Pattern Recognition
Elements of Decision Theory

Francesco Tortorella

University of Cassino and Southern Latium

Cassino, Italy

Foundations of PR

- Pattern Recognition lies on two pillars:
  - Probability
  - Decision Theory

Why Probability ?

- A key concept in the field of pattern recognition is that of uncertainty. It arises both through noise on measurements, as well as through the finite size of data sets.
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for pattern recognition.

Why Decision Theory?

- When combined with probability theory, decision theory allows us to make optimal decisions in situations involving uncertainty such as those encountered in pattern recognition

Decision Theory: characteristics

- The goal of the decision theory is to make a quantitative comparison among different classification decisions by using probability arguments and the costs related to the particular decisions
- Basic assumptions:
  - <u>The decision problem is cast in probabilistic terms</u>
  - <u>All the probability functions relevant to the problem are known</u>

Principles

- Let's consider a problem with C classes, with labels ω_j with j=1,2,…,C.
- Let's call α_i i=1,2,…,A the decisions we can take (A<>C ?).
- Initially, let's suppose to know the probability $P(\omega_j)$ that a sample belongs to the class ω_j (*a priori* probability or *prior*).

Principles

- We must take a decision about the class of a sample s (s is described by a feature vector X with size N)
- If we have not any other source of information, the decision rule should be entirely based on the priors $P(\omega_j)$
- Who wins?

Principles

- Let's add some information about the classes
- It is available in the form of the *class-conditional density* or *likelihood* $p(x|\,\omega_j)$, i.e. the probability of having the value x, *when we know* that it belongs to the class ω_j

$$p(x|\,\omega_j)$$

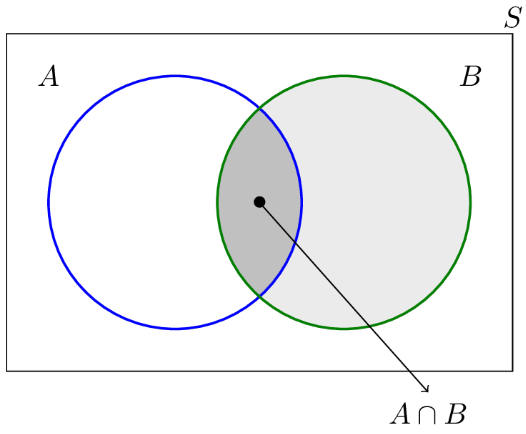
uncertainty

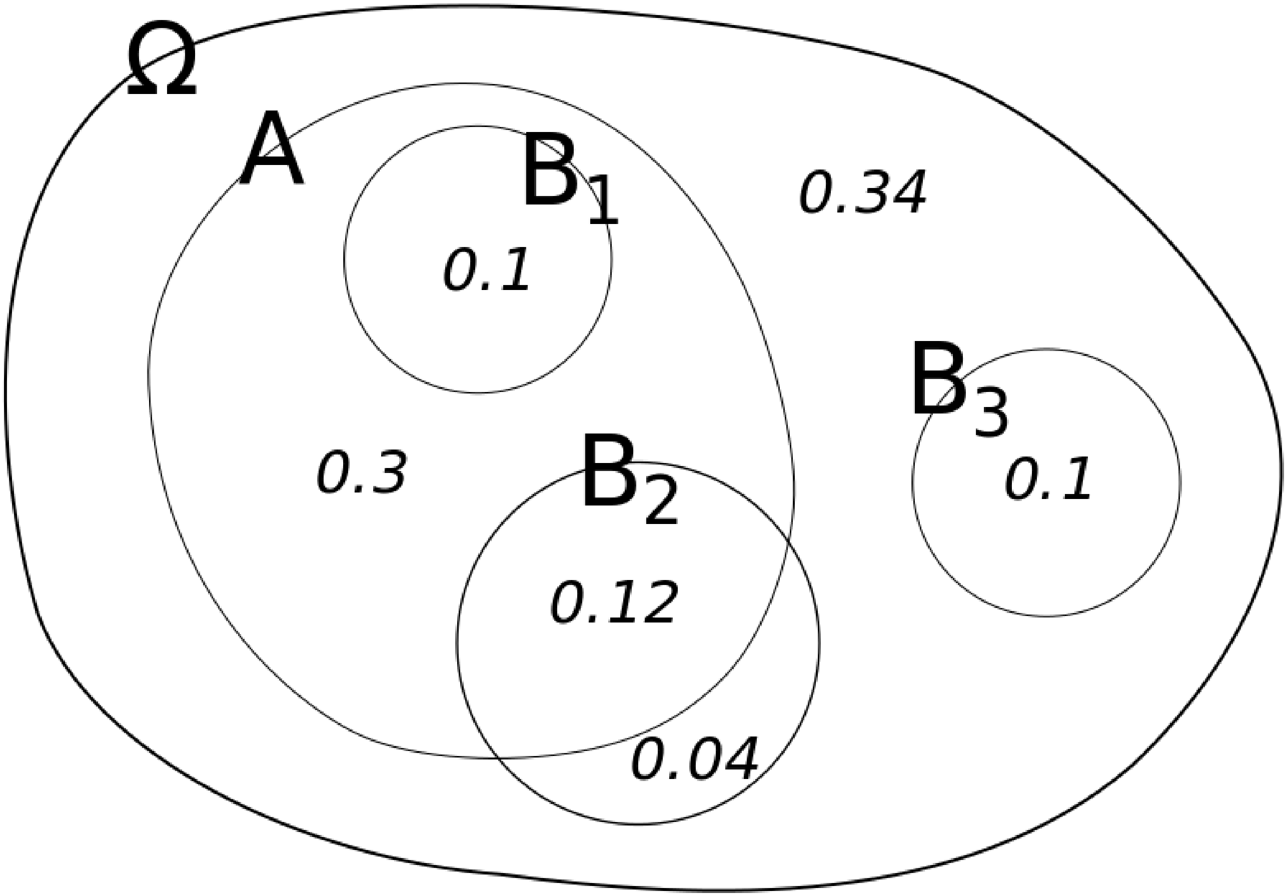
knowledge

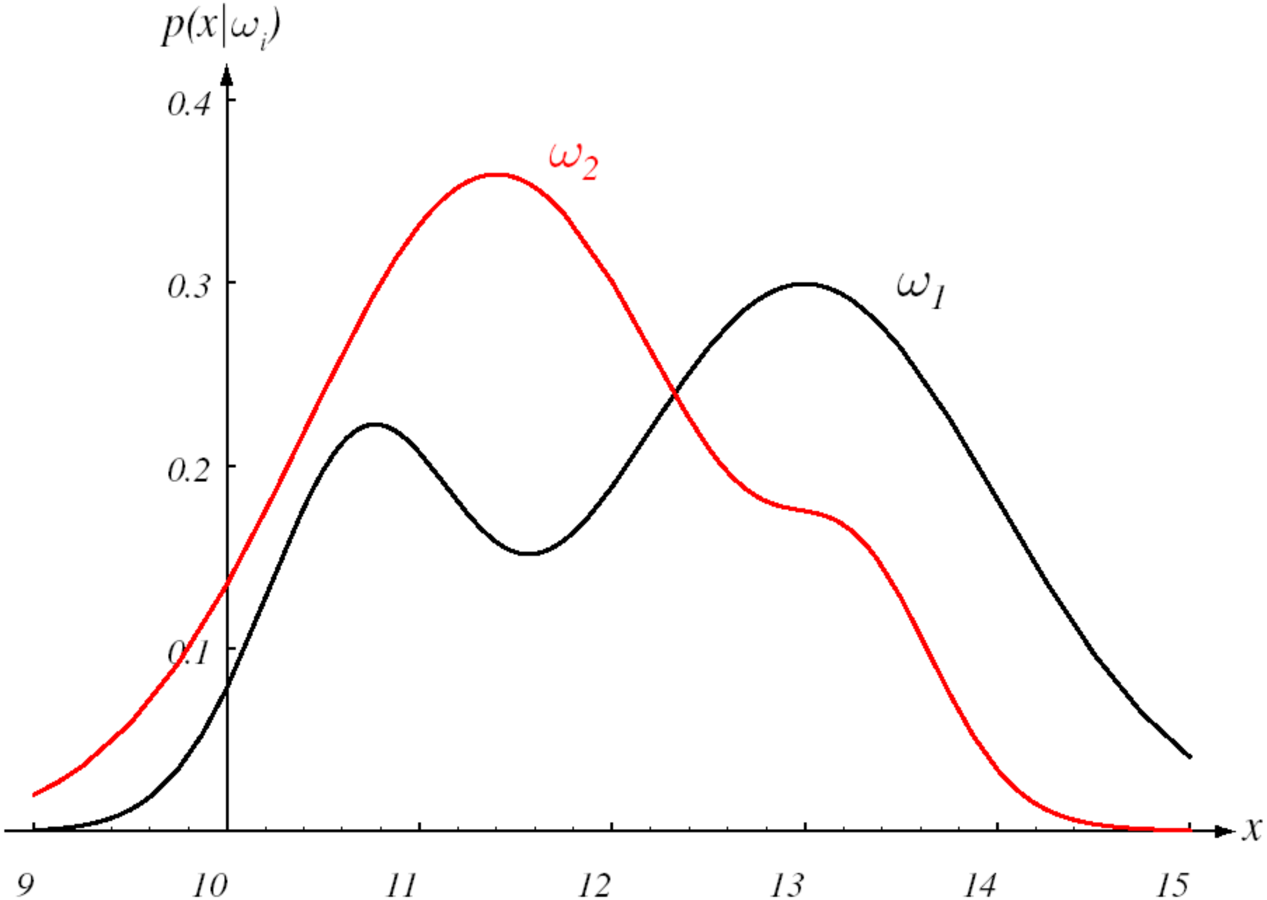
Conditional probability

- P(A|B) is the probability of the event A given that (by assumption, presumption, assertion or evidence) the event B has occurred and is defined as:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$




Ω
A
B1
0.34
0.1
B3
0.3
B2
0.1
0.12
0.04



$p(x|\omega_i)$
0.4
0.3
0.2
0.1
ω_2
ω_1
x
9
10
11
12
13
14
15

What we have?
What we want?

- In our case, we have the values of the feature vector x (knowledge) and we want to decide about the class ω_j it belongs to (uncertainty).
- Thus we have to consider another probability:

$$P(\omega_j|x)$$

uncertainty knowledge

Help, rev. Bayes !!!

Bayes' Theorem

Thanks to Bayes' Theorem we are able to evaluate the probability $P(\omega_j|x)$ that the observed f.v. x was produced by a sample of the class ω_j (*a posteriori* probability) if we know the priors $P(\omega_j)$ and the likelihoods $p(x|\omega_j)$.

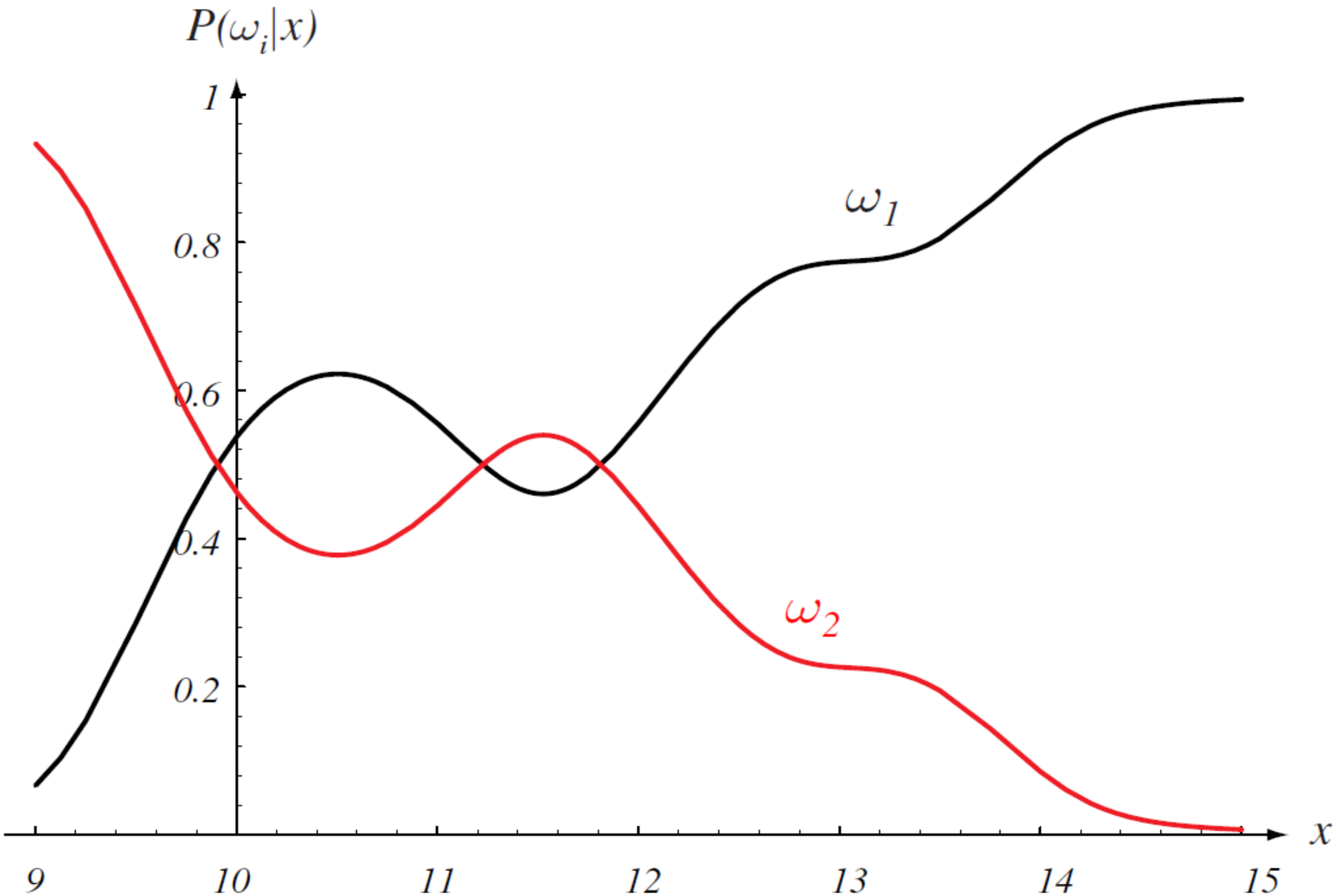
Rev. Thomas Bayes
b. 1702, London
d. 1761, Tunbridge Wells, Kent

Bayes' Theorem

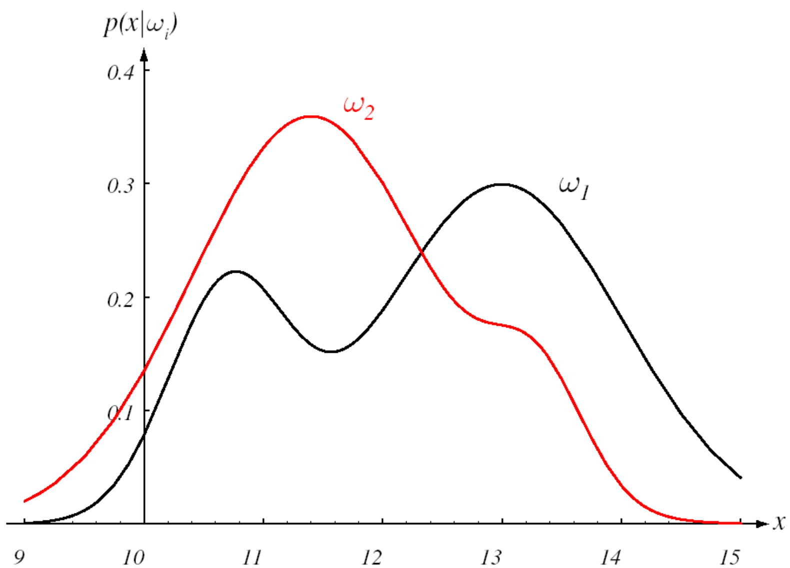
It says that:

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

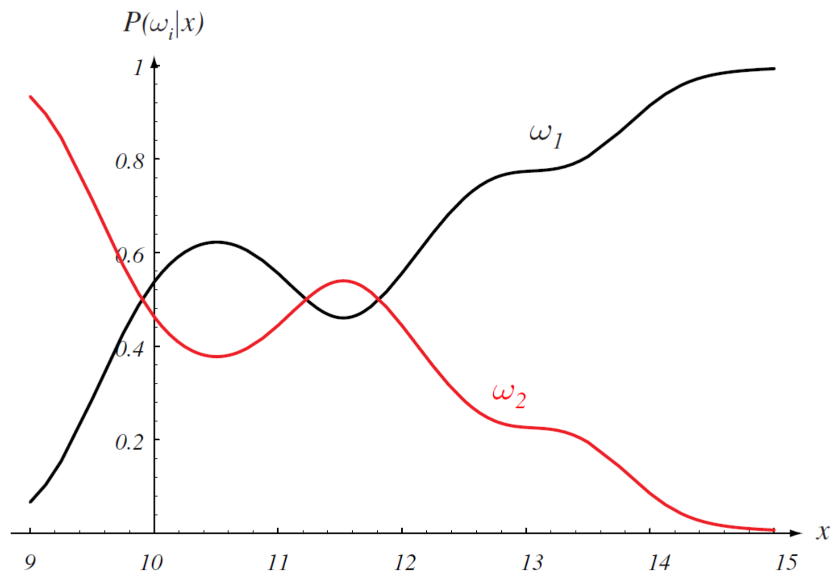
where $p(x) = \sum_{j=1}^{C} p(x|\omega_j) \cdot P(\omega_j)$ is the *unconditional density function* of the f.v. x



Post probabilities when $P(\omega_1)=2/3$ and $P(\omega_2)=1/3$.



p(x|ωi)
0.4
0.3
0.2
0.1
ω2
ω1
9
10
11
12
13
14
15
x



P(ωi|x)
1
0.8
0.6
0.4
0.2
ω1
ω2
9
10
11
12
13
14
15
x

Now, let's decide!

- *Reasonably* the decision is toward the class with the highest post probability:
  Choose ω_1 if $P(\omega_1|x) > P(\omega_2|x)$
  otherwise choose ω_2
- Maximum a Posteriori (MAP) rule
- Actually this rule minimizes the probability of error:

$$P(error|x)=\min\{P(\omega_1|x)\ ,\ P(\omega_2|x)\}$$

Error probability

- Let's consider the sources of error for a classifier.
- In the two-class problem, the classifier has divided the space T into two (possibly nonoptimal) regions R_1 and R_2 ($T=R_1 \cup R_2$).
- Two possible errors:
  - $x \in \omega_1$ but it falls in R_2
  - $x \in \omega_2$ but it falls in R_1

Error probability

- Thus the value of the error probability P_e can be written:

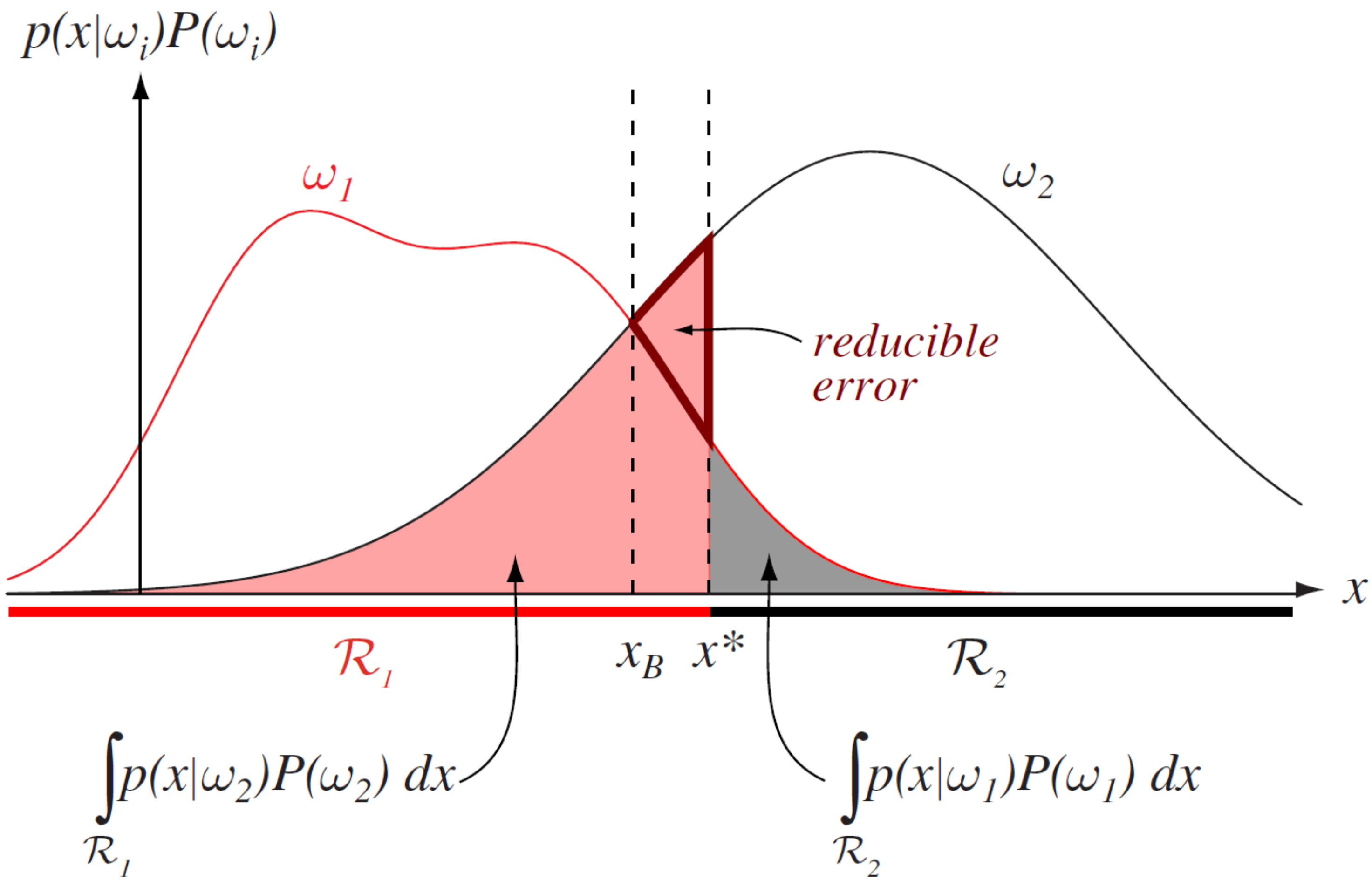
$$\mathrm{P}_e = p(x \in R_2, \omega_1) + p(x \in R_1, \omega_2) =$$

$$p(x \in R_2|\omega_1) \cdot P(\omega_1) + p(x \in R_1|\omega_2) \cdot P(\omega_2) =$$

$$\int_{R_2} p(x|\omega_1)dx \cdot P(\omega_1) + \int_{R_1} p(x|\omega_2)dx \cdot P(\omega_2) =$$

$$\int_{R_2} p(x|\omega_1)P(\omega_1)dx + \int_{R_1} p(x|\omega_2)P(\omega_2)dx$$

Decision regions and errors



Lowest error probability

- The error probability is bounded below:

$$P_e = \int_{R_2} p(x|\omega_1)P(\omega_1)dx + \int_{R_1} p(x|\omega_2)P(\omega_2)dx \geq$$

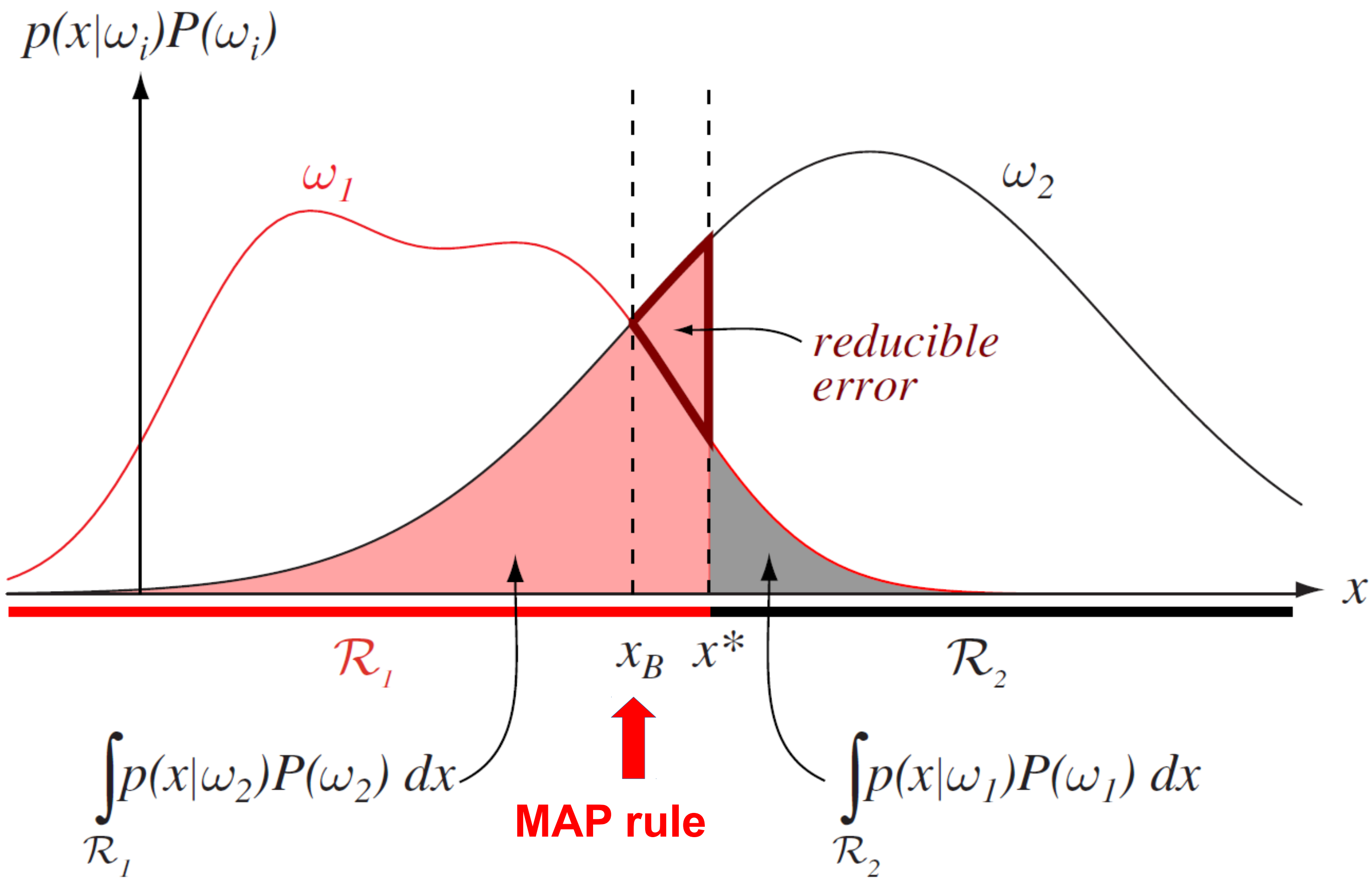
$$\int_{R_2} min\{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\}dx +$$

$$\int_{R_1} min\{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\}dx =$$

$$\int_{T} min\{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\}dx$$

MAP rule

Decision regions and errors



Decision

- From a practical point of view, $p(x)$ does not affect the decision, thus the rule can be written as:

  Choose ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$
  otherwise choose ω_2

- Particular situations:
  - if $p(x|\omega_1) = p(x|\omega_2)$ the knowledge of the value of the f.v. x does not add further information to what we know from the priors
  - if $P(\omega_1) = P(\omega_2)$ the decision is made only on the base of the likelihood

Bayes is the best!

- In the case of multiclass problems the decision rule becomes $\alpha(x) = \text{argmax } \{P(\omega_i|x)\}$.
- Also in this case the decision rule minimizes the error probability
- In summary, the *Maximum A Posteriori* (MAP) rule provides the optimal classifier.

How much this costs?

- Now let's suppose to know some information about the consequences of our decisions.
- This is given by a *loss function* $\lambda(\alpha_i | \omega_j)$ that provides the cost produced by the decision α_i when the sample belongs to the class ω_j.
- The cost of the decision α_i for the sample x is

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

- This is the *conditional risk* or *conditional cost*

Minimum risk decision

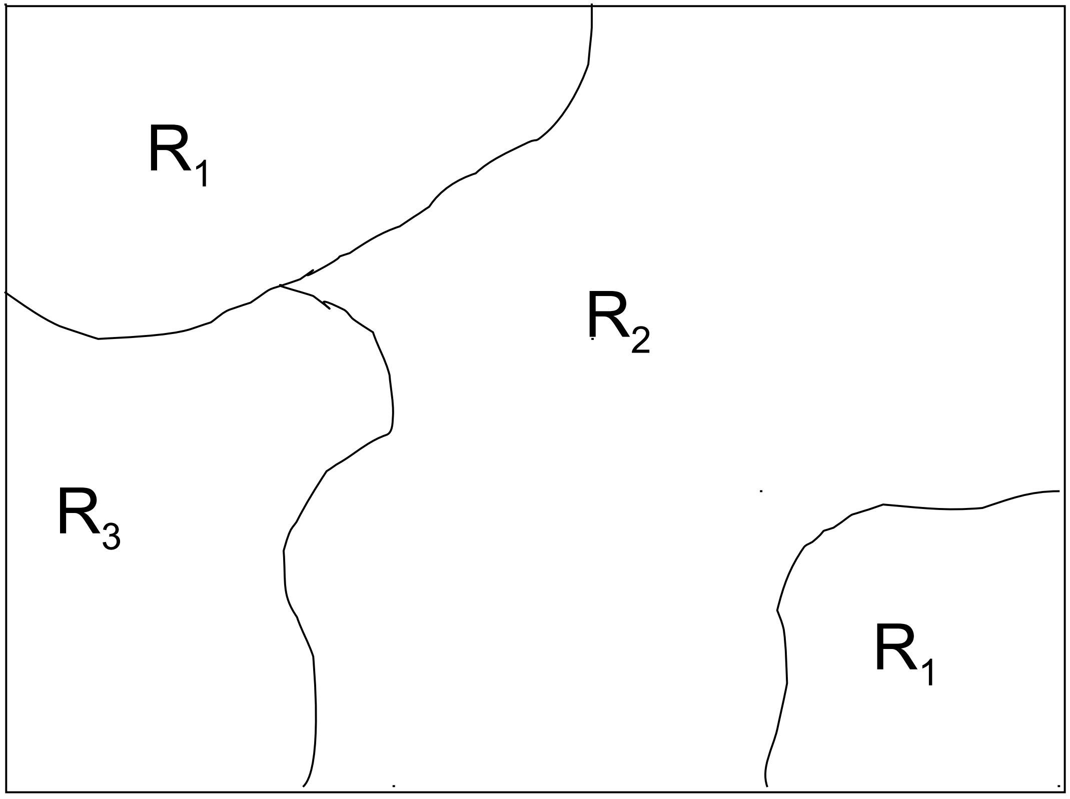
- In this setting the most reasonable decision rule is to minimize the risk
- Thus we have:

$$\alpha(x) = \underset{1 \leq j \leq A}{\operatorname{argmin}} \; \mathrm{R}(\alpha_i | x)$$

Decision regions

The decision rule induces in the feature space a set of decision regions

$x \in R_i \Leftrightarrow \alpha(x) = \alpha_i$



2 class problems

- As a particular case, consider a problem with 2 classes (the worst ones!!) and call α_i the decision for the class ω_i with i=1,2 (A==C)
- If $\lambda_{ij}= \lambda(\alpha_i|\omega_j)$, the conditional risk for the two decisions are:

$$R(\alpha_1|x)= \lambda_{11}P(\omega_1|x)+ \lambda_{12}P(\omega_2|x)$$
$$R(\alpha_2|x)= \lambda_{21}P(\omega_1|x)+ \lambda_{22}P(\omega_2|x)$$

- *Reasonable* values for the costs are such that $\lambda_{ii} < \lambda_{ij}$ with $j \neq i$

2 class problems

- Choose ω_1 if $R(\alpha_1|x) < R(\alpha_2|x)$, i.e. if:

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

  equivalent to:

  $$(\lambda_{11}-\lambda_{21})P(\omega_1|x) \; < \; (\lambda_{22}-\lambda_{12})P(\omega_2|x)$$

- Since $(\lambda_{11}-\lambda_{21})<0$ e $(\lambda_{22}-\lambda_{12})<0$, we can multiply both members by -1 and change the sign in the inequality:

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

  or:

  $$\frac{P(\omega_1|x)}{P(\omega_2|x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$$

2 class problems

- If we recall the Bayes' theorem, the rule can be written:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

  where the first term is the *likelihood ratio*

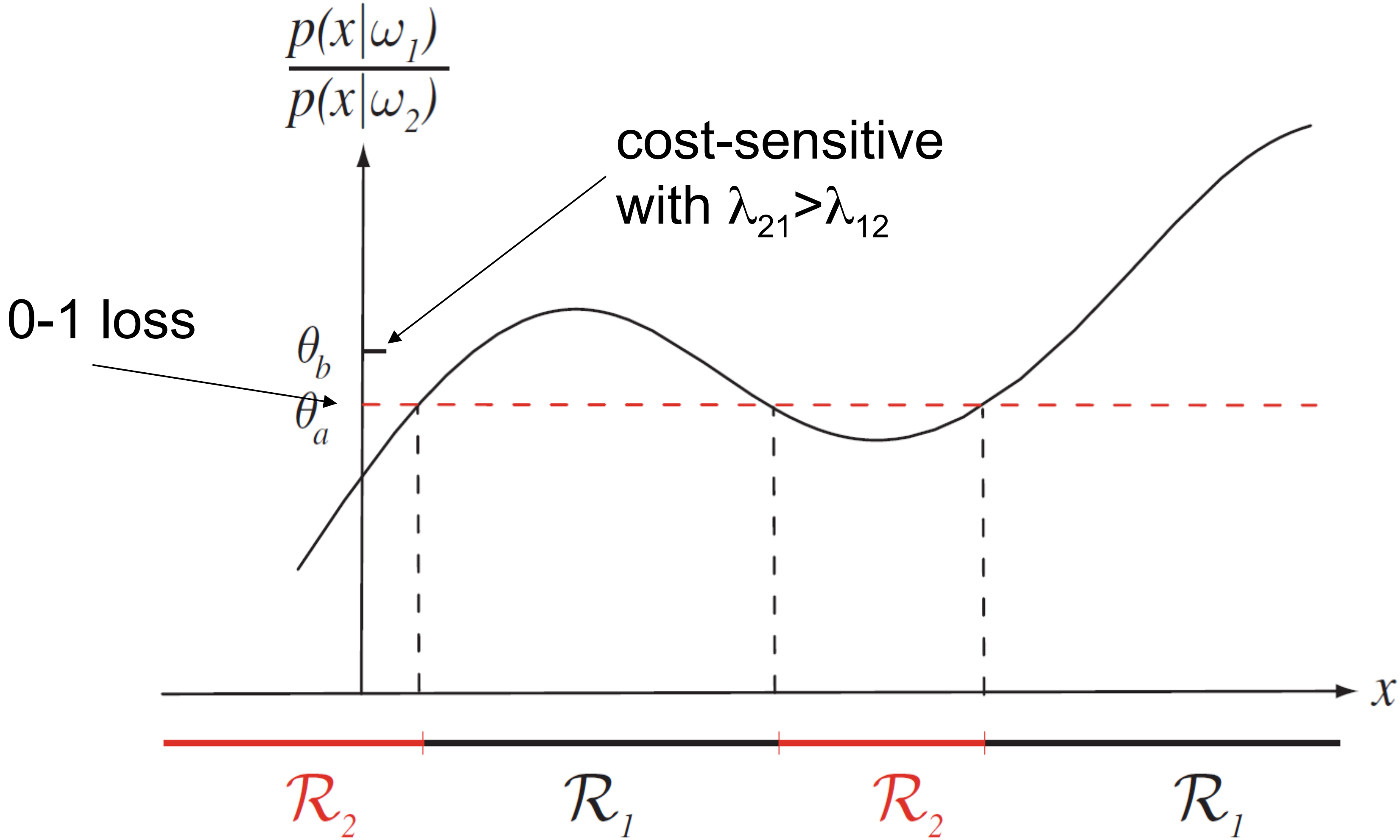
- *Likelihood Ratio Test* (LRT)

2 class problems

- The minimum error decision rule can be derived by the minimum risk rule by assigning $\lambda_{21}=\lambda_{12}=1$ e $\lambda_{11}=\lambda_{22}=0$ (*zero-one loss*).
- The LRT becomes:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

2 class problems



Neyman-Pearson decision rule

- We have seen two decision criteria:
  - Minimum probability of error
  - Minimum risk
- In some cases, instead of minimizing an overall penalty (risk or error), we need to fix a bound on the error on one class while minimizing the error on the other class.
- Example: we want the probability of error ε_2 on the class ω_2 is lower than α and that the probability of error ε_1 on the class ω_1 is minimum.
- This is the *Neyman-Pearson decision rule*

Reject Option

- When using the minimum error rule, there can be cases where the probability of error, although minimum, is too high to be accepted.
- In these cases, it is more convenient abstaining from the decision rather than running the risk of providing a wrong answer.
- In other words, we add another possible decision: the “no decision” (or *reject*)

Reject Option

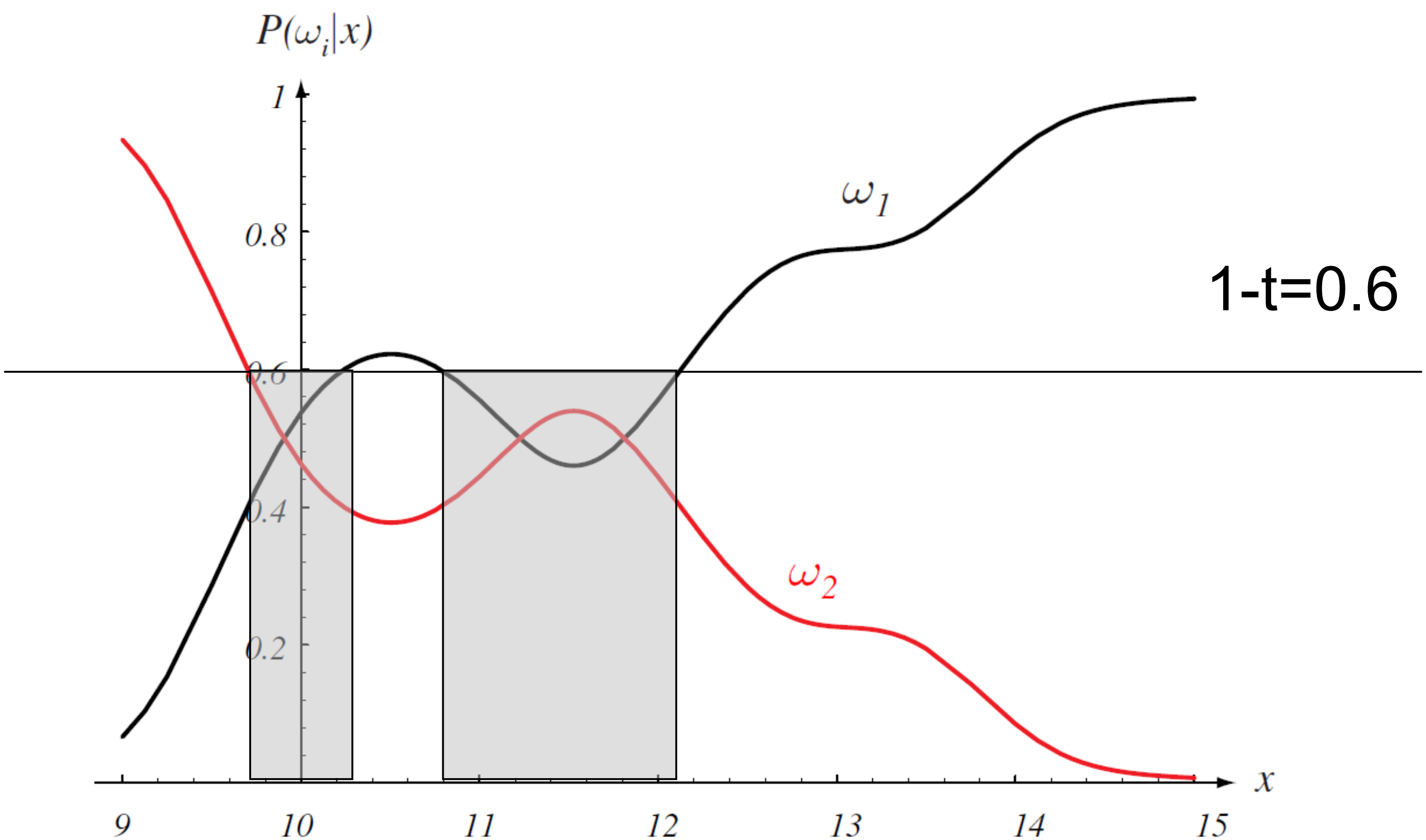
- With the minimum error rule, the probability of error when classifying a sample x is $P_e(x) = 1\text{-}\max\{P(\omega_i|x)\}$.
- Suppose we cannot accept the decision if P_e is higher than a threshold t.

Minimum error decision rule with reject

- The decision rule is now:

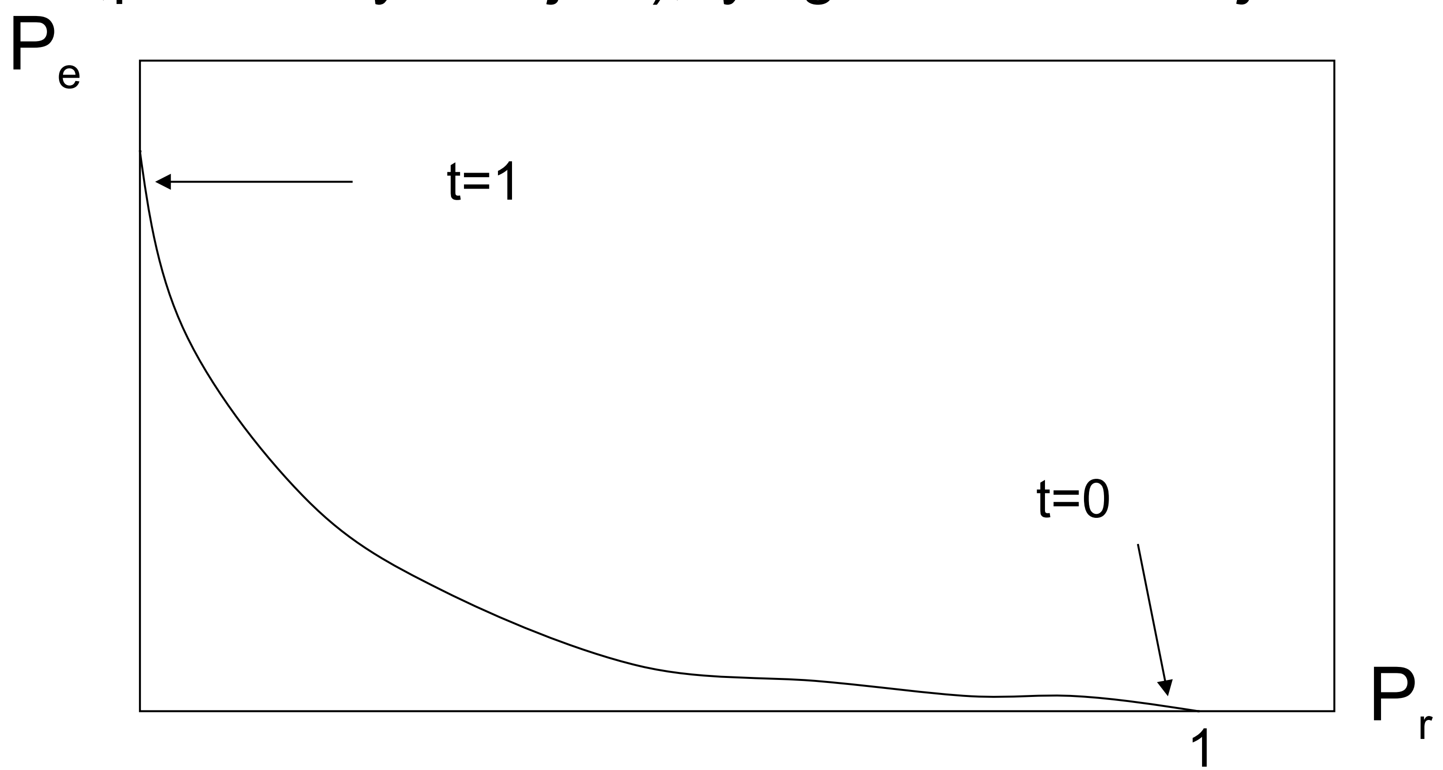
$$\alpha(x) = \begin{cases} \omega_i & \text{if } P(\omega_i|x) > P(\omega_j|x) \;\; \forall i \neq j \;\text{ and} \\ & P(\omega_i|x) > 1\text{-}t \\ \text{'reject'} & \text{otherwise} \end{cases}$$

Reject region



Error/reject curve

When t varies, we obtain different pairs (P_e,P_r) (probability of error,probability of reject), lying on an *error/reject curve*



Minimum risk decision rule with reject (uniform costs)

- The reject option can be applied also in the cost-sensitive setting
- In this case the reject has a proper cost:

$$\lambda_{ij} = \begin{cases} c \text{ if } i=j \\ e \text{ if } i \neq j \\ r \text{ if } i=\text{'reject'} \end{cases}$$

Reasonable costs:
$c < e$
$c < r$
$r < e$

Minimum risk decision rule with reject (uniform costs)

- The conditional risk is:

$$R(\alpha|x) = \begin{cases} r \text{ if } \alpha = \text{`reject'} \\ c\, P(\omega_i|x) + e\,(1 - P(\omega_i|x)) \text{ if } \alpha = \omega_i \end{cases}$$

- Thus the decision rule becomes:

$$\alpha(x) = \begin{cases} \omega_i & \text{if } P(\omega_i|x) > P(\omega_j|x)\ \forall i \neq j \text{ and} \\ & P(\omega_i|x) > (e-r)/(e-c) \\ \text{`reject'} & \text{otherwise} \end{cases}$$

Chow's Rule

Minimum risk decision rule with reject
(2 classes, non-uniform costs)

- Consider now a 2-class cost-sensitive problem with non-uniform costs
- How is the reject option applied?
- Consider the conditional risks:

$$R(\alpha_0) = \lambda_0$$
$$R(\alpha_1) = \lambda_{11} P(\omega_1|\mathbf{x}) + \lambda_{12} P(\omega_2|\mathbf{x})$$
$$R(\alpha_2) = \lambda_{21} P(\omega_1|\mathbf{x}) + \lambda_{22} P(\omega_2|\mathbf{x})$$

Minimum risk decision rule with reject
(2 classes, non-uniform costs)

- We decide for class ω_1 if $R(\alpha_1) = \min\{R(\alpha_1), R(\alpha_2)\}$ and $R(\alpha_1) \le R(\alpha_0)$ that leads to:

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \ge \frac{\lambda_{12} - \lambda_0}{\lambda_0 - \lambda_{11}} \frac{P_2}{P_1}$$

- While for the deciding for class ω_2:

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \le \frac{\lambda_0 - \lambda_{22}}{\lambda_{21} - \lambda_0} \frac{P_2}{P_1}$$

with $R(\alpha_2) = \min\{R(\alpha_1), R(\alpha_2)\}$

Minimum risk decision rule with reject
(2 classes, non-uniform costs)

- The condition for rejecting the sample is:

$$\frac{\lambda_0 - \lambda_{22}}{\lambda_{21} - \lambda_0}\frac{P_2}{P_1} < \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} < \frac{\lambda_{12} - \lambda_0}{\lambda_0 - \lambda_{11}}\frac{P_2}{P_1}$$