BAYESIAN DECISION THEORY

Introduction

All the patterns to recognize belong to J different classes, j=1,2,..., J:

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_J\}$$

Prior probability

We assume that there is some a priori probability (***prior***)

$$P(\omega_j) = \pi_j, \quad j = 1, 2, \ldots, J$$

(Knowledge we have about one class before carrying out an experiment).

Properties:

1. $\pi_j \geq 0, \quad j = 1, 2, \ldots, J$, y
2. $\sum_{j=1}^{J} \pi_j = 1$

Decision rule using only the value of the prior probabilities:

Decide ω_1 if $\pi_1 > \pi_2$
Decide ω_2 if $\pi_1 < \pi_2$

The probability of having an error in classification is the lower value of π_1, π_2.

3

Example of classification using prior probability

Example:

Classification problem: discriminate between healthy people or people with anemia *(Blutarmut)*

Prior knowledge:

- 90% of the people is healthy: $\pi_1 = 0.9$
- 10% of the people is ill: $\pi_2 = 0.1$

If we have to classify a new patient, which is his/her class?

- Decision rule: THAT CLASS WITH GREATER PRIOR PROBABILTY
- Decide ω_1 as $\pi_1 = 0.9 > \pi_2 = 0.1$

If we have no other information, we have to take this decision.

4

Example of classification using prior probability

Example:

Classification problem: discriminate between healthy people or people with anemia *(Blutarmut)*

However, nobody will be satisfied if our family doctor would decide our state of health without looking at us or asking us a blood test.

It is necessary to use more information: we need **measurements** relative to the patterns.

5

Class-conditional probability density

Usually, there is additional information: the value of the observation to classify, **x**.

Considerations:

- Pattern values relative to a class must be essentially different to that of the other classes.
- Patterns of the same class are not exactly equal. (variability of class patterns)

6

Class-conditional probability density

The variability of the measurements is expressed as a random variable **x**, and its probability density function depends on the class ω_j.

$p(x|\omega_j)$ is the ***class-conditional probability density*** function, the probability function for **x** given that the class is ω_j.

For each class ω_j: $\int_x p(x|\omega_j) = 1$

7

Example of classification using class-conditional probability

Example:

Classification problem: discriminate between healthy people or people with anemia *(Blutarmut)*

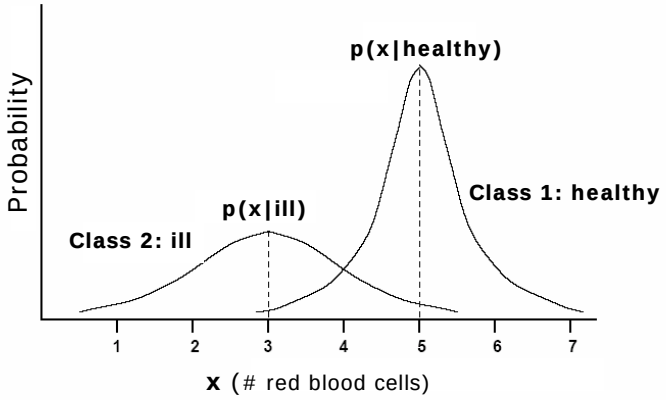
- We have the results of a blood test, so we know the amount of red blood cells.
- The amount of red blood cells is the random variable (**x**) (We do not have the same number of red blod cells than other people).
- This variable has a gaussian distribution.

8

Example of classification using class-conditional probability

Example:

Classification problem: discriminate between healthy people or people with anemia *(Blutarmut)*



9

Example of classification using class-conditional probability

Example:

Classification problem: discriminate between healthy people or people with anemia *(Blutarmut)*

Blood test: 4.500.000 red blood cells

So, the patient is healthy.

p(x=4.500.000 | healthy) > p(x=4.500.000 | ill)

If we consider the patient is healty, the probaility he has 4.5 milion red blood cells is higher than if we consider he is ill, with this number of red blood cells.

10

Posterior probability

Suppose that we know both the prior probabilities and the conditional densities:

$$\pi_1,\ \pi_2 \qquad p(x\ |\omega_j),\quad j=1,2,$$

And suposse further that we measure an observation **x**

How we can use all this information together? → Bayes formula

Posterior probability

$$P(\omega_j\ |x)\ =\ \frac{p(x\ |\omega_j)\ \pi_j}{p(x)} \qquad p(x)=\sum_{i=1}^{J} p(x\ |\omega_i)\ \pi_i$$

Bayes formula shows that by observing the value of **x** we can convert the prior probability π_j to the a posteriori probability (or posterior) **$P(\omega_j|x)$.**

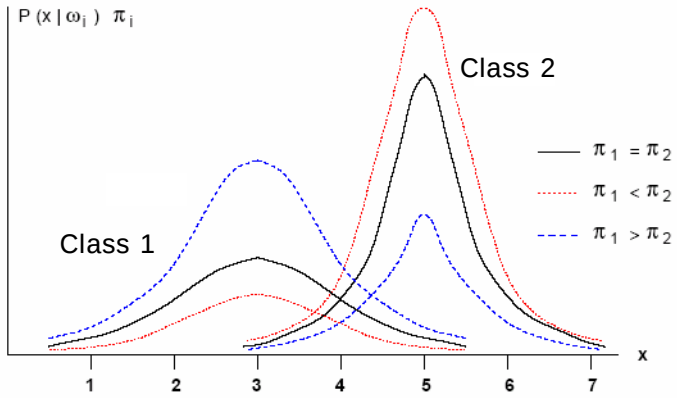
$P(\omega_j|x)$: the probability of the pattern belonging to class ω_j given that the feature value **x** has been measured.

We call **$p(x|\omega_j)$** the ***likelihood*** of ω_j with respect to **x**; π_j is the ***prior***; and **p(x)** is called ***evidence.***

Notice that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability; the evidence factor, **p(x)** , can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one, as all good probabilities must.

Posterior probability

Effect of the ***prob. a priori*** over the ***prob. a posteriori***



13

Bayes decision rule

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$

Decide ω_2 if $P(\omega_2|x) > P(\omega_1|x)$

14

Bayes decision rule

Or an equivalent rule:

Decide ω_1 if $p(x|\omega_1)\pi_1 > p(x|\omega_2)\pi_2$

Decide ω_2 if $p(x|\omega_2)\pi_2 > p(x|\omega_1)\pi_1$

15

Maximun likelihood rule

If $\pi_1 = \pi_2$:

Decide ω_1 if $p(x|\omega_1) > p(x|\omega_2)$

Decide ω_2 if $p(x|\omega_2) > p(x|\omega_1)$

16

Example of classification using the **Bayes rule**

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$
Decide ω_2 if $P(\omega_2|x) > P(\omega_1|x)$

Example:

Classification problem: discriminate between healthy people or people with anemia *(Blutarmut)*

Blood test: 3.500.000 red blood cells (**x**)

Prior knowledge:

- 90% of the people is healthy: $\pi_1 = 0.9$
- 10% of the people is ill: $\pi_2 = 0.1$

Bayes rule

$$* \; P(\omega_1 \,|x = 3.5 \cdot 10^6) \;\propto\; p(x = 3.5 \cdot 10^6 \,|\omega_1)\; \pi_1$$

$$p(x = 3.5 \cdot 10^6 \,|\omega_1)\; \pi_1 \;=\; 0.02 \cdot 0.9 \;=\; \mathbf{0.018}$$

$$* \; P(\omega_2 \,|x = 3.5 \cdot 10^6) \;\propto\; p(x = 3.5 \cdot 10^6 \,|\omega_2)\; \pi_2$$

$$p(x = 3.5 \cdot 10^6 \,|\omega_2)\; \pi_2 \;=\; 0.6 \cdot 0.1 \;=\; \mathbf{0.06}$$

So, the patient is ***ill.***

17

Extension to multidimensional patterns

These concepts can be extended to study more complex and realistic situations:

1. $x \;\Rightarrow\; X$
2. $\Omega = \{\omega_1, \omega_2\} \;\Rightarrow\; \Omega \;=\; \{\omega_1, \omega_2, \ldots, \omega_J\}$

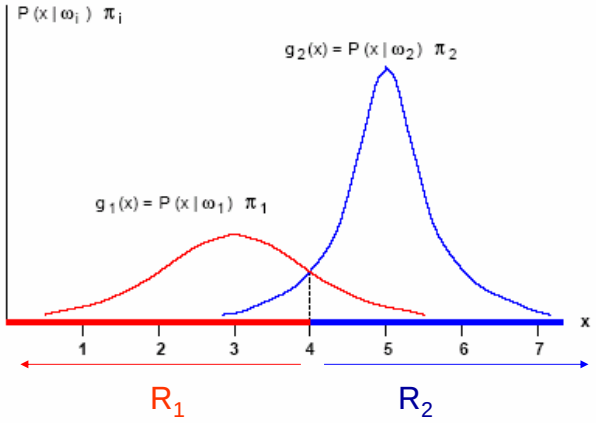
Bayes rule or Bayes classifier:

Select ω_i if $\mathbf{P(\omega_i|x) > P(\omega_j|x)}$ for all $\mathbf{j \neq i}$

18

Decision boundary

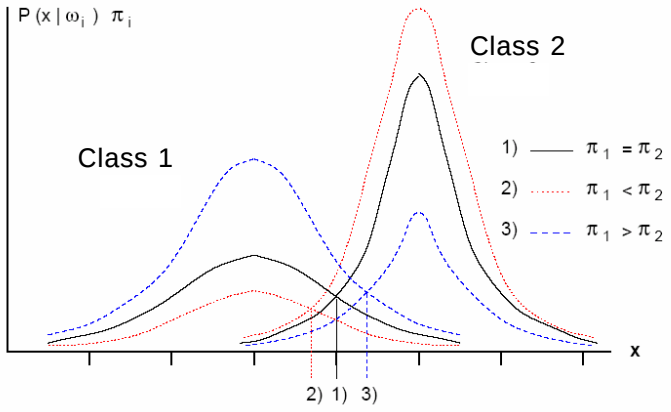
In an uni-dimensional case, the decision boundary is just one point, and the decision regions are intervals in the x-axis.



19

Decision boundary

Different values of 'prior probabilities' make the decision boundary to move.



20

Error in classification

In a two-category classification and one dimension patterns, the possible errors are:

- **x** is in $\mathbf{R_1}$ and its real class is ω_2
- **x** is in $\mathbf{R_2}$ and its real class is ω_1

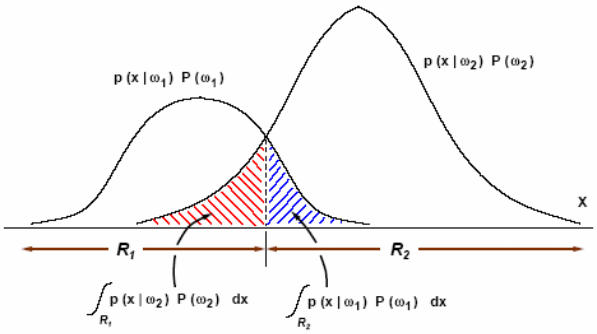
21

Error in classification

$$\begin{aligned} P(\text{error}) &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\ &= P(x \in R_2 | \omega_1)\pi_1 + P(x \in R_1 | \omega_2)\pi_2 \\ &= \int_{R_2} P(x|\omega_1)\pi_1 dx + \int_{R_1} P(x|\omega_2)\pi_2 dx \end{aligned}$$

22

Error in classification

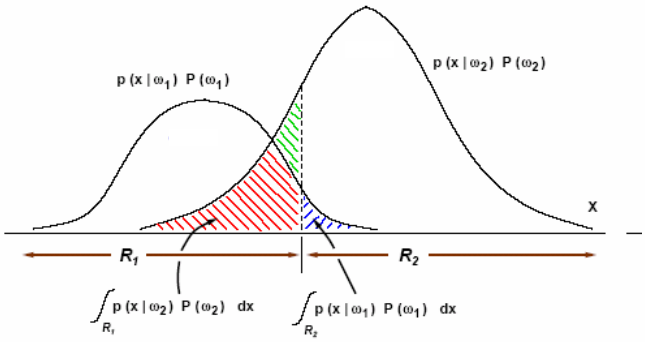


MINIMUM ERROR

23

Error in classification

If the decision boundary is not chosen correctly, the error is not minimum.



ERROR NOT MINIMUM

24

Error in classification

The Bayes classifier minimizes the average probability of error, so the best choice is to use the Bayes rule as the classifier of the pattern recognition system.

However, in most practical cases, the class-conditional probabilities are not known, and that fact makes impossible the use of the Bayes rule.

25

An example

PATTERN RECOGNITION SYSTEM:

CONSENT A MORTGAGE OR NOT

Variables:	**Classes:**
•*Age*	•*Yes, he/she will return the money (GOOD PAYER)*
•*Wage bill*	•*No, he/she will NOT return the money (BAD PAYER)*

We have 20 examples, 10 per each class

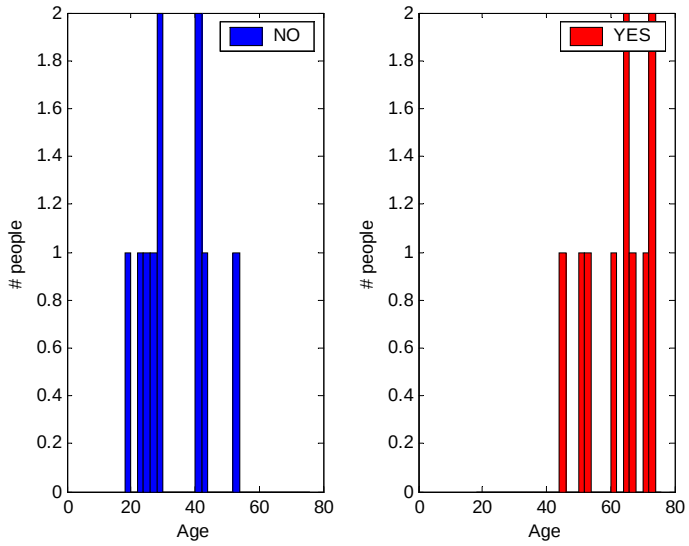
26

An example

	age	wage bill (euros)	class
1	25	106	BAD PAYER
2	41	195	BAD PAYER
3	19	259	BAD PAYER
4	26	1424	BAD PAYER
5	41	212	BAD PAYER
6	29	964	BAD PAYER
7	28	420	BAD PAYER
8	43	344	BAD PAYER
9	52	1113	BAD PAYER
10	22	857	BAD PAYER
11	72	2431	GOOD PAYER
12	51	1244	GOOD PAYER
13	64	2150	GOOD PAYER
14	72	1895	GOOD PAYER
15	60	1191	GOOD PAYER
16	70	986	GOOD PAYER
17	66	1388	GOOD PAYER
18	52	2454	GOOD PAYER
19	65	2422	GOOD PAYER
20	45	1293	GOOD PAYER

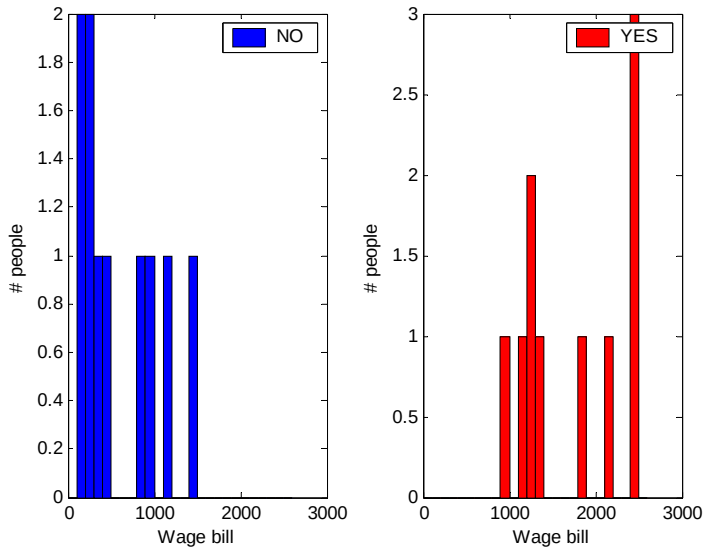
27

An example



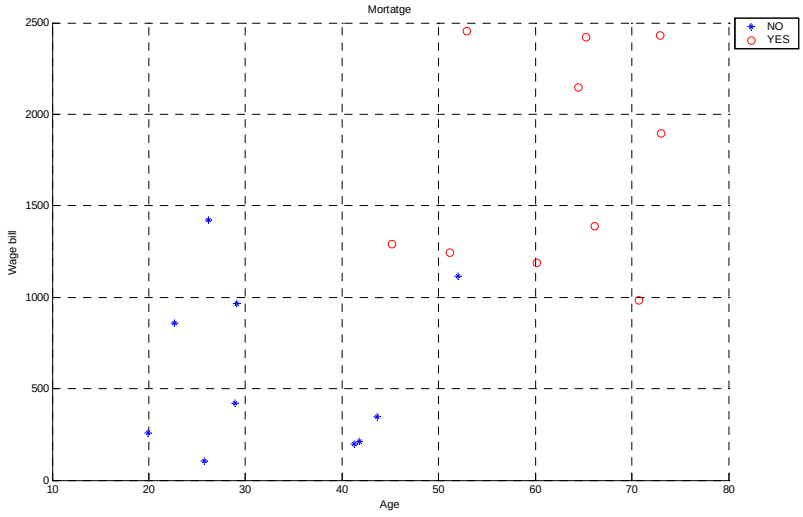
28

An example



29

An example



30

An example

Prior probability: 0.5 for both classes

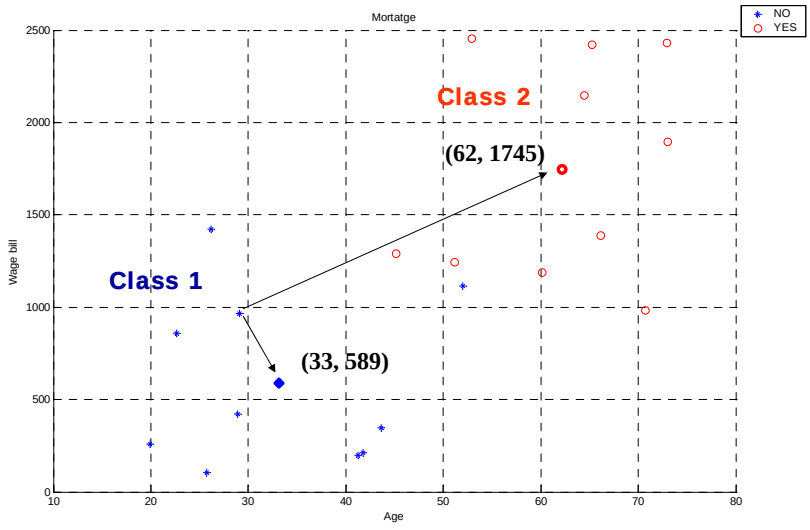
We don´t know the class-conditional probability density functions of the variables, so we can´t use the Bayes rule.

We will use a simple classifier:

Euclidean distance to the mean of each class (nearest neighbour)

31

An example



32

An example

	age	wage bill	class	distance to the class 1 mean	distance to the class 2 mean	output classifier
1	25	106	BAD PAYER	483	1639	BAD PAYER
2	41	195	BAD PAYER	394	1550	BAD PAYER
3	19	259	BAD PAYER	330	1486	BAD PAYER
4	26	1424	BAD PAYER	834	323	GOOD PAYER
5	41	212	BAD PAYER	377	1533	BAD PAYER
6	29	964	BAD PAYER	374	782	BAD PAYER
7	28	420	BAD PAYER	169	1325	BAD PAYER
8	43	344	BAD PAYER	245	1401	BAD PAYER
9	52	1113	BAD PAYER	524	632	BAD PAYER
10	22	857	BAD PAYER	268	888	BAD PAYER
11	72	2431	GOOD PAYER	1842	685	GOOD PAYER
12	51	1244	GOOD PAYER	654	501	GOOD PAYER
13	64	2150	GOOD PAYER	1560	404	GOOD PAYER
14	72	1895	GOOD PAYER	1306	150	GOOD PAYER
15	60	1191	GOOD PAYER	601	554	GOOD PAYER
16	70	986	GOOD PAYER	397	759	BAD PAYER
17	66	1388	GOOD PAYER	799	357	GOOD PAYER
18	52	2454	GOOD PAYER	1864	708	GOOD PAYER
19	65	2422	GOOD PAYER	1832	676	GOOD PAYER
20	45	1293	GOOD PAYER	703	452	GOOD PAYER

33

An example

Confusion Matrix

Class	1	2	Total	Success	Error
1	**9**	**1**	10	90%	10%
2	**1**	**9**	10	90%	10%
Total	10	10	20	**90%**	**10%**

Recognition Rate

34

Performance of the classifier

Measuring the performance of the classifier on the training images is not fair (*too optimistic*):

- The goal is to know the expected performance of the classifier when it is fed with new data.
- Example: the goal is to know if a face recognition system trained with images of a certain person will recognize him/her the following day (using a previously unseen image).
- In brief, the performance of a classifier has to be tested with previously unseen examples (not used for training).

35

Performance of the classifier

Training set: the set of examples used for training the classifier.

Test set: the set of examples used to measure the expected classification accuracy.

Images (or examples) of the test set should never belong to the training set.

36