Bayesian Decision Theory

Computational Social Intelligence - Lecture 12

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

http://www.dcs.gla.ac.uk/vincia
Alessandro.Vinciarelli@glasgow.ac.uk

Texts (see Moodle)

This lecture is based on the following text (available on Moodle):

- Chapter 5 of F.Camastra and A.Vinciarelli, "Machine Learning for Audio, Image and Video Processing", Springer Verlag, 2008.

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Thomas Bayes (1701-1761)

Hence the name Bayesian Decision Theory

Bayesian Decision Theory

- Bayesian Decision Theory is a statistical approach for the formalisation of common sense;
- It is one of the main approaches that Artificial Intelligence technologies adopt to "make decisions";
- Bayesian Decision Theory has nothing to do with the way humans make decisions.

The Feature Vectors

The data is represented with D-dimensional vectors

The D components are physical measurements extracted from the data

$$\vec{x} \in \mathbb{R}^D$$

$$\vec{x} = (x_1, \ldots, x_j, \ldots, x_D)$$

The D physical measurements are called features

The Decisions

The set C includes M possible decisions

The number M of possible decisions is finite

$$C = \{C_1, \ldots, C_i, \ldots, C_M\}$$

The M decision are called classes

AI Decisions

In AI, decision means to map a feature vector into a decision

$$\vec{x} \rightarrow C_j$$

$$C_j \in C$$

The decision must belong to the set of the M possible decisions

Toy Example (Logic)

The vector represents a student

AC and Ex are Assessed Exercise and Exam scores, respectively

$$\vec{x} = (AC, Ex)$$

$$\mathcal{C} = \{passed, failed\}$$

The decisions is easy because there is a rule (e.g., both AC and Ex above C3)

The machine must decide whether it is passed or failed

Toy Example (AI)

The vector represents a patient

T and P are temperature and blood pressure, respectively

$$\vec{x} = (T, P)$$

$$C = \{ill, healthy\}$$

The difficulty is that there is uncertainty, it is not possible to solve the problem with a rule

The machine must decide whether a person is ill or healthy

AI and Decisions

- When all the necessary information is available and there is a rule, logic is sufficient and no decision theory is necessary (typical of IT);

- When the information is partial and there is uncertainty, logic is not sufficient and Bayesian Decision Theory is necessary (typical of AI);

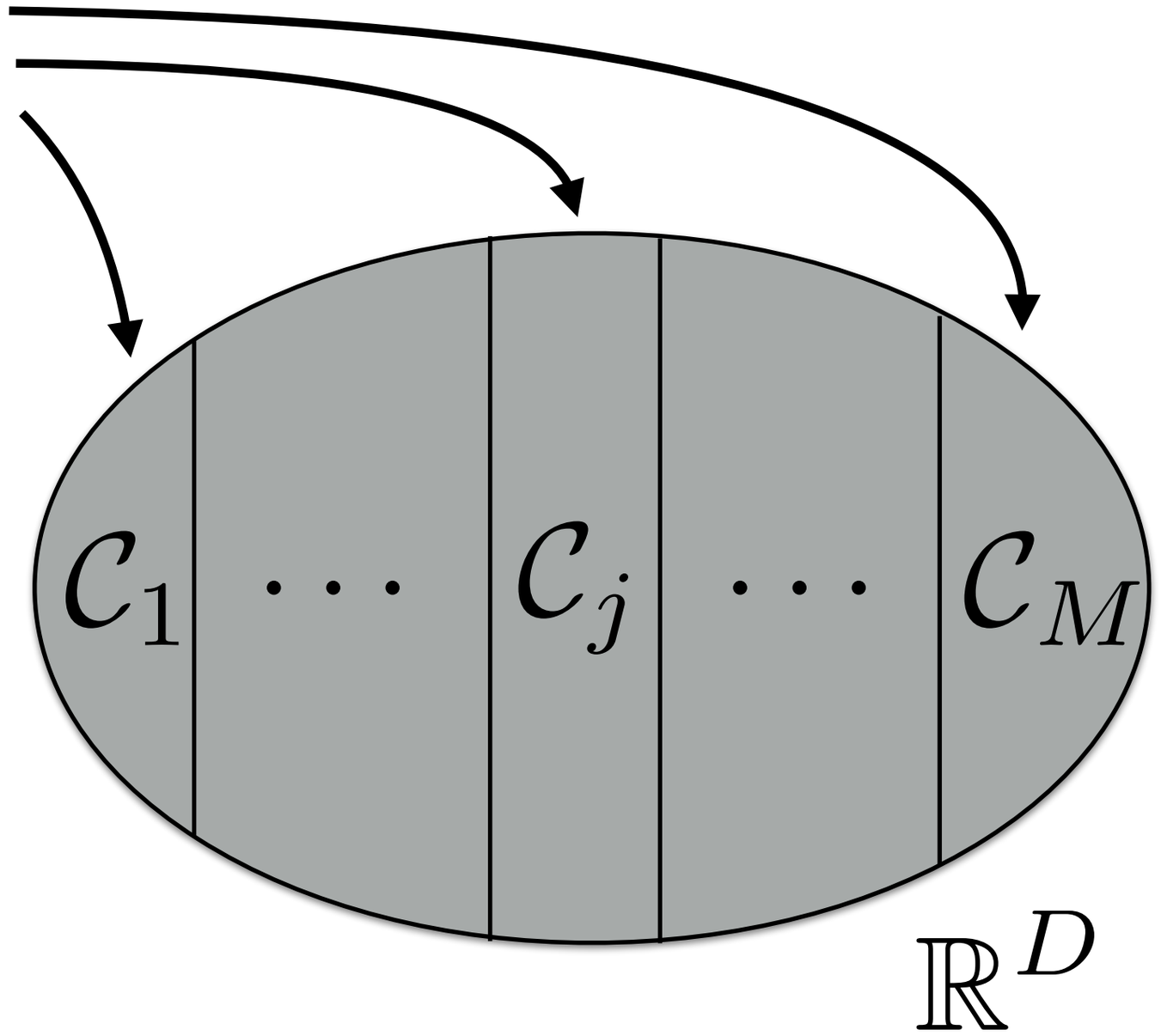
- The decision process is often referred to as "classification" (the decisions can be thought of as classes or categories).

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

From Decisions to Classes (I)

Every decision corresponds to a subset of the space, a class



$$\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$$

The classes are disjoint

From Decisions to Classes (II)

This is the right decision for the vector

$\vec{x} \in C_i$

It is equivalent to saying that the vector belongs to class i

A-Priori Probability

A-priori probability of class i

$$p(\vec{x} \in \mathcal{C}_i) = p(\mathcal{C}_i)$$

$$\sum_{k=1}^{M} p(\mathcal{C}_k) = 1$$

The classes correspond to mutually exclusive events

Prior Decision Rule

The class is the one with the highest a-priori probability

$$C^* = \arg\max_{C_k \in C} p(C_k)$$

The decision approach takes into account all possible decisions

Toy Example (I)

There are two mutually exclusive classes

$$\mathcal{C} = \{\mathcal{C}_1 = woman, \mathcal{C}_2 = man\}$$

$$p(\mathcal{C}_1) = 0.15$$

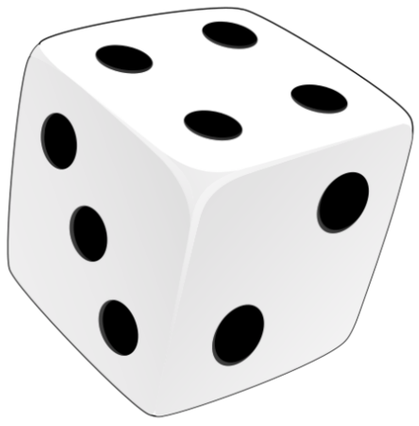
$$p(\mathcal{C}_2) = 0.85$$

The class with the highest a-priori probability is “man”

Toy Example (II)

The problem is to decide what is the outcome after rolling the dice

$$\mathcal{C} = \left\{ \begin{array}{l} \mathcal{C}_1 = 1 \\ \mathcal{C}_2 = 2 \\ \mathcal{C}_3 = 3 \\ \mathcal{C}_4 = 4 \\ \mathcal{C}_5 = 5 \\ \mathcal{C}_6 = 6 \end{array} \right\}$$



Each class corresponds to one of the possible outcomes

Toy Example (II)

The a-priori probability is the same for all classes

$$p(\mathcal{C}_i) = \frac{1}{6}$$

$$i = 1, 2, \ldots, 6$$

$$\arg\max_{\mathcal{C}_i \in \mathcal{C}} p(\mathcal{C}_i) = ?$$

It is not possible to identify the highest a-priori probability

Uninformative Prior

In some cases, all classes have the same a-priori probability

$$p(\mathcal{C}_1) = \ldots = p(\mathcal{C}_i) = \ldots = p(\mathcal{C}_M)$$

The prior rule is like rolling a dice and the prior is uninformative

Recap

- The prior rule can be used when the only available information is the a-priori probability of the classes;
- The common sense suggests to make the decision corresponding to the class with the highest a-priori probability;
- The prior rule does not take into account the features.

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Taking the Features into Account

The joint probability of class and feature vector

$$p(C_i, \vec{x}) = p(C_i|\vec{x})p(\vec{x}) = p(\vec{x}|C_i)p(C_i)$$

The Product Law allows one to write the joint probability in two ways

The Bayes Theorem

A-posteriori probability (or posterior) of class i

Likelihood of class i with respect to the feature vector

$$p(C_i|\vec{x}) = \frac{p(\vec{x}|C_i)p(C_i)}{p(\vec{x})}$$

The evidence

The a-priori probability of class i

The Evidence

Evidence

$$p(\vec{x}) = \sum_{i=1}^{M} p(\vec{x}|\mathcal{C}_i)p(\mathcal{C}_i)$$

A combination of product and addition laws

The Bayes Theorem

The sum over the posteriors is 1

$$\sum_{i=1}^{M} p(\mathcal{C}_i|\vec{x}) = \frac{\sum_{i=1}^{M} p(\vec{x}|\mathcal{C}_i)p(\mathcal{C}_i)}{\sum_{k=1}^{M} p(\vec{x}|\mathcal{C}_k)p(\mathcal{C}_k)} = 1$$

The evidence is a normalisation constant

Outline

- Introduction
- Bayesian Decision Theory
- Conclusions

Error Probability

Probability of error if the right decision is j

$$p(err|\vec{x}) = \sum_{i \neq j} p(C_i|\vec{x})$$

The sum over all posteriors except the posterior of j

Error Probability

The class that corresponds to the highest posterior

$$C^* = \arg\max_{C_k \in C} p(C_k|\vec{x})$$

The value of the posterior is checked for all possible classes

The posterior takes the features into account

Posterior Rule

The expression of the priors according to the Bayes Theorem

$$C^* = \arg\max_{C_k \in C} \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} =$$

$$= \arg\max_{C_k \in C} p(\vec{x}|C_k)p(C_k)$$

The evidence is the same for all classes and it can be eliminated

Toy Example

The feature vector represents a student

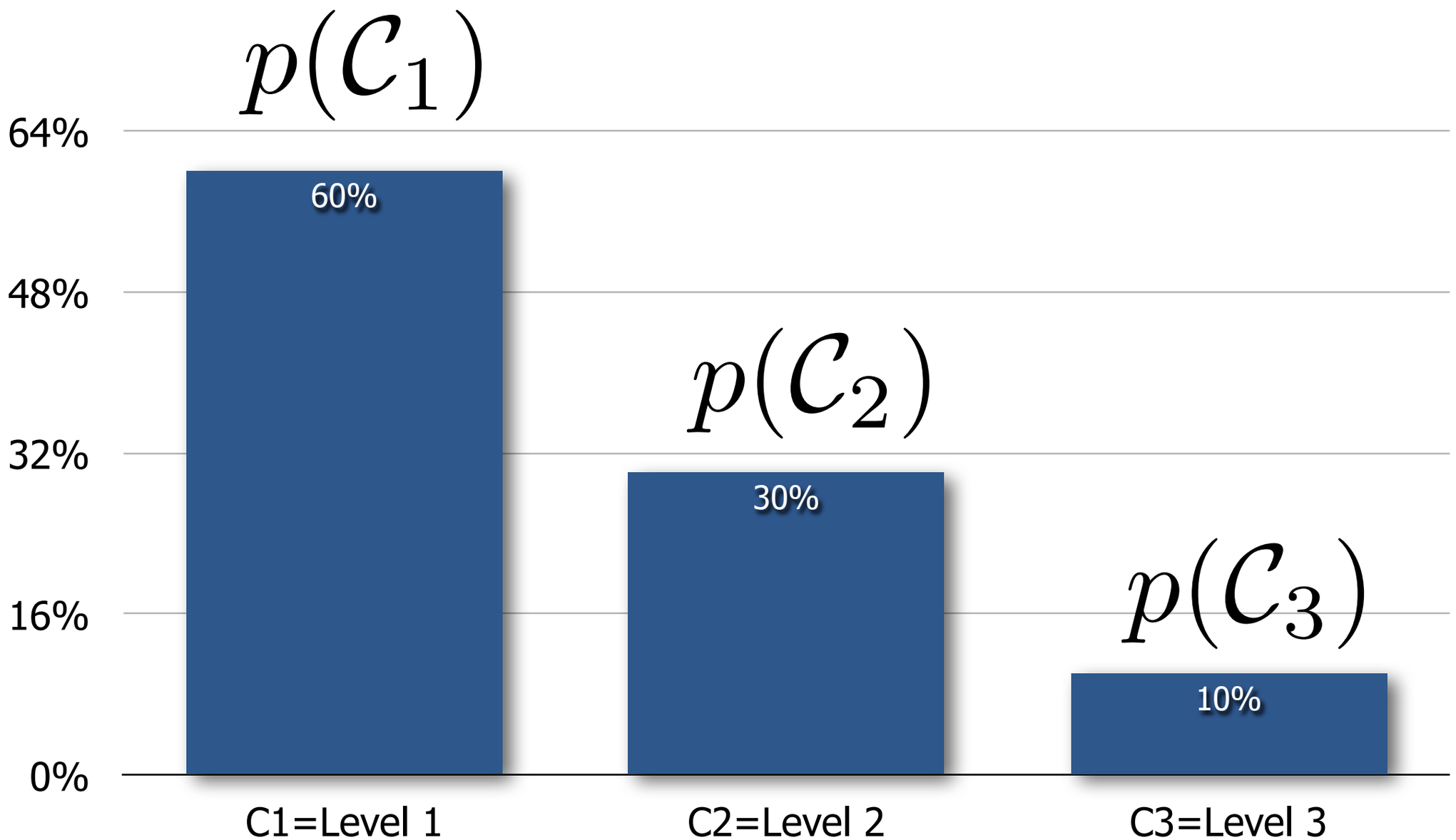
The only feature is the age of the student

$$\vec{x} = (age)$$

$$C = \{L1, L2, L3\}$$

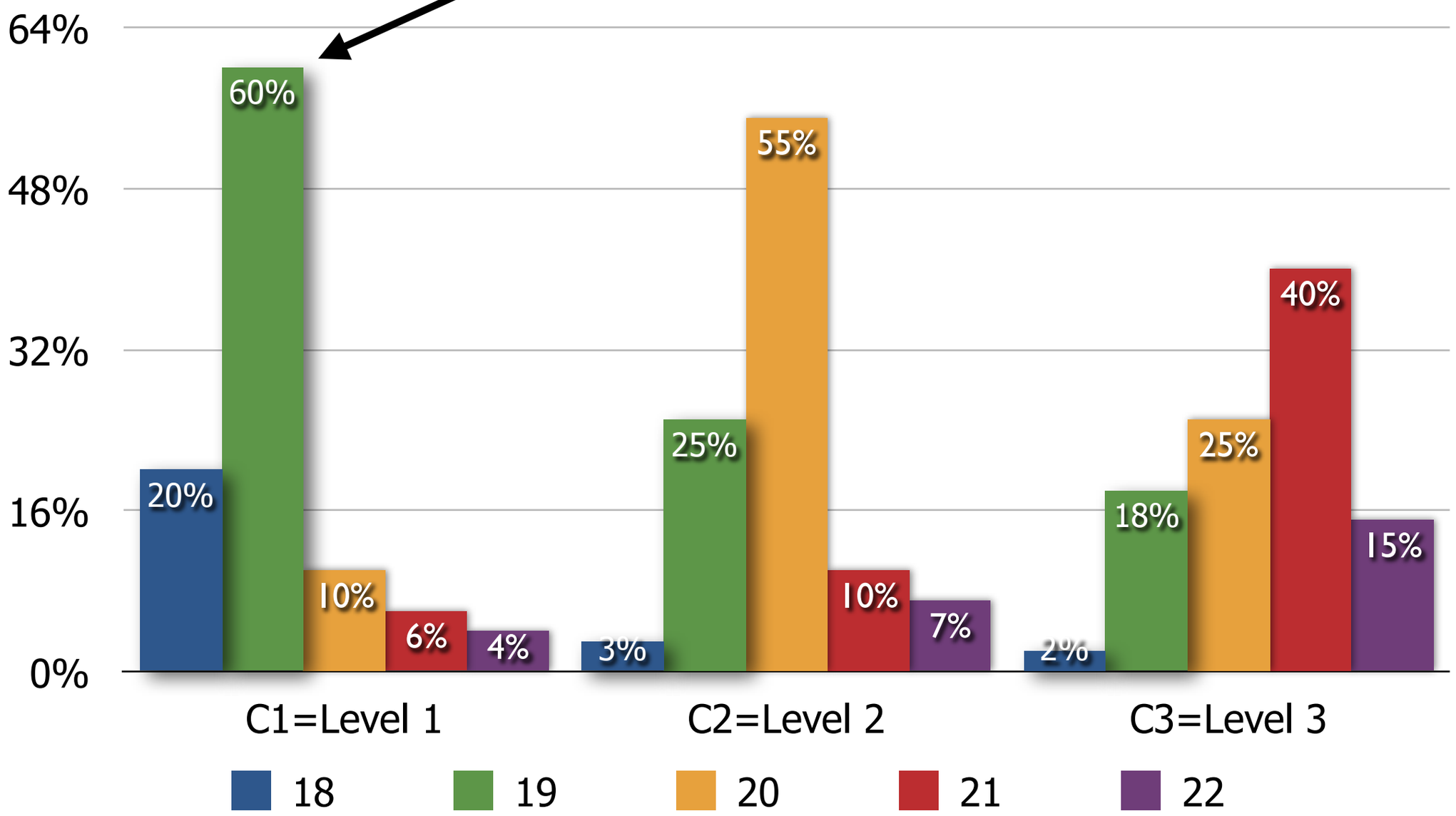
The classes correspond to the Levels 1, 2 and 3

Priors



Likelihoods

$$p(x = 19|\mathcal{C}_1)$$



Decision

The age is known, but the level is not

$$\vec{x} = (21)$$

$$p(\vec{x}|\mathcal{C}_1)p(\mathcal{C}_1) = 0.036$$
$$p(\vec{x}|\mathcal{C}_2)p(\mathcal{C}_2) = 0.030$$
$$p(\vec{x}|\mathcal{C}_3)p(\mathcal{C}_3) = 0.040$$

$$\mathcal{C}_3 = \arg\max_{\mathcal{C}_i \in \mathcal{C}} p(\vec{x}|\mathcal{C}_i)p(\mathcal{C}_i)$$

Class C3 corresponds to the highest posterior

Recap

- Unlike the prior rule, the posterior rule takes into account the features;
- The goal of the posterior rule is to minimise the error probability (in this respect it formalises common sense);
- The main problem left open is how to estimate the a-priori and a-posteriori probabilities involved in the problem.

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Conclusions

- In the Bayesian Decision Theory, making a decision means to map a feature vector into one of M predefine decisions;
- The set of the feature vectors for which the right decision is the same can be thought of as a class (a subset of the feature space);
- The decision process is often referred to as classification.