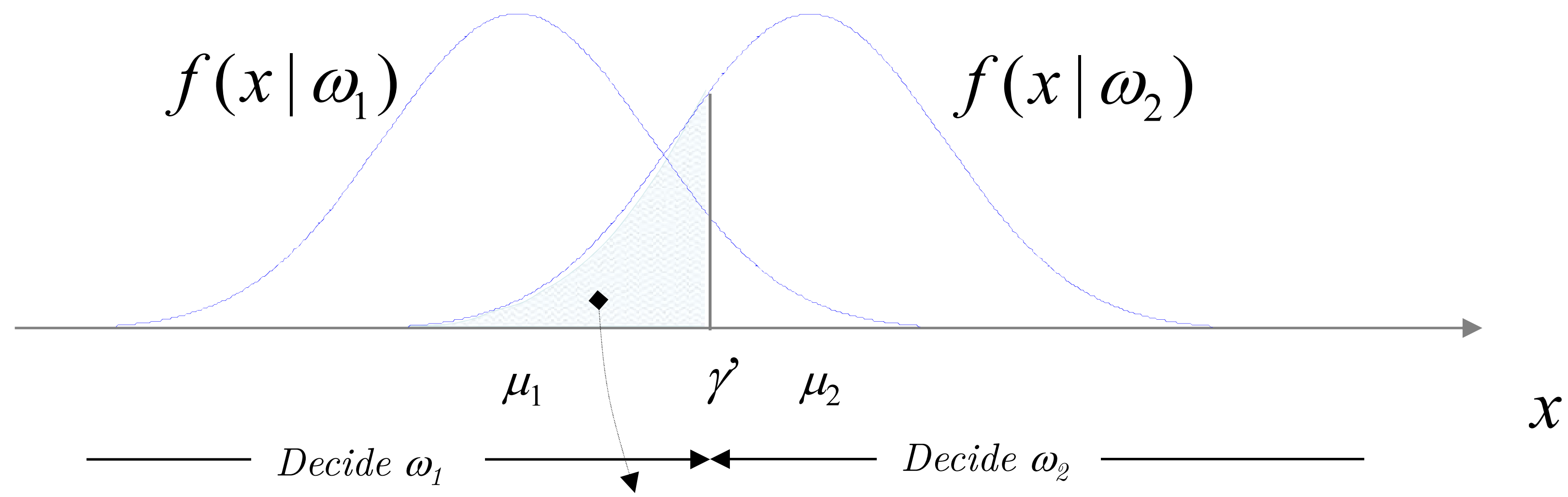
$$P_D = \Pr\left(x \geq \gamma' | \omega_2\right) = \int_{\gamma'}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_2^{\,2}}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^{\,2}}\right) dx$$

Probability of false alarm, for Gaussian model :



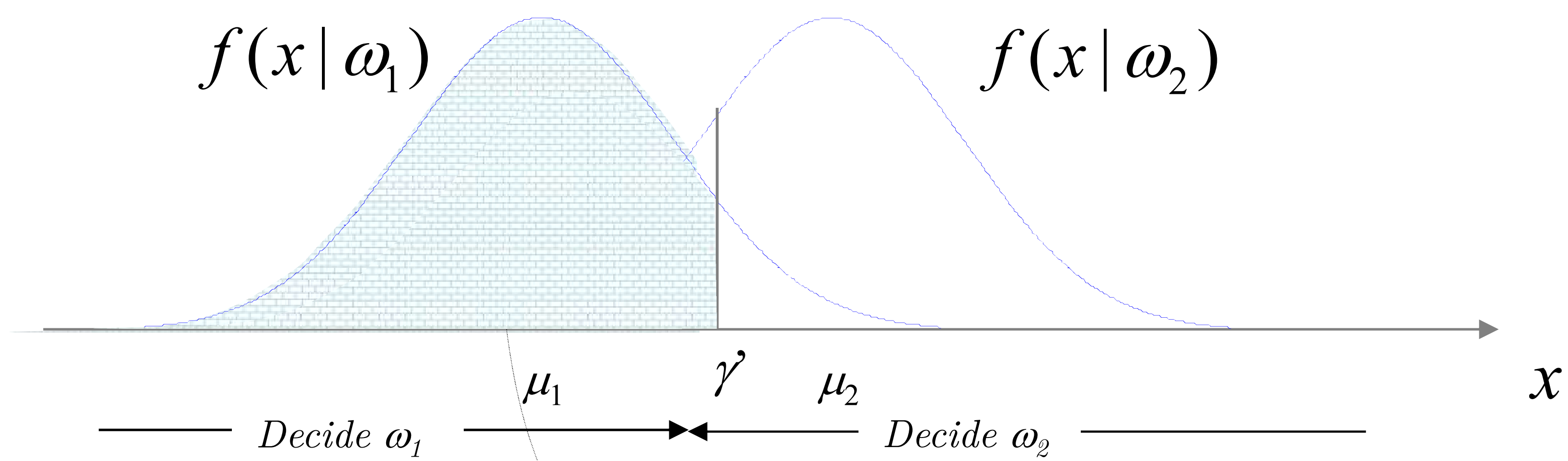
$$P_{FA} = \Pr\left(x \geq \gamma' \,|\, \omega_1\right) = \int_{\gamma'}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_1^{\,2}}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^{\,2}}\right) dx$$

Probability of miss, for Gaussian model:



$$P_P = \Pr\left(x \le \gamma' | \omega_2\right) = \int_{-\infty}^{\gamma'} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right) dx$$

Probability of correct reject (specificity), for Gaussian model:



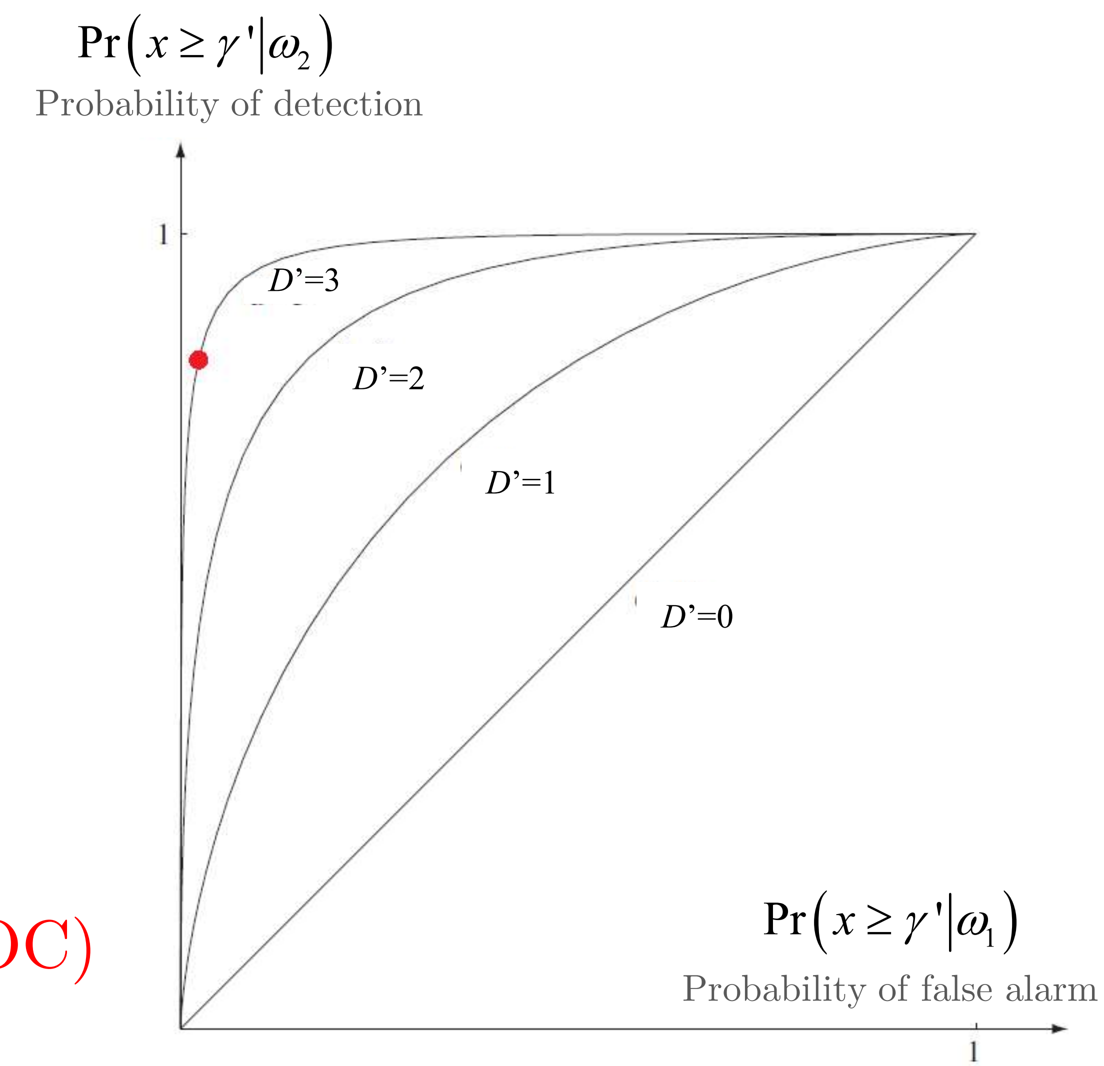
$$P_R = \Pr\left(x \le \gamma' \mid \omega_1\right) = \int_{-\infty}^{\gamma'} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) dx$$

The receiver operating characteristic (ROC)

- ROC represents probability of detection vs. probability of false alarm.
- It depends on the discriminability among clases.
- A measure of discriminability, if variances for each class are equal:

$$D' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

- The area under the ROC (AUROC) is often used as a measure of discriminability (or classifer performance).



- For more than two clases ($c > 2$), for a given probability of detection there are several possible values for the false alarm probability.

- A simple extension of discriminability for $d > 1$

$$D(\omega_i, \omega_j) = \sum_{k=1}^{d} \left| \frac{(\boldsymbol{\mu}_i)_k}{(\sigma_i)_k} - \frac{(\boldsymbol{\mu}_j)_k}{(\sigma_j)_k} \right|$$

- Discriminability based on the Mahalanobis distance:

$$D_M(\mathbf{x} \in \omega_i, \omega_j) = \sum_{\mathbf{x} \in \omega_i} (\mathbf{x} - \boldsymbol{\mu}_j) \mathbf{C}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)$$

Error metrics in more common wording...

True positive: Sick people correctly diagnosed as sick $\Pr\left(\mathbf{x} \in R_2 \middle| \omega_2\right)$

False positive: Healthy people incorrectly identified as sick $\Pr\left(\mathbf{x} \in R_2 \middle| \omega_1\right)$

True negative: Healthy people correctly identified as healthy $\Pr\left(\mathbf{x} \in R_1 \middle| \omega_1\right)$

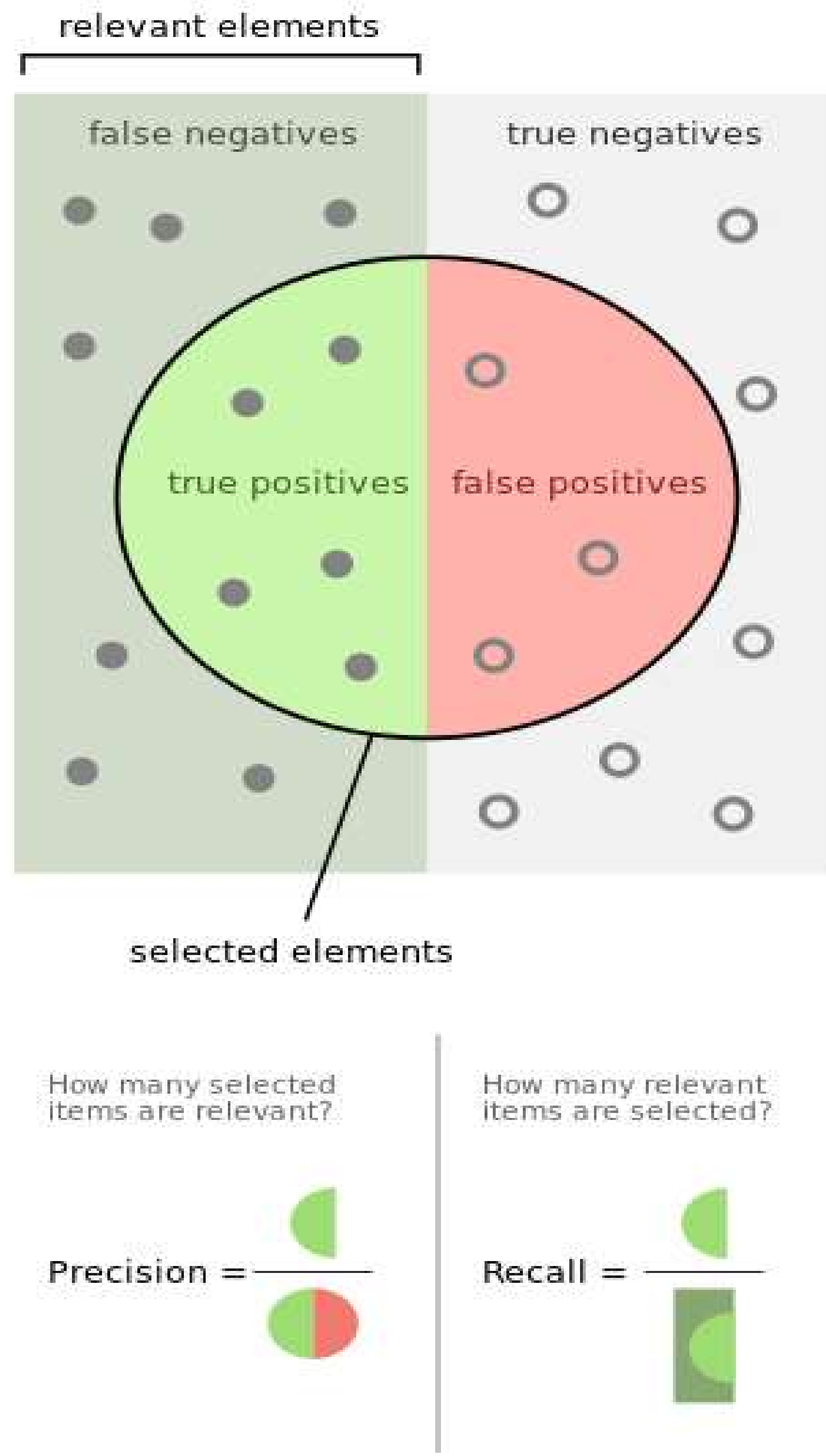
False negative: Sick people incorrectly identified as healthy $\Pr\left(\mathbf{x} \in R_1 \middle| \omega_2\right)$

- **Sensitivity** **(the true positive probability, or "recall", *R*)**: Probability that a sick patient be diagnosed as sick (true positive rate)
- **Specificity**: **(the true negative probability, *S*)**: Probability that a non-infected patient be diagnosed as healthy (true negative rate)
- **Precision** **(also "positive predictive value", or *P*)**: Out of all patients decided sick, what fraction is actually sick? $\Pr\left(\omega_2 \middle| \mathbf{x} \in R_2\right)$
- **F-score**: It measures the test accuracy considering both the precision and the recall

$$F_{score} = \frac{2P \cdot R}{P + R}$$

- **Accuracy**: fraction of total correct decisions for all classes.

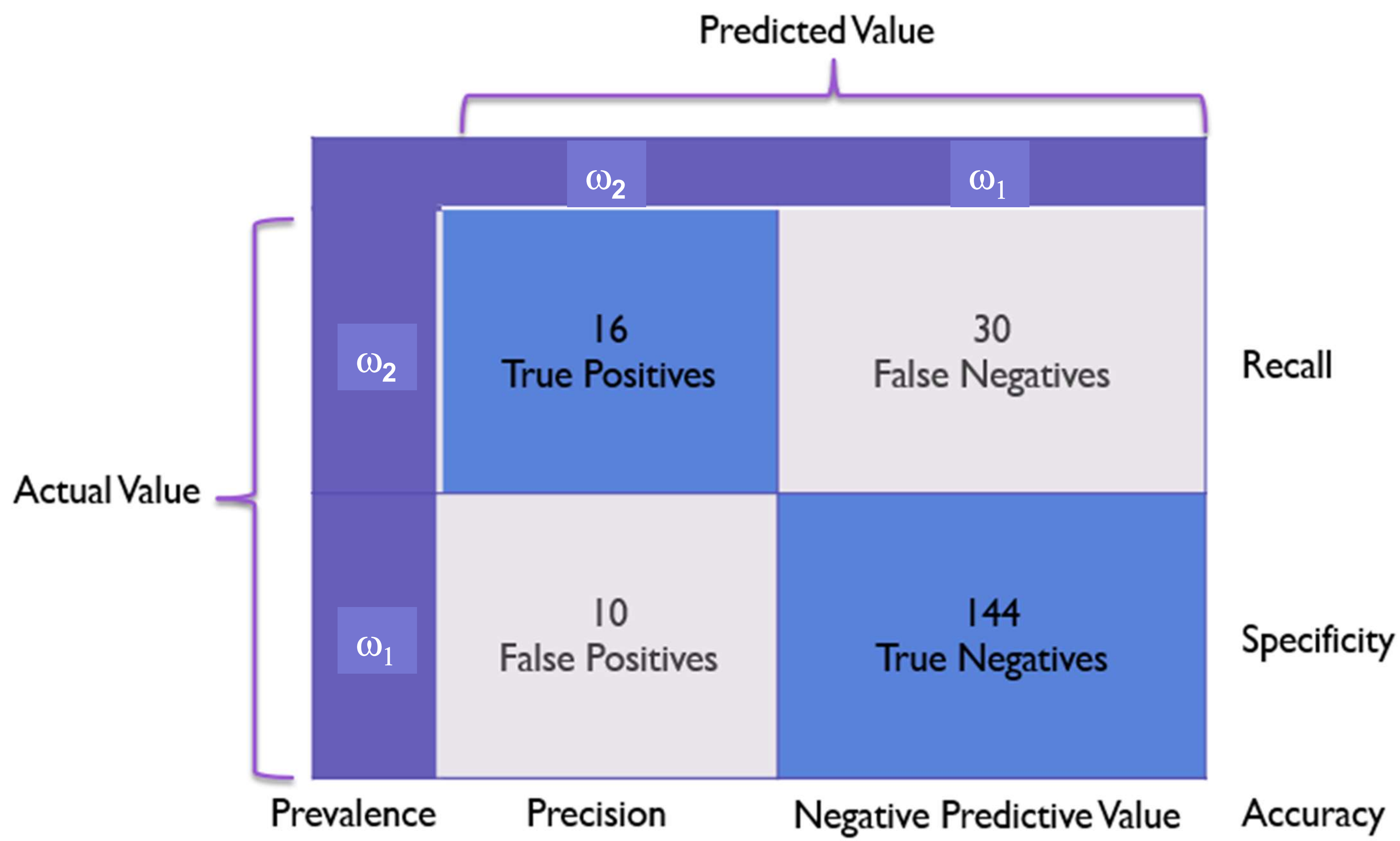
$$A = \Pr\left(\omega_1\right) S + \Pr\left(\omega_2\right) R$$


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many selected items are relevant?
How many relevant items are selected?
Precision =
Recall =

Different combinations of precision (P) and recall (R) have the following meanings:

- **high recall + high precision** : class ω_2 is perfectly handled by the model
- **low recall + high precision** : the model cannot detect class ω_2 well but is highly trustable when it does
- **high recall + low precision** : class ω_2 is well detected but the model also include points of classes ω_1 in it
- **low recall + low precision** : class ω_2 is poorly handled by the model

The confusion matrix



Sensitivity

Precision

It is crucial to distinguish between $\Pr\left(\mathbf{x} \in R_2 \middle| \omega_2\right)$ and $\Pr\left(\omega_2 \middle| \mathbf{x} \in R_2\right)$

and between $\Pr\left(\mathbf{x} \in R_1 \middle| \omega_1\right)$ and $\Pr\left(\omega_1 \middle| \mathbf{x} \in R_1\right)$

Specificity

How are they related?

When getting the result of a medical test on a disease, what can be said about the patient's condition? Take as random variables:

Patient condition: $\omega \in$ {healthy, sick}
Result of the test: $y \in \{-,+\}$

and the prevalence of the disease: Pr(sick)

Use Bayes' theorem to determine the chances you are sick given that the result of the test is positive:

Prob. of getting a positive test if you are sick

Prevalence

$$\Pr(sick \mid +) = \frac{\Pr(+ \mid sick)\Pr(sick)}{\Pr(+ \mid healthy)\Pr(healthy) + \Pr(+ \mid sick)\Pr(sick)}$$

Clinical trials at the pharmaceutical lab...

ω : healthy

TEST

y : – – – – – + – – – –

$\Pr(- \mid \textit{healthy})$

Specificity

ω : sick

TEST

y : + – + + + + + + + –

$\Pr(+ \mid \textit{sick})$

Sensitivity

Me at the doctor's...

TEST

y : +

$\Pr(\textit{sick} \mid +)$?

Get values of specificity and sensitivity for AIDS test from this link, and values of prevalence from AIDS prevalence data in Spain.

Using Bayes' theorem

$$\Pr(sick|+) = \frac{\overbrace{\Pr(+|sick)}^{0.997}\overbrace{\Pr(sick)}^{0.001}}{\underbrace{\Pr(+|healthy)\Pr(healthy)+\Pr(+|sick)\Pr(sick)}_{(1-0.985)\times 0.999+0.997\times 0.001}} = 0.062$$

Chances are very low!

What are the chances if you belong to a risk population from AIDS prevalence data in Spain?

Now, what are the chances of being sick if a second independent test is positive too, $+_2$? Let us apply Bayes' theorem again...

… but note that the prior is now $\Pr(sick|+_1) = 0.062$, as obtained from the first test:

$$\Pr\left(sick\,|+_1,+_2\right) = \frac{\Pr\left(+_2\,|sick,+_1\right)\Pr\left(sick\,|+_1\right)}{\Pr\left(+_2\,|healthy,+_1\right)\Pr\left(healthy\,|+_1\right)+\Pr\left(+_2\,|sick,+_1\right)\Pr\left(sick\,|+_1\right)} =$$

$$= \frac{\Pr\left(+_2\,|sick\right)\Pr\left(sick\,|+_1\right)}{\Pr\left(+_2\,|healthy\right)\Pr\left(healthy\,|+_1\right)+\Pr\left(+_2\,|sick\right)\Pr\left(sick\,|+_1\right)} = 0.82$$

Now the probability is much higher!

1.8 CONCLUSIONS

1. Classification is done by maximising a discriminant function:

$$\hat{\omega}_i = \max_i \{ g_i(\mathbf{x}) \} \qquad i = 1, \dots, c$$

   In the MAP case $g_i(\mathbf{x}) = \Pr(\omega_i | \mathbf{x})$

2. Linear boundaries are only optimal in some specific cases.

3. Clusters of vectors in the d-dimensional space identify classes. The shape of clusters depend on the eigenvalues of the covariance matrix.

4. The ROC is useful to measure the behaviour of the classifier. It can be used to compute the optimal threshold in a experimental way, from a labeled data base. The area under the ROC and Mahalanobis distance provide a measure of the discriminability of our problem.

CONTENTS

2.1 INTRODUCTION

Bayesian decision requires precise knowledge of $f_{\mathbf{x}}(\mathbf{x}\,|\omega_i)$ and $\Pr(\omega_i)$. The computation of these magnitudes require:

- Having a previously-labeled reliable data base (train data base).
- Having an estimator of the pdf and the a priori probabilities.

The estimation of $f_{\mathbf{x}}(\mathbf{x}\,|\omega_i)$ requires many data unless we can use of function that depends on a few parameters $\boldsymbol{\theta}_i$.

Gaussian case: $\boldsymbol{\theta}_i$ contains the mean and the covariance matrix

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

We have two possibilites:

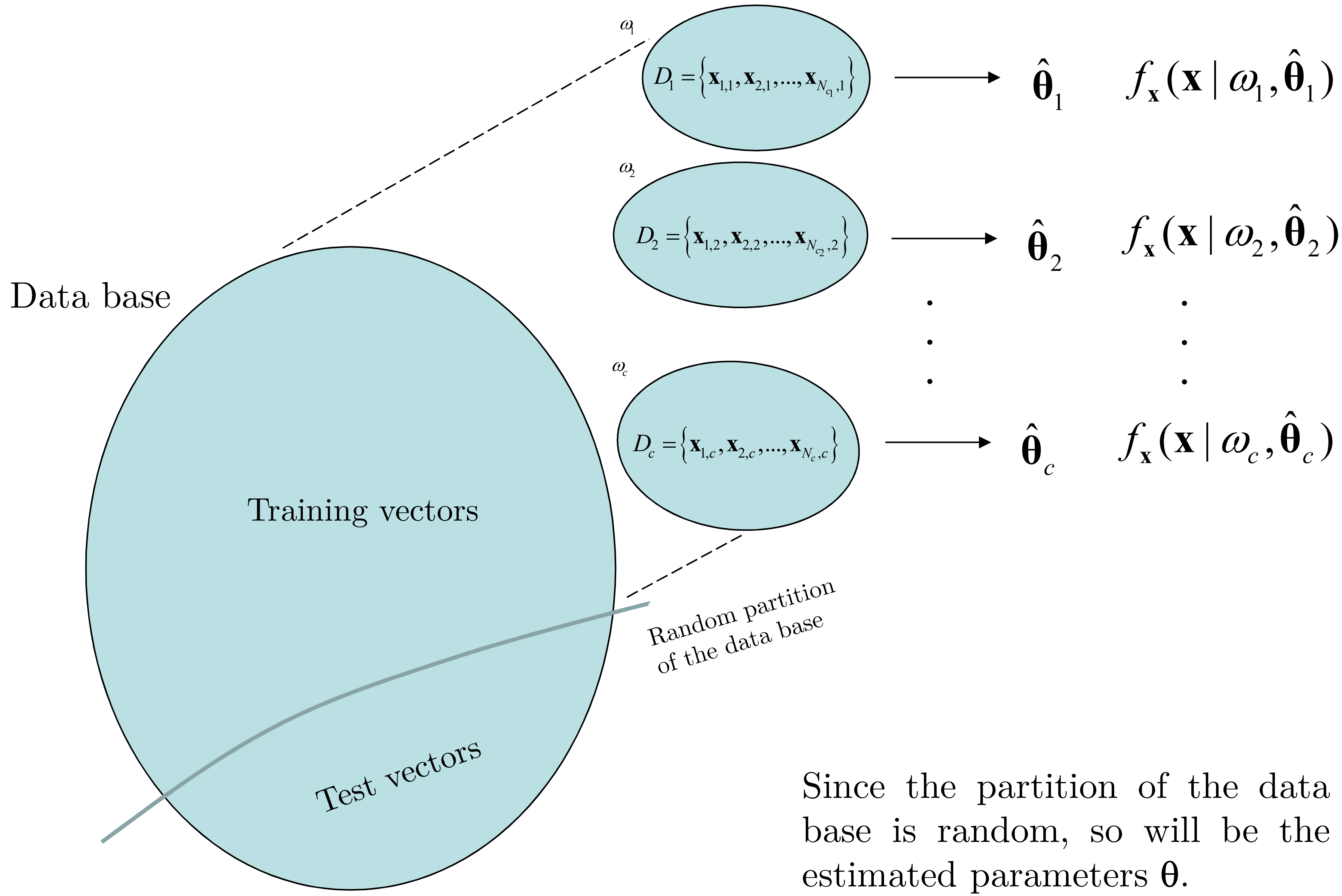
1. Maximum likelihood estimation (ML). The parameters are considered deterministic (although unknown).

2. Bayesian estimation. The parameters are random variables of which an a-priori knowledge is available (related to the concept of "belief") in the form of a pdf.

In all cases, we will assume that a labeled data base is available. Using a partition (the **training data base**) we have to determine $f_{\mathbf{x}}(\mathbf{x}|\omega_i)$.

The rest of vectors will be used to evaluate the performance of the classifier (the **test data base**).

See over...



Since the partition of the data base is random, so will be the estimated parameters $\boldsymbol{\theta}$.

2.2 MAXIMUM LIKELIHOOD ESTIMATION (ML)

Assume that, on each class i, the observed vectors $\mathbf{x}_{k,i} \in D_i$ are statistically independent. The likelihood function is given by:

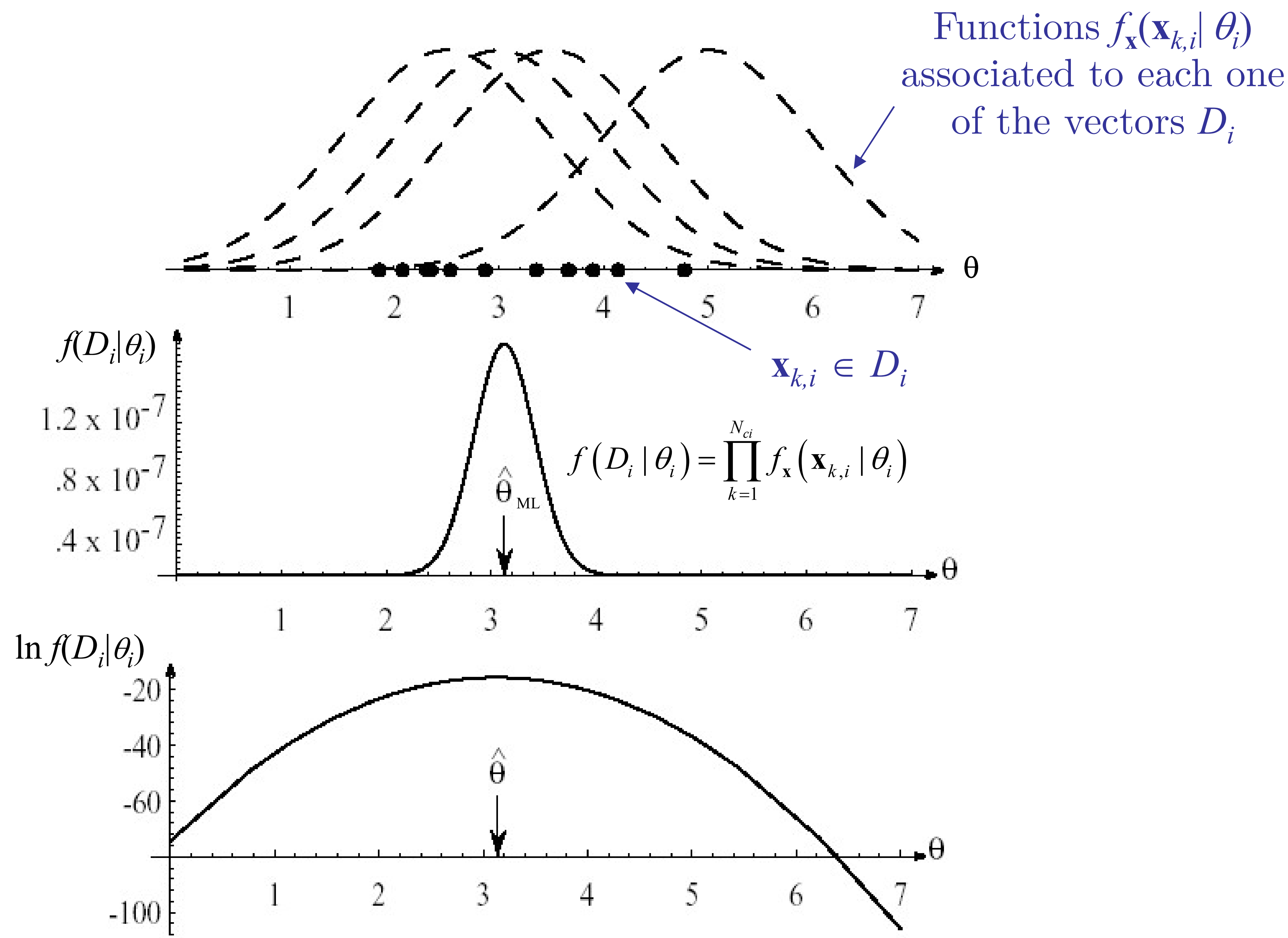
$$f\left(D_i \mid \boldsymbol{\theta}_i\right) = \prod_{k=1}^{N_{c_i}} f_{\mathbf{x}}(\mathbf{x}_{k,i} \mid \boldsymbol{\theta}_i)$$

The ML estimator maximizes this function (or a non-deacresing function of it):

$$\hat{\boldsymbol{\theta}}_{i,ML} = \arg\max_{\boldsymbol{\theta}_i} f\left(D_i \mid \boldsymbol{\theta}_i\right) = \arg\max_{\boldsymbol{\theta}_i} \ \ln f\left(D_i \mid \boldsymbol{\theta}_i\right)$$

A necessary condition to obtain the estimator is therefore given by:

$$\nabla_{\boldsymbol{\theta}_i} \ln f\left(D_i \mid \boldsymbol{\theta}_i\right) = \mathbf{0}$$



Functions $f_{\mathbf{x}}(\mathbf{x}_{k,i}|\,\theta_i)$ associated to each one of the vectors D_i
θ
1
2
3
4
5
6
7
$\mathbf{x}_{k,i} \in D_i$
$f(D_i|\theta_i)$
1.2 x 10-7
.8 x 10-7
.4 x 10-7
$f(D_i\,|\,\theta_i) = \prod_{k=1}^{N_{ci}} f_{\mathbf{x}}(\mathbf{x}_{k,i}\,|\,\theta_i)$
$\hat{\theta}_{ML}$
θ
1
2
3
4
5
6
7
$\ln f(D_i|\theta_i)$
-20
-40
-60
-100
$\hat{\theta}$
θ
1
2
3
4
5
6
7

Characterization of an estimator

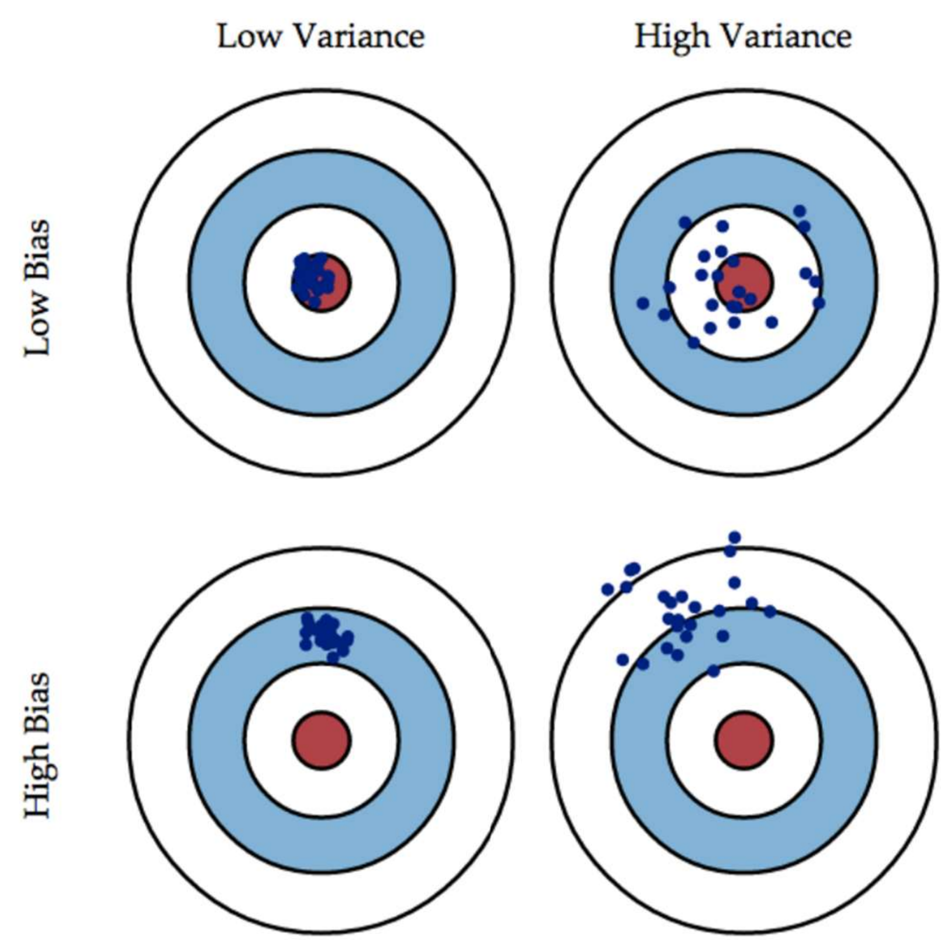
An estimator is a function that applies on an ensambles of feature vectors $\mathbf{x}_{k,i}$ from the training data base. If the selection of vectors is random, so will be the outcomes of the estimator: for each possible partition l of the data base, we obtain a diffent estimate $\hat{\boldsymbol{\theta}}_{l,i}$

1. **Bias**: is the difference between the true value of the parameter and the average of all posible estimates obtained from all posible random partitions of the data base. It measures the sistematic error of the estimator.

$$B\left\{\hat{\boldsymbol{\theta}}_i\right\} = \boldsymbol{\theta} - \frac{1}{L}\sum_{l=1}^{L}\hat{\boldsymbol{\theta}}_{l,i}$$

2. **Variance**: is the deviation of the estimated values from the average value. It measures how the outcome of the estimator depends on a specific partition of the data base. For a scalar parameter:

$$\operatorname{var}\left\{\hat{\theta}_i\right\} = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_{l,i} - \frac{1}{L}\sum_{s=1}^{L}\hat{\theta}_{s,i}\right)^2$$



Properties of the ML estimator:

1. It is asymptotically unbiased (in many cases it is unbiased for low value of N)
2. It is asymptotically efficient (for large N, its variance attains the minimum variance given by the Crámer-Rao bound)

However…

1. It does not necessarily provides the least misclassification error if used in

$$f_{\mathbf{x}}(\mathbf{x} \mid \omega_i, \hat{\boldsymbol{\theta}}_{i,ML})$$

2. If the assumed pdf is far from the real one, the estimations may be of very low quality.

Example 1:

ML estimator of the mean $\boldsymbol{\mu}_i$ if the covariance matrix $\mathbf{C}_i$ is known, in the multivariate Gaussian case. Prove that:

$$\hat{\boldsymbol{\mu}}_{i,ML} = \frac{1}{N_{c_i}} \sum_{k=1}^{N_{c_i}} \mathbf{x}_k$$

Example 2:

ML estimator of both the mean $\boldsymbol{\mu}_i$ and the covariance matrix $\mathbf{C}_i$ in the multivariate Gaussian case. Prove that:

$$\hat{\boldsymbol{\mu}}_{i,ML} = \frac{1}{N_{c_i}} \sum_{k=1}^{N_{c_i}} \mathbf{x}_k \qquad \hat{\mathbf{C}}_{i,ML} = \frac{1}{N_{c_i}} \sum_{k=1}^{N_{c_i}} \left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{i,ML}\right)\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{i,ML}\right)^T$$

Example 3:

ML estimator of the probability p_k for '1' in each component of a binary-valued vector $\mathbf{x} \in \{0,1\}^d$:

$$f_{\mathbf{x}}(D \mid \omega, \mathbf{p}) = \prod_{j=1}^{N_i} \prod_{k=1}^{d} p_k^{\,x_{k,j}} (1 - p_k)^{1 - x_{k,j}}$$

$$\mathbf{p} = \left[p_1, \ldots, p_d \right]$$

2.3 BAYESIAN ESTIMATION

Sometimes we can take advantage of a priori knowledge about the possible values of θ_i. This knowledge will be included in $f(\theta_i)$, which has all properties of a pdf and expresses our "belief" about the possible values of θ_i. Two approaches are possible:

1. Improve the ML estimation of θ_i (using MAP principles)

$$\hat{\boldsymbol{\theta}}_{i,MAP} = \arg\max_{\boldsymbol{\theta}_i} f\left(D_i \mid \boldsymbol{\theta}_i\right) f\left(\boldsymbol{\theta}_i\right) = \arg\max_{\boldsymbol{\theta}_i} \left[\ln f\left(D_i \mid \boldsymbol{\theta}_i\right) + \ln f\left(\boldsymbol{\theta}_i\right)\right]$$

2. Directly estimate the a posteriori probabilities $\mathbf{Pr}(\omega_i|\mathbf{x})$

Computing $f_{\mathbf{x}}(\mathbf{x}|\omega_i)$ and $\mathbf{Pr}(\omega_i)$. This is the recommended procedure in a classification application.

ML AND BAYESIAN ESTIMATION

Comparison:

The function $f(D_i \mid \boldsymbol{\theta}_i)$ will peak around $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$, the more as larger is N_i.

If $f(\theta_i)$ is non-zero and it does not change much around $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$, then

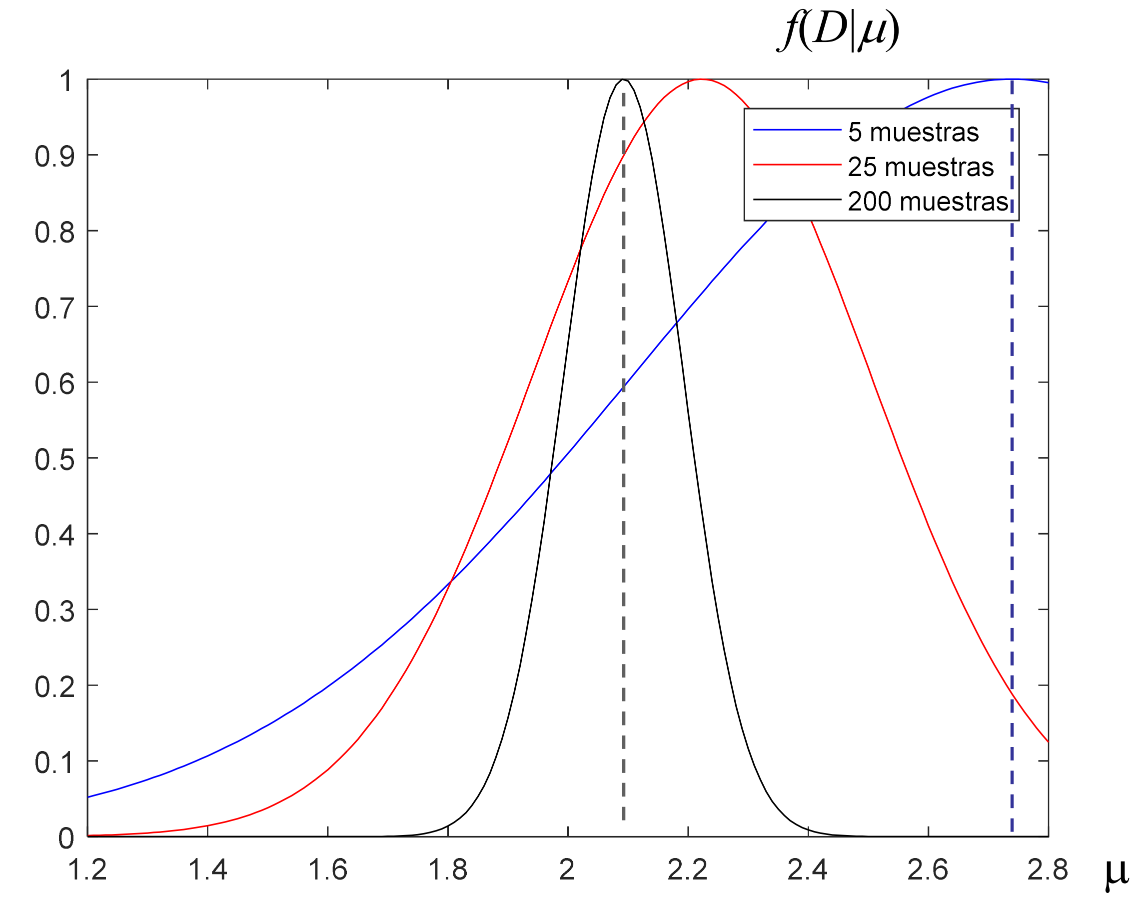
$$f(\boldsymbol{\theta}_i \mid D_i) = \frac{f(D_i \mid \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i)}{f(D_i)}$$

also peaks in $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$ and the estimates obtained through ML and Bayesian principles coincide.

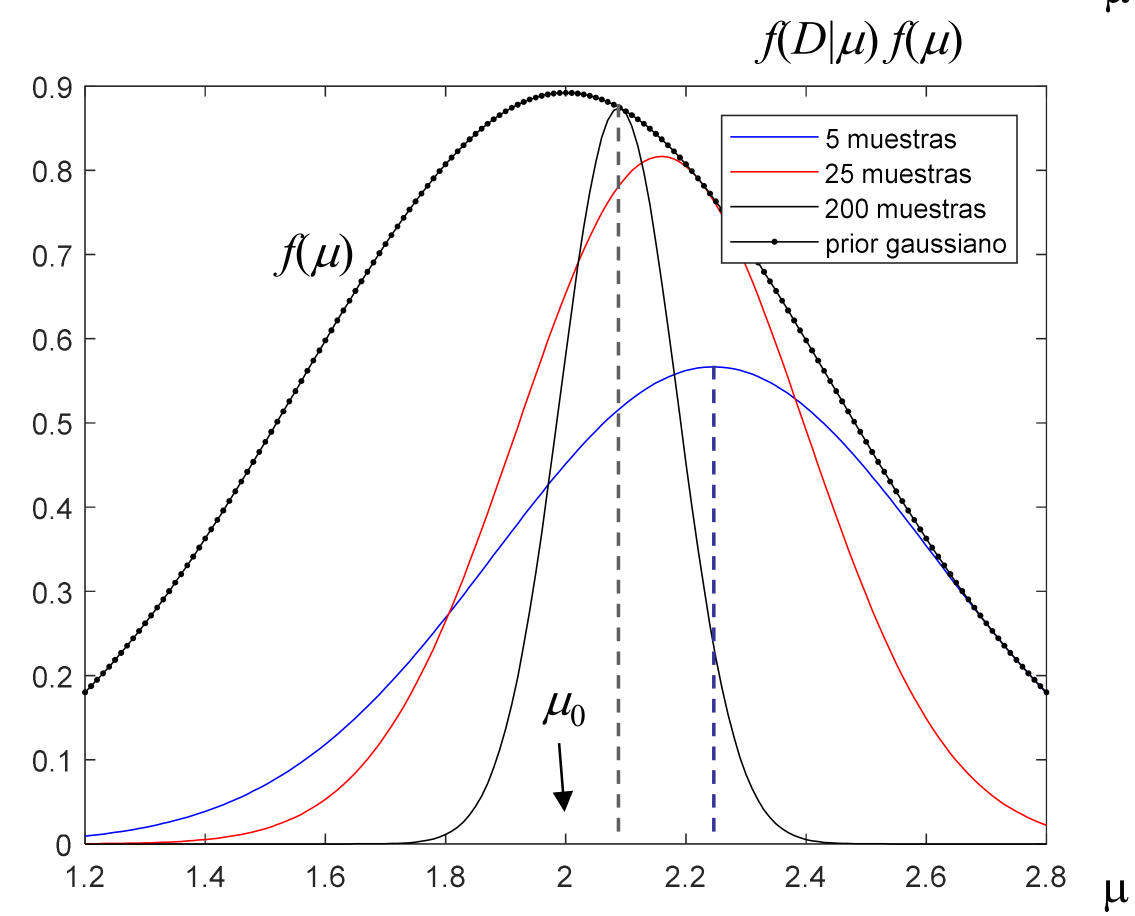
In practice, if the number of vectors in D_i is small, Bayesian estimate is preferred. When many vectors are available, both coincide...

Example 5:

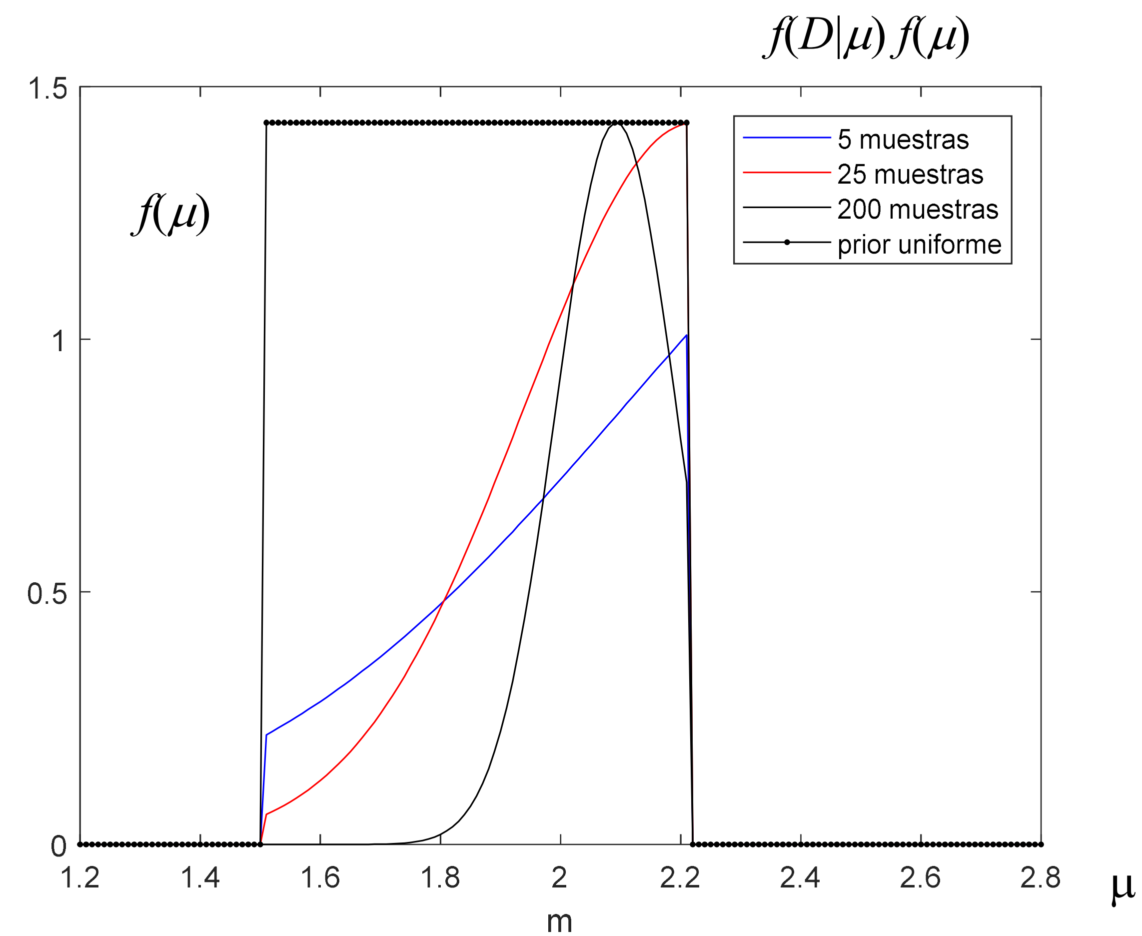
ML estimation of the mean (μ_0=2) obtained from a number of Gaussian simples (likelihood functions have been normalised to the máximum for clarity)



Bayesian estimate of the mean (μ_0=2) obtained from a number of Gaussian samples.
The a priori pdf of μ is Gaussian.



Bayesian estimate of the mean (μ_0=2) obtained from a number of Gaussian samples.
The a priori pdf of μ is uniform.



2. Directly estimate the a posteriori probabilities $\mathbf{Pr}(\omega_i|\mathbf{x})$

Assumptions

- The shape of $f_{\mathbf{x}}(\mathbf{x}|\theta_i)$ is known, but not the parameter θ_i
- Our a priori knowledge of θ_i is in $f(\theta_i)$
- The rest of our knowledge on θ_i is given by data in D_i

Procedure:

1. Average the likelihood function with respect to the a posteriori probability of our parameter:

$$f_{\mathbf{x}}(\mathbf{x}\,|\,\omega_i) \cong f_{\mathbf{x}}(\mathbf{x}\,|\,D_i) = \int f(\mathbf{x}\,|\,\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i\,|\,D_i) d\boldsymbol{\theta}_i$$

2. Compute the a posteriori probability of our parameter as:

$$f(\boldsymbol{\theta}_i\,|\,D_i) = \frac{f(D_i\,|\,\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i)}{\int f(D_i\,|\,\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i} \propto f(D_i\,|\,\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i)$$

3. Assuming independence of data in D_i

$$f(D_i\,|\,\boldsymbol{\theta}_i) = \prod_{k=1}^{N_i} f(\mathbf{x}_{k,i}\,|\,\boldsymbol{\theta}_i)$$

Example 4:

Bayesian estimate of $f_{\mathbf{x}}(\mathbf{x}|D)$ if

$$f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\mu}) = N(\boldsymbol{\mu}, \mathbf{C}) \qquad f(\boldsymbol{\mu}) = N(\boldsymbol{\mu}_0, \mathbf{C}_0)$$

where $\boldsymbol{\mu}_0$, $\mathbf{C}_0$ and $\mathbf{C}$ are known, and feature vectors are available $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \omega$

From **2** and **3** we can write:

$$f(\boldsymbol{\mu}|D) = \alpha \prod_{k=1}^{N} f_{\mathbf{x}}(\mathbf{x}_k \mid \boldsymbol{\mu}) f(\boldsymbol{\mu}) =$$

$$= \alpha' \exp\left[-\frac{1}{2} \left(\boldsymbol{\mu}^T \left(N\mathbf{C}^{-1} + \mathbf{C}_0^{-1} \right) \boldsymbol{\mu} + 2\boldsymbol{\mu}^T \left(\mathbf{C}^{-1} \sum_{k=1}^{N} x_k + \mathbf{C}_0^{-1} \boldsymbol{\mu}_0 \right) \right) \right]$$

This equation can also be written as:

$$f(\boldsymbol{\mu}|D) = \alpha'' \exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)^T \mathbf{C}_N^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)\right]$$

And equating both expressions:

$$\boldsymbol{\mu}^T\left(N\mathbf{C}^{-1}+\mathbf{C}_0^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\mathbf{C}^{-1}\sum_{k=1}^{N}x_k + \mathbf{C}_0^{-1}\boldsymbol{\mu}_0\right) = \boldsymbol{\mu}^T\mathbf{C}_N^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}_N^T\mathbf{C}_N^{-1}\boldsymbol{\mu} + K$$

where the terms that do not depend on $\boldsymbol{\mu}$ and other constants are lumped into K. Comparing the quadratic term in $\boldsymbol{\mu}$:

$$\mathbf{C}_N^{-1} = N\mathbf{C}^{-1} + \mathbf{C}_0^{-1} \tag{1}$$

And comparing the linear terms in $\boldsymbol{\mu}$:

$$\mathbf{C}_N^{-1}\boldsymbol{\mu}_N = \mathbf{C}^{-1}\sum_{k=1}^{N}x_k + \mathbf{C}_0^{-1}\boldsymbol{\mu}_0 \tag{2}$$

From (1) y using the equality: $\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}=\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}$

$$\mathbf{C}_N=\mathbf{C}_0\left(\mathbf{C}+N\mathbf{C}_0\right)^{-1}\mathbf{C} \qquad (3)$$

Having in mind that if $\mathbf{A}$ y $\mathbf{B}$ are invertible

$$\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}=\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}$$

we can use (3) in (2) to obtain

$$\boldsymbol{\mu}_N=\mathbf{C}_0\left(\mathbf{C}_0+\frac{1}{N}\mathbf{C}\right)^{-1}\mathbf{m}_N+\frac{1}{N}\mathbf{C}\left(\mathbf{C}_0+\frac{1}{N}\mathbf{C}\right)^{-1}\boldsymbol{\mu}_0$$

$$\mathbf{m}_N=\frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k$$

Note that the estimated mean is a linear combination of a priori knowledge of the mean $\boldsymbol{\mu}_0$ and the information provided by the data $\mathbf{m}_N$. Integrating equation (1):

$$f_{\mathbf{x}}(\mathbf{x}\,|\,\omega) \cong f_{\mathbf{x}}(\mathbf{x}\,|\,D) = \int f(\mathbf{x}\,|\,\boldsymbol{\mu})\, f(\boldsymbol{\mu}\,|\,D)\; d\boldsymbol{\mu} \sim N\left(\boldsymbol{\mu}_N, \mathbf{C}+\mathbf{C}_N\right)$$

When $N \to \infty$ the estimation of $\boldsymbol{\mu}$ from $f(\boldsymbol{\mu}|D)$ tends to be ML

$$\boldsymbol{\mu}_N = \mathbf{m}_N \qquad\qquad \mathbf{C}_N = \frac{1}{N}\mathbf{C}$$

2.3 CONCLUSIONS

- If we can asume a parametric function for $f_{\mathbf{x}}(\mathbf{x}|\omega_i)$ then the training phase of the classifier reduces to the estimation of the parameters.

- We can use two approaches for the estimation: ML (computationally simpler) or Bayesian (if we have a priori knowledge of parameters)