EDAMI laboratory

Lab5 – classification

Basic notions

- Object – an entity described by a set of attributes
  - Nominal attributes
  - Numerical attributes
- Database (DB) – a set of objects.

Classification

Classification is a process of assigning a given object to a certain class (group) from a predefined set of classes.

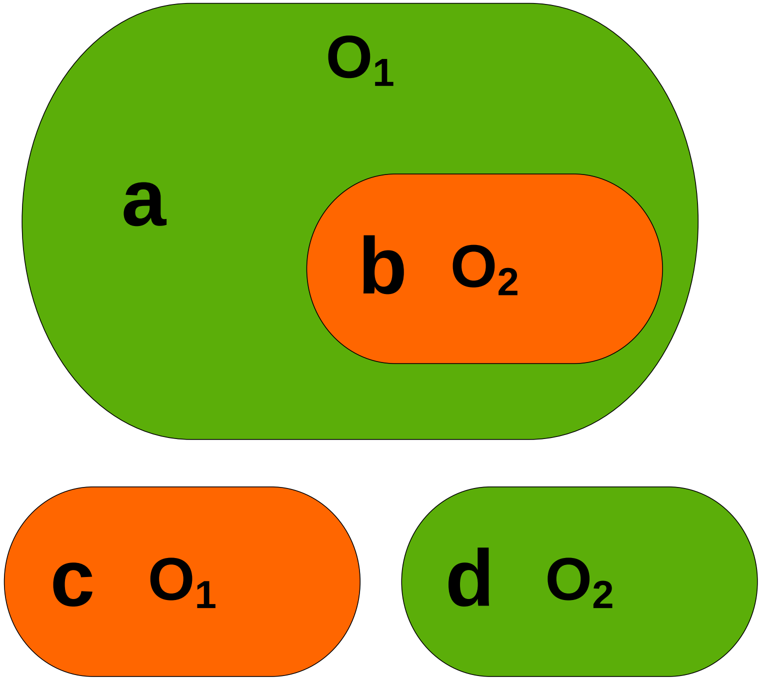
Typically, only one class is assigned.

Classification in the data mining area refers to the methods of automatic building of classifiers based on training data including labelled objects.

The purpose of the classification is to find the rule (rules) which assigns any object *w* its class *c* .

Quality of classification

- a – number of objects classified correctly to a given class
- b – number of objects classified erroneously to a given class
- c – number of objects erroneously not classified to a given class.
- d – number of objects correctly not classified to a given class.



Evaluation of classification quality

- precision

$$precision=\frac{a}{a+b}$$

- recall

$$recall=\frac{a}{a+c}$$

- accuracy

$$accuracy=\frac{a+d}{a+b+c+d}$$

- error

$$error=\frac{b+c}{a+b+c+d}$$

F- measure

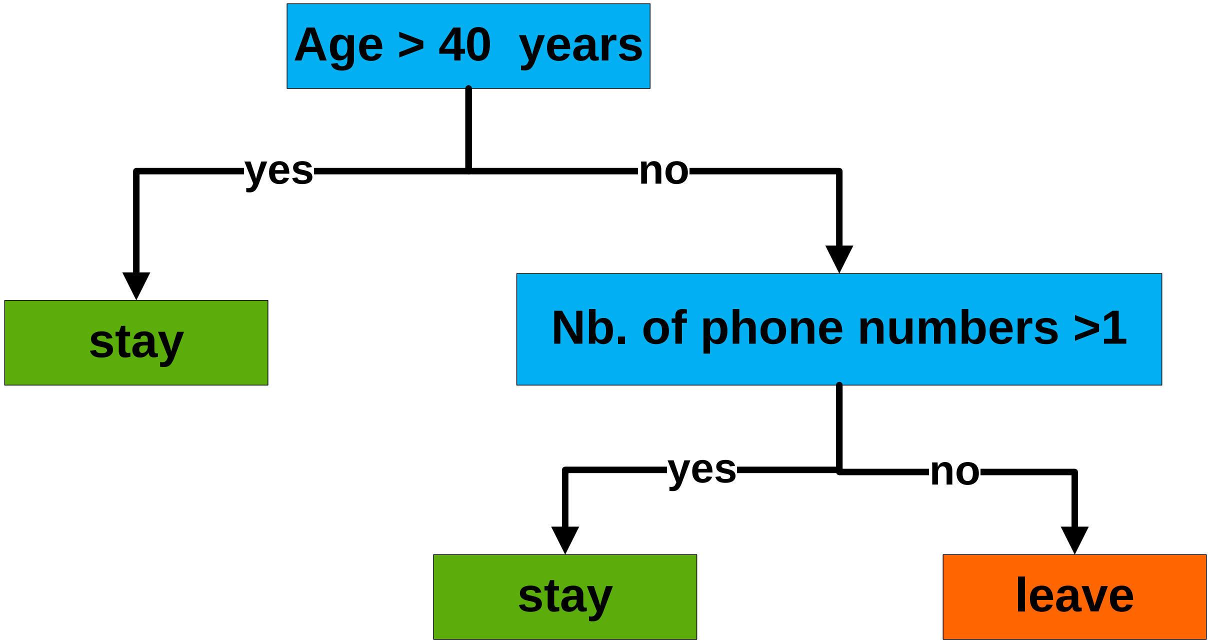
$$F_{\beta}=(1+\beta 2)\frac{precision*recall}{\beta 2*precision+recall}$$

β - the factor indicating how much recall is more essential than precision.

If recall and precision are equally important the equation is in the following form:

$$F_{\beta}=2\frac{precision*recall}{precision+recall}$$

Decision tree - example



Decision tree: problems

1. How to partition the data at each step?
2. When to stop partitioning?
3. How to predict the value of a class/category for each object in a partition?

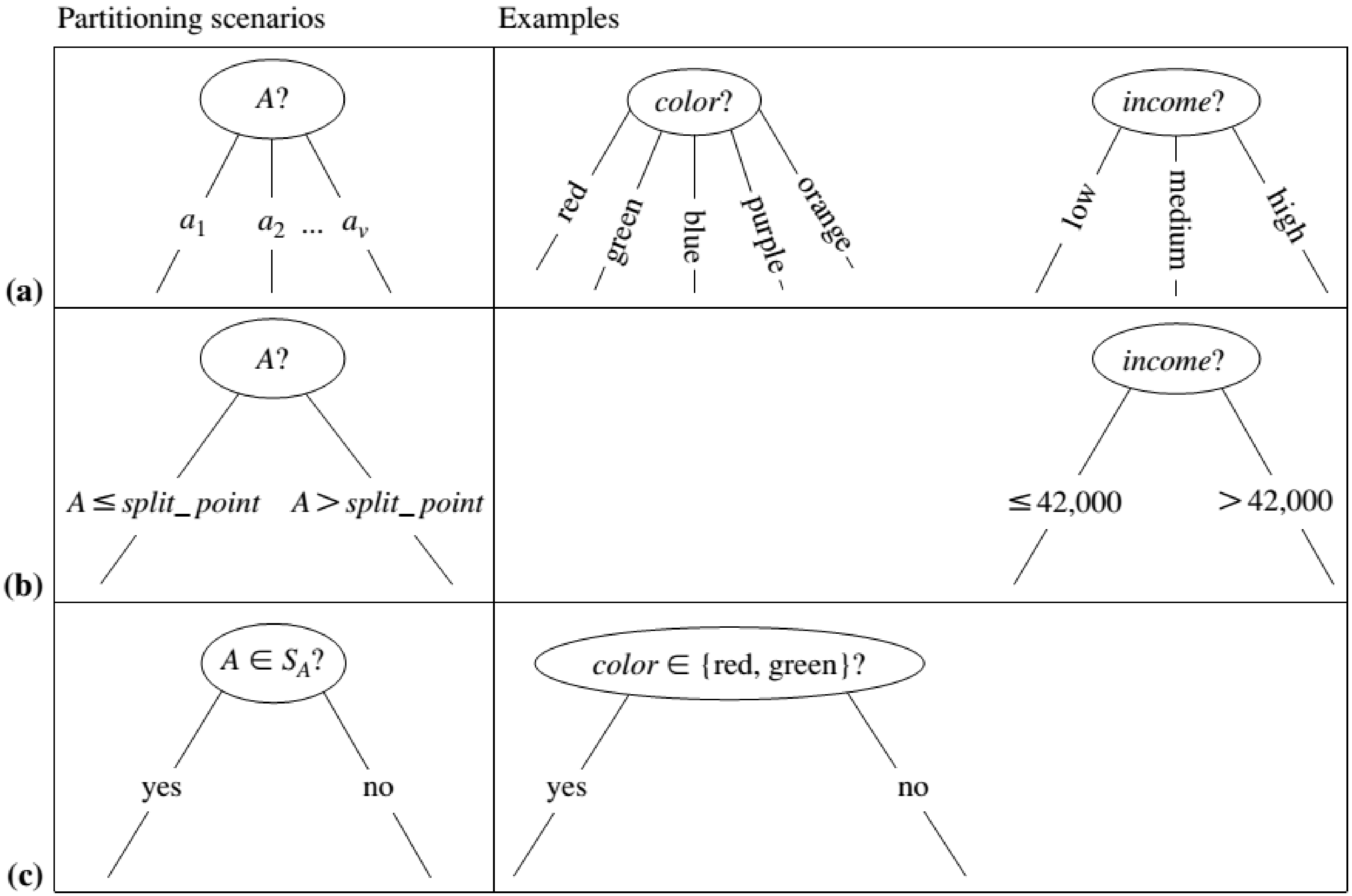
Decision tree

- A decision tree is built in two phases:
  - building,
  - pruning to avoid overfitting.
- A decision tree is built by recursively splitting input sets in nodes until:
  - input data includes only objects belonging to one class,
  - the number of objects in an input set is small enough.

Split selection

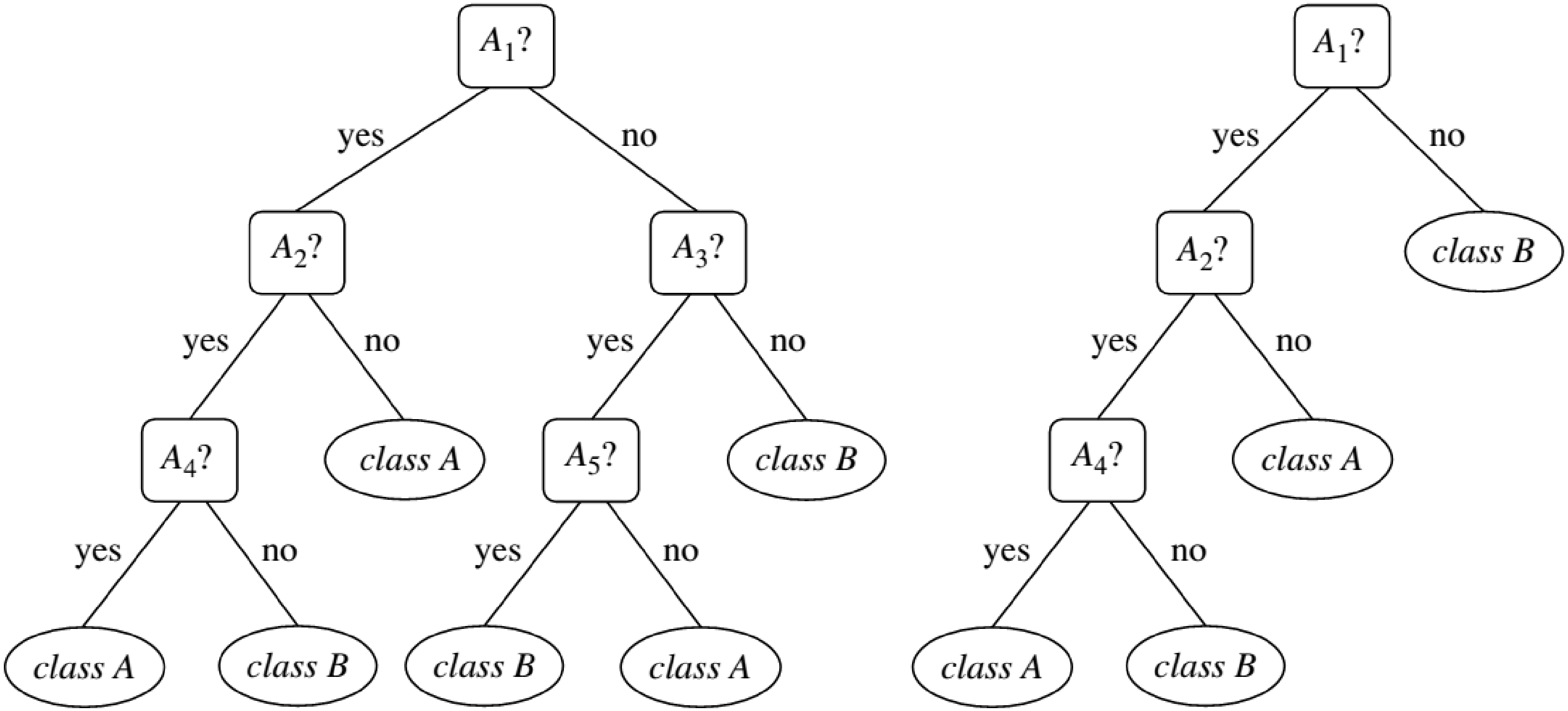
- Information-Gain (ID3)
- Information-Gain Ratio (C4.5)
- Gini Index

Example of tests for splitting



a) Discrete attributes
b) Numerical attributes
c) Discrete attributes – binary split

Tree before and after pruning



Naive Bayes Classifier

- Assumes the values of attributes are statistically independent.
- Assigns a class C to an object *w* such that conditional probability $P(C|w)$ is the biggest.

Classification in R

- Package party
  - ctree(…) - for building trees
  - print(ctree), plot(ctree) – for presentation of trees parameters
- Package rpart
  - rpart(…) - for building trees
  - prune(rpart,…) – for pruning trees
  - print(rpart), summary(rpart), plot(rpart), text(rpart) - for presentation of trees parameters
- Package e1071
  - naiveBayes – for building Naive Bayes classifier
  - print(naiveBayes) for presentation of a model

Laboratory task

Define the classification problem related to one of the datasets listed below. Find the best classifier for the selected dataset according to the chosen way of classifier quality evaluation.

exomple : prop. da guardare

CLASSES
WINE → GOOD / NOT GOOD

ACCURACY 50%

PRIMA DI TUTTO GUARDA PREC. POI ACCURACY

PRECISION → ACC.

Datasets:

- http://archive.ics.uci.edu/ml/datasets/Wine+Quality
- http://archive.ics.uci.edu/ml/datasets/Wine+Quality
- http://archive.ics.uci.edu/ml/datasets/Abalone