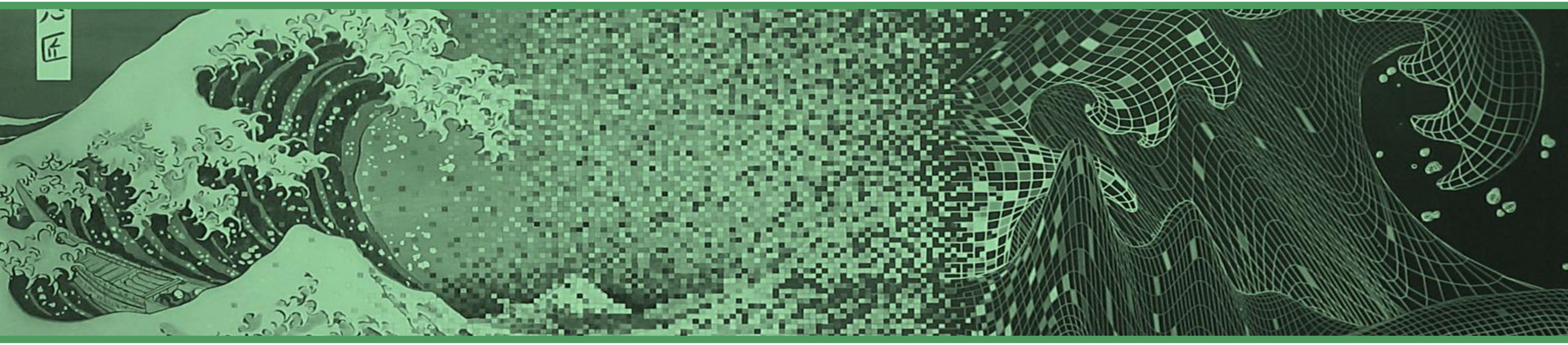
TÉCNICO LISBOA

Decision trees

Associative learning and pattern mining



Aprendizagem 2023

Rui Henriques rmch@tecnico.ulisboa.pt
Andreas Wischert andreas.wichert@tecnico.ulisboa.pt

Outline

- **Decision trees**
  – associative learning
  – information gain
  – continuous variables
  – addressing overfitting
  – variants: ID3, C4.5, CART
- Advanced aspects (*optional*)
  – ensembles
  – pattern mining

Associative learning

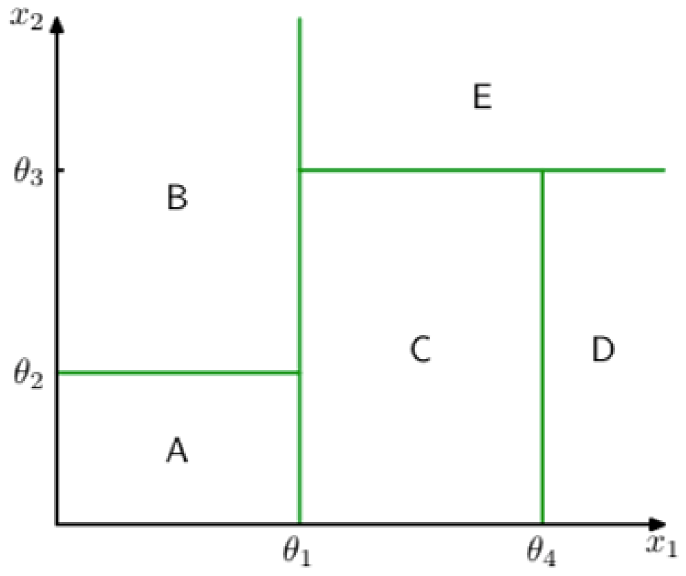
- Finding relevant **associations** within data
  - e.g. $\text{gender} = male \wedge \text{age} > 50 \wedge \text{BMI} > 35 \wedge \text{infection} = positive \Rightarrow hospitalization$
  - a central notion in ML

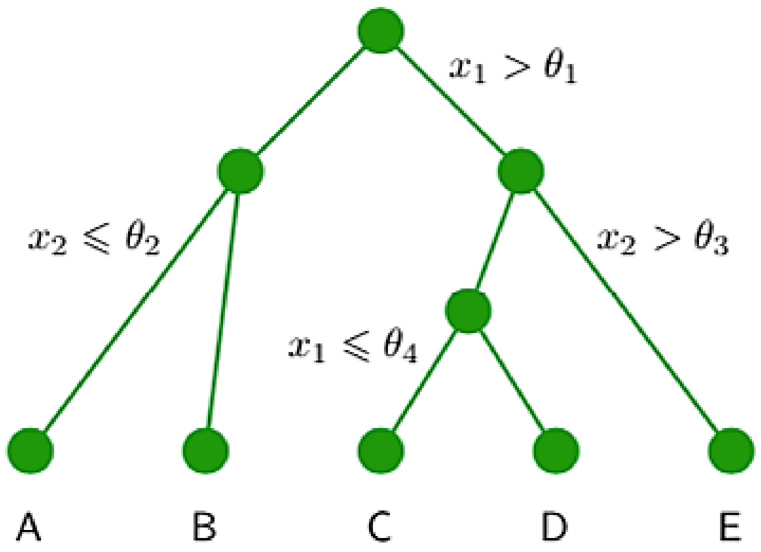
- **Predictive models**
  - **decision trees**
  - **ensembles**: random forests, XGBoost...

- **Descriptive models**
  - pattern mining
  - subspace clustering

Decision trees

- Predictive model given by a tree
  - relationship between discriminative features and outcomes
    - applicable to both categoric output variables (classification) and numeric (regression)
  - each path from root to leaf is an association rule





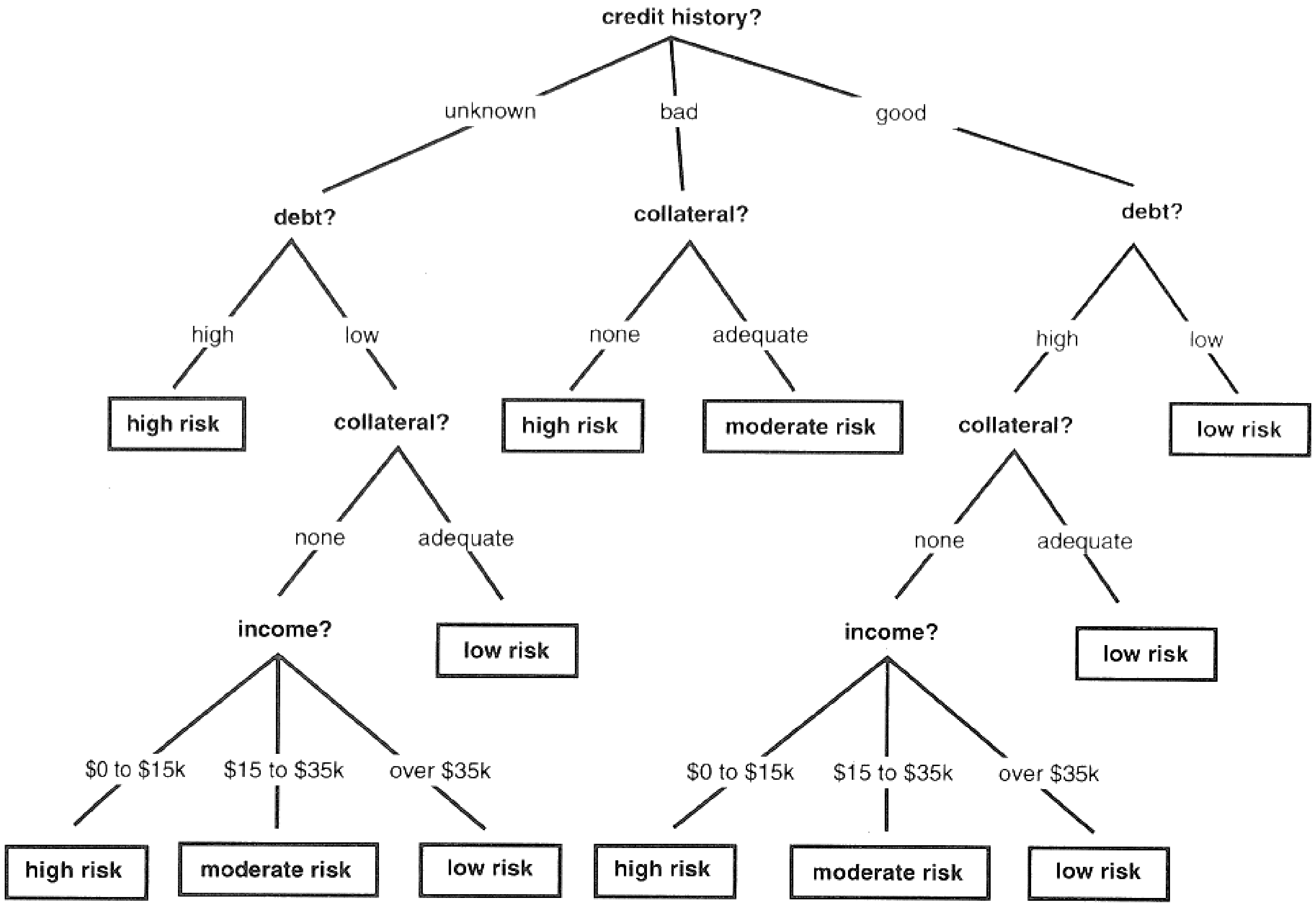
Example: credit risk

- Let us consider the credit risk assessment domain

- Exercise:
  - draw a decision tree able to correctly classify all observations

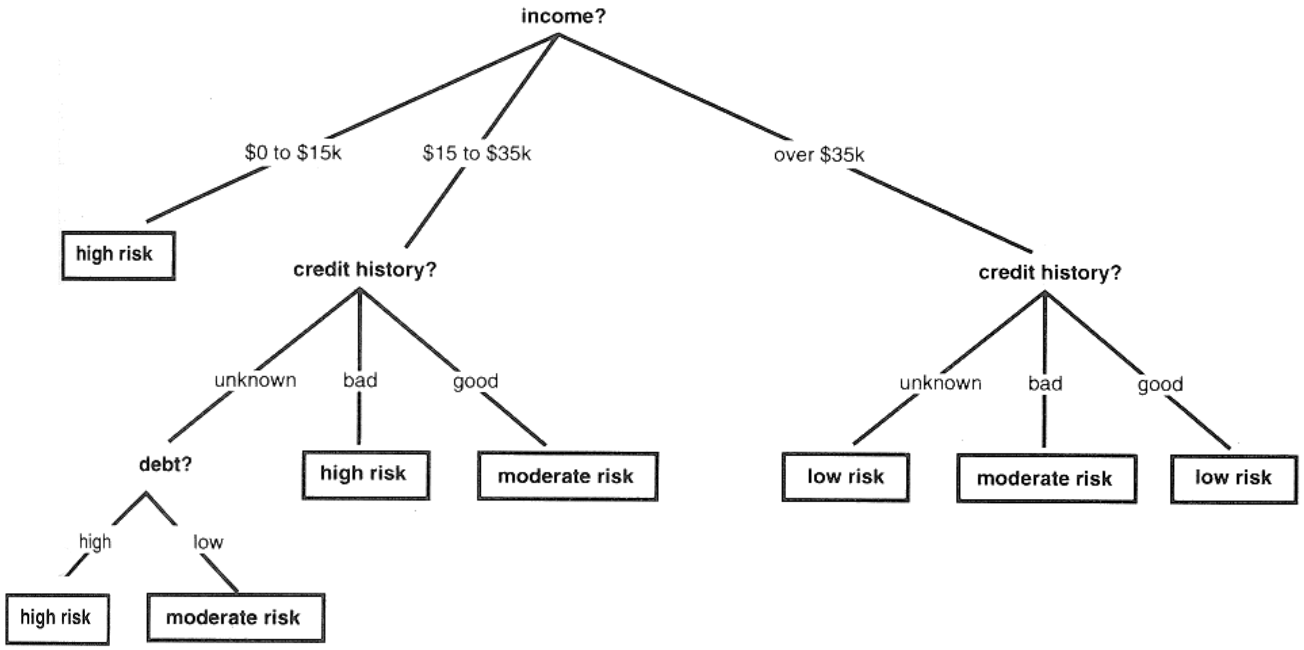
	risk	credit history	debt	collateral	income
$x1$	high	bad	high	none	\$0-\$15k
$x2$	high	unknown	high	none	\$15k-\$35k
$x3$	moderate	unknown	low	none	\$15k-\$35k
$x4$	high	unknown	low	none	\$0-\$15k
$x5$	low	unknown	low	none	>\$35k
$x6$	low	unknown	low	adequate	>\$35k
$x7$	high	bad	low	none	\$0-\$15k
$x8$	moderate	bad	low	adequate	>\$35k
$x9$	low	good	low	none	>\$35k
$x10$	low	good	high	adequate	>\$35k
$x11$	high	good	high	none	\$0-\$15k
$x12$	moderate	good	high	none	\$15k-\$35k
$x13$	low	good	high	none	>\$35k
$x14$	high	bad	high	none	\$15k-\$35k

Decision tree for credit risk



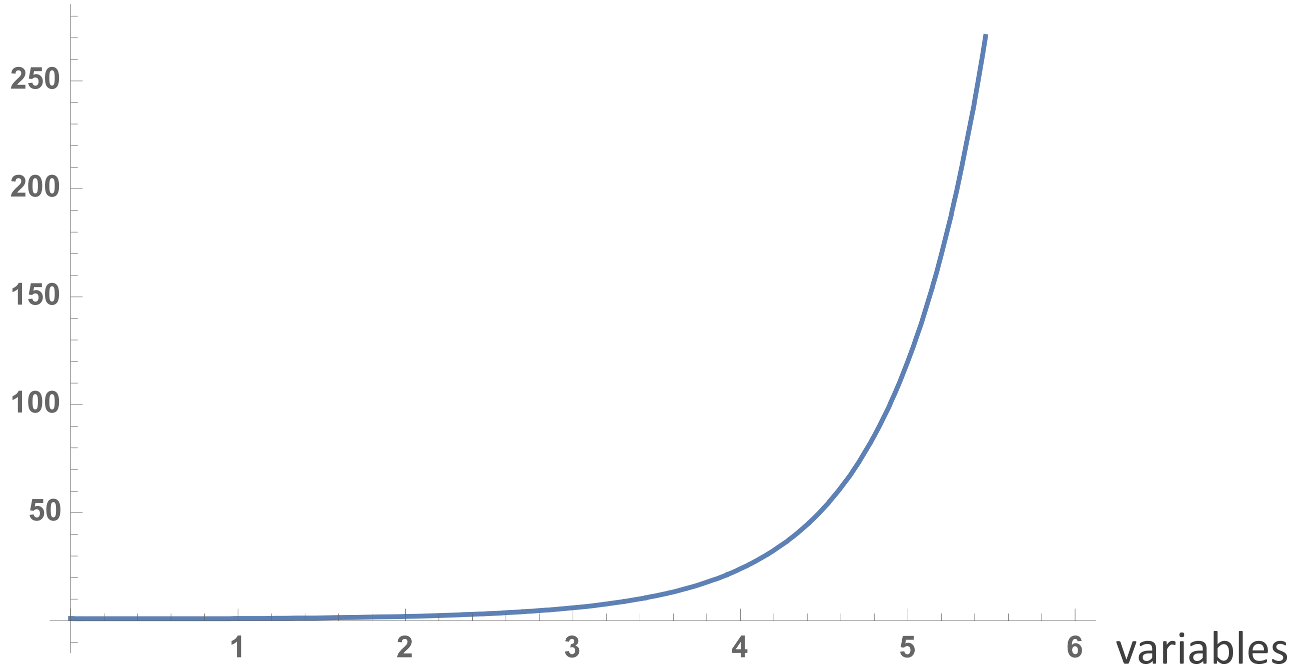
Decision trees

- The size of a tree necessary to classify a given set of observation varies…
  - …according to the order with which variables are tested
- Given a set of different decision trees, we may ask:
  - which tree has the greatest ability to classify the population?
  - example: simplified decision tree for credit risk assessment able to correctly classify all observations



Decision trees

- How many different decision tree exist?
  - there exist $m!$ different ordering of variables, $m!$ different decision trees
- Algorithm: compute all $m!$ decision trees and chose the smallest one
  - blind search finds the global minima, the smallest decision tree (optimal)
- Problem: computational complexity!
  - $m!$ grows extremely fast



Best decision trees

- Decision tree learning generally assumes that a good decision tree is the ***simplest*** decision tree
  - *heuristic*: preferring simplicity and avoiding unnecessary assumptions
  - in accordance with Occam Razor principle:
    - *“all other things being equal, the simplest model is the best”*

Occam Razor was first articulated by the medieval logician William of Occam in 1324:

“vain do with more what can be done with less..”

Outline

- **Decision trees**
  - associative learning
  - **information gain**
  - continuous variables
  - addressing overfitting
  - variants: ID3, C4.5, CART
- Advanced aspects
  - ensembles
  - pattern mining

Heuristic function: information gain

- Then how to learn a *good* and *simple* decision tree?
  - considering the **most discriminative variables**…

    … while accounting for compactness (dispersion)
  - how to assess the discriminative power of a variable?
    - **information gain** (*coming next*!)

- Decision tree learning
  - variable with highest discriminative power against target is selected and fixed as the root node
  - for each possible value on the node, create a value-conditional dataset and learn a subtree
  - stop tree growing when instances in the conditional dataset are correctly classified
    or no more variables available

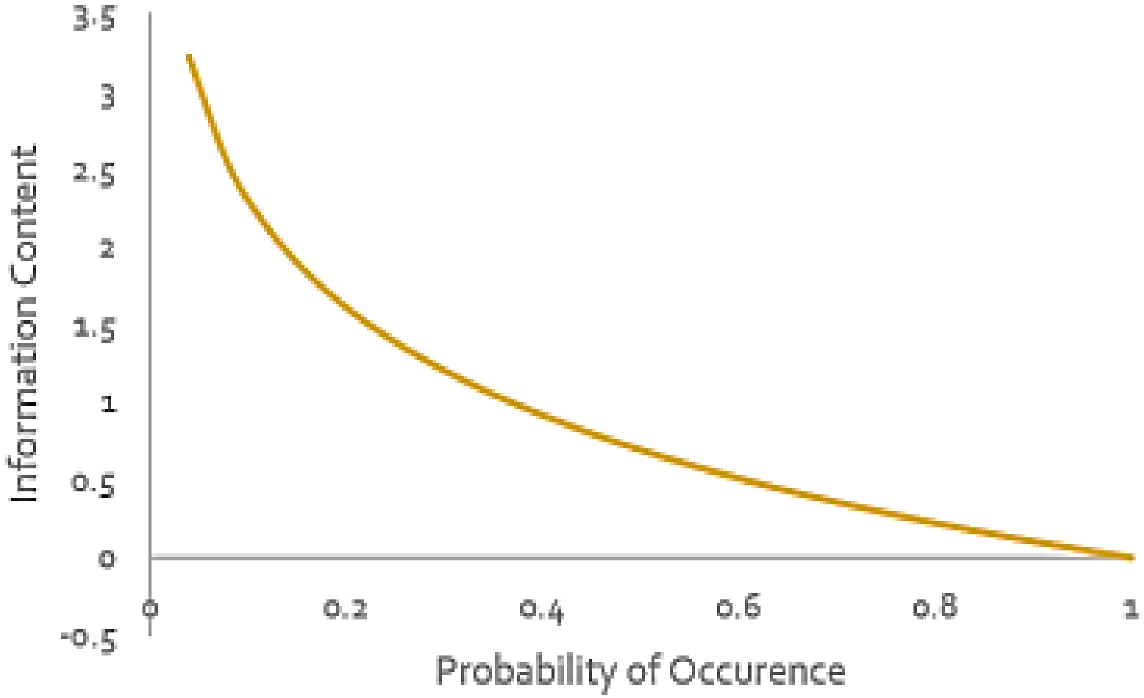
Information theory

- ***"Dog bites man"***
  - no surprise
  - quite common
  - not very informative

- ***"Man bites dog"***
  - most unusual
  - seldom happens
  - worth a headline!

- **Information** inversely related to **probability**
  - on logarithmic scale:

$$I = \log(1/p) = -\log(p)$$



Information gain (1/3)

- Given a universe of messages $M = \{m_1, m_2, \dots, m_n\}$ and a probability $p(m_i)$ for the occurrence of each message, the **information** content (also called **entropy**) is given by

$$H(M) = \sum_{i=1}^{n} -p(m_i)\log_2(p(m_i))$$

- The **credit risk** in the loan table has following information
  - $p(risk = high) = 6/14, p(risk = moderate) = 3/14, p(risk = low) = 5/14$
  - hence...

$$H(credit\ risk) = -\frac{6}{14}\log_2\frac{6}{14} - \frac{3}{14}\log_2\frac{3}{14} - \frac{5}{14}\log_2\frac{5}{14} = 1.531 bits$$

Information gain (2/3)

- Information needed to complete the tree: weighted average of information content of each subtree
  - let X be the training set and a target variable z (e.g. *risk*)
  - if variable y_j (e.g. *income*) has k values, X can be divided into **subsets** $\{X_1, X_2, .., X_k\}$ according to y_j
  - expected information needed to complete the tree after making y_j root

$$H(z \mid y_j) = \sum_{i=1}^{k} \frac{|X_i|}{|X|} H(z \mid X_i)$$

- **Information gain**
  - the amount of information needed to complete the classification after performing the test

$$IG(y_j) = H(z) - H(z \mid y_j)$$

Information gain (3/3)

- In the credit risk table, we make *income* the property tested at the root
  - makes the division into $X_1 = \{\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_{11}\}, X_2 = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_{12}, \mathbf{x}_{14}\}$ and $X_3 = \{\mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}, \mathbf{x}_{13}\}$

$$H(risk \mid income) = \frac{4}{14} I(X_1) + \frac{4}{14} I(X_2) + \frac{6}{14} I(X_3) = \frac{4}{14} 0 + \frac{4}{14} 1.0 + \frac{6}{14} 0.65 = 0.564 bits$$

$$IG(income) = H(risk) - H(risk \mid income) = 1.531 - 0.564 = 0.967\ bits$$

$$IG(credit\ history) = 0.266$$

$$IG(debt) = 0.581$$

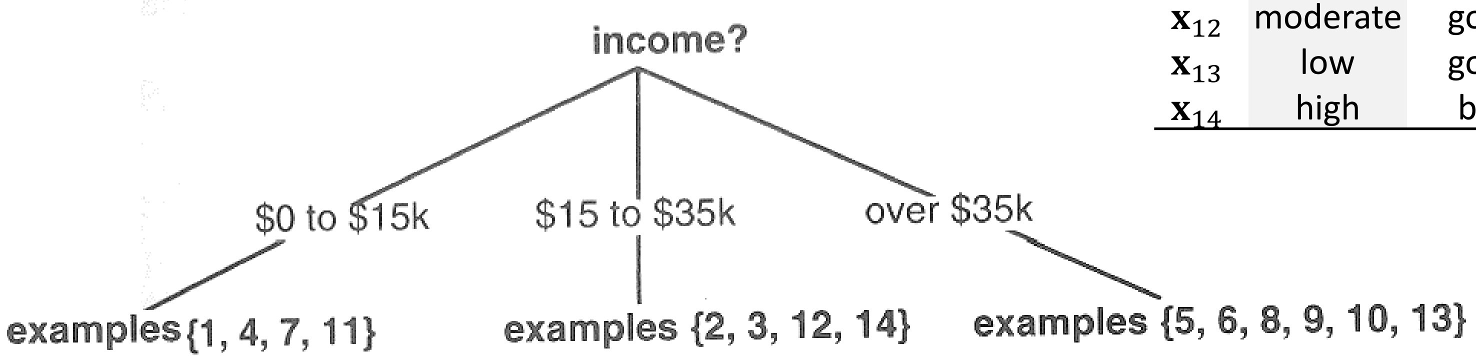
$$IG(collateral) = 0.756$$

  - *income* provides the greatest information gain, hence it is select as the root of the tree
  - the algorithm continues to apply this analysis recursively to each *subtree*, until completion

Tree learning: example

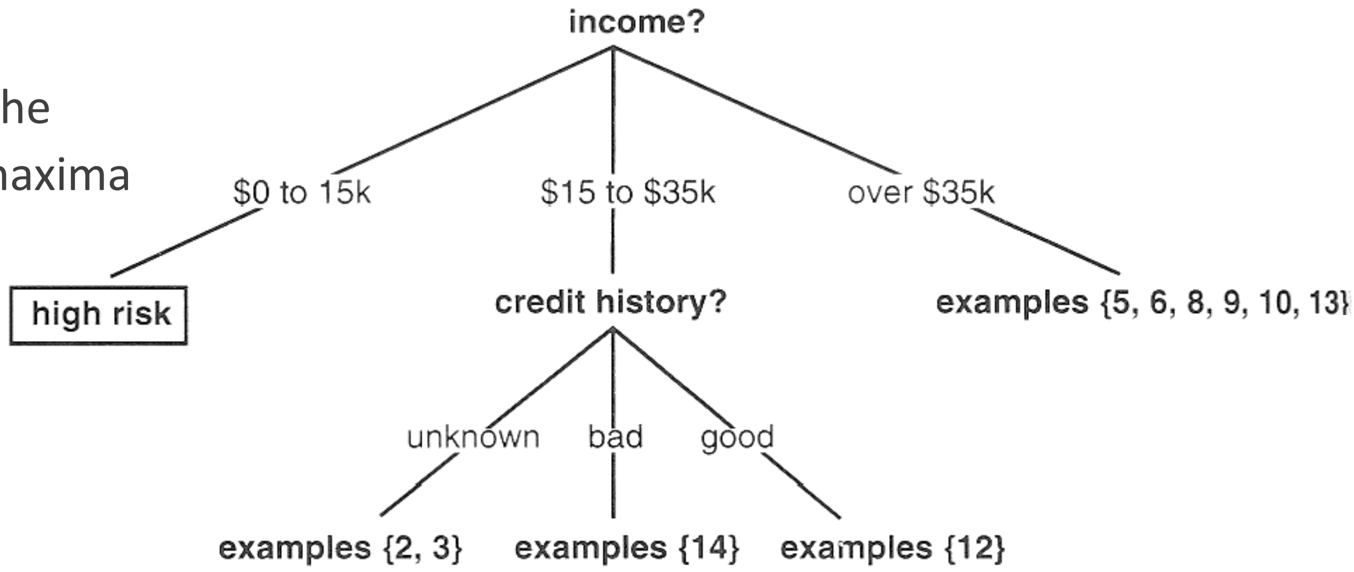
- Recovering credit risk assessment
  - income has the highest information gain
    - selected as root
    - resulting in three data partitions $X_1 = \{\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_{11}\}$, $X_2 = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_{12}, \mathbf{x}_{14}\}$ and $X_3 = \{\mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}, \mathbf{x}_{13}\}$
    - restart the process for each partition

	risk	credit history	debt	collateral	income
$\mathbf{x}_1$	high	bad	high	none	\$0-\$15k
$\mathbf{x}_2$	high	unknown	high	none	\$15k-\$35k
$\mathbf{x}_3$	moderate	unknown	low	none	\$15k-\$35k
$\mathbf{x}_4$	high	unknown	low	none	\$0-\$15k
$\mathbf{x}_5$	low	unknown	low	none	>\$35k
$\mathbf{x}_6$	low	unknown	low	adequate	>\$35k
$\mathbf{x}_7$	high	bad	low	none	\$0-\$15k
$\mathbf{x}_8$	moderate	bad	low	adequate	>\$35k
$\mathbf{x}_9$	low	good	low	none	>\$35k
$\mathbf{x}_{10}$	low	good	high	adequate	>\$35k
$\mathbf{x}_{11}$	high	good	high	none	\$0-\$15k
$\mathbf{x}_{12}$	moderate	good	high	none	\$15k-\$35k
$\mathbf{x}_{13}$	low	good	high	none	>\$35k
$\mathbf{x}_{14}$	high	bad	high	none	\$15k-\$35k



Tree learning: example

- partition $X_1 = \{\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_{11}\}$ consists entirely of high-risk individuals, a class leaf is created
- credit history has the highest IG for partition $X_2 = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_{12}, \mathbf{x}_{14}\}$
  - selected as the root of the subtree
  - partition is further divided into $\{\mathbf{x}_2, \mathbf{x}_3\}$, $\{\mathbf{x}_{14}\}$ and $\{\mathbf{x}_{12}\}$
- this is a form of hill climbing in the space of all possible trees using a heuristic function
  - note that it does not guarantee to find the smallest decision tree, can find a local maxima



Outline

- **Decision trees**
  - associative learning
  - information gain
  - **continuous variables**
  - addressing overfitting
  - variants: ID3, C4.5, CART
- Advanced aspects
  - ensembles
  - pattern mining

Continuous input variables

- Problem? Previous principles only applicable to discrete variables
  – how to handle **continuous variables**?

- Solutions:
  – **variable discretization**
    – e.g. income in the credit risk example
  – leave numeric values as-is and let the tree learning approach identify the **best binarization threshold**
    – when selecting a continuous variable, examine possible split points for the real values
    – the split point that maximizes the discriminative power (information gain) is taken as a candidate

Continuous input variables

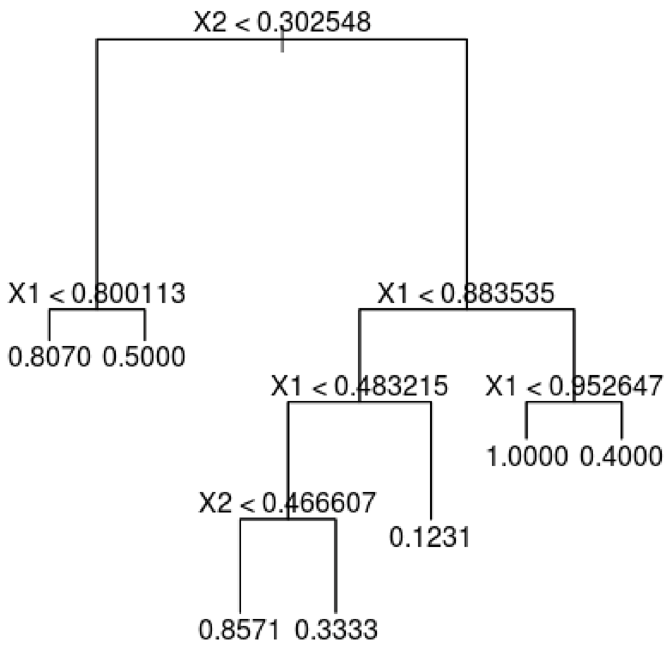
- Principle
  - select the variable y_j whose ***splitting point*** Θ produces the greatest separation in the target
    - $y_j = \Theta$ is called a "split"
    - if $y_j < \Theta$ then send the data to the left; otherwise, to the right
  - now repeat same process on these two "nodes"

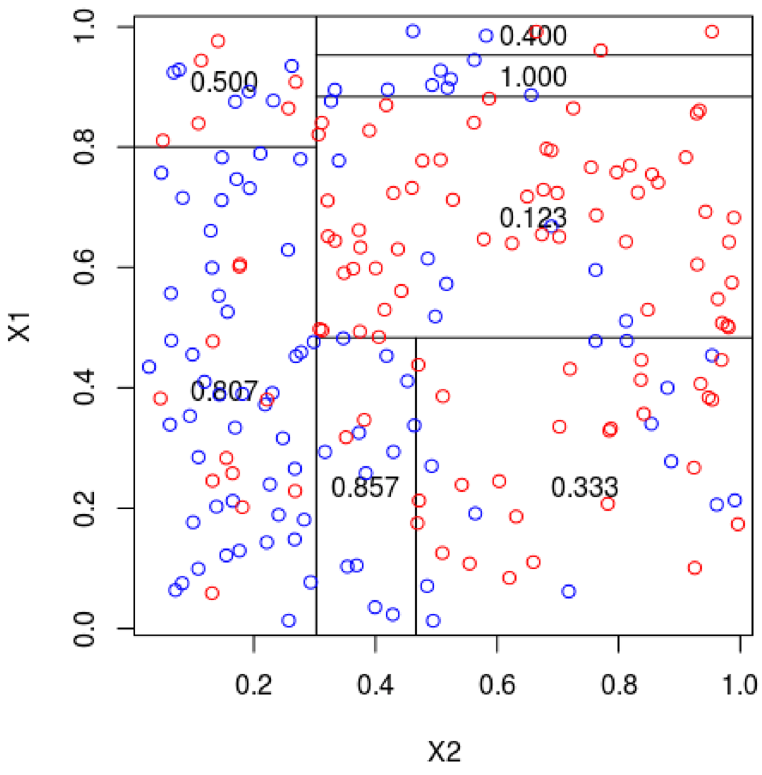
- Example:
  - the best split in age is between 59 and 65
  - the best split in BMI is between 28 and 33
  - considering both splitting ranges, age has the highest IG

age	BMI	hospitalization
33	17	Y
65	33	Y
68	35	Y
19	28	N
44	37	N
53	25	N
59	22	N

Decision tree regressors: numeric targets

- Principles to handle a numeric target?
  - can we recover splitting points from input towards a continuous output variable? How?





Outline

- **Decision trees**
  - associative learning
  - information gain
  - continuous variables
  - **addressing overfitting**
  - variants: ID3, C4.5, CART
- Advanced aspects
  - ensembles
  - pattern mining

Overfiting

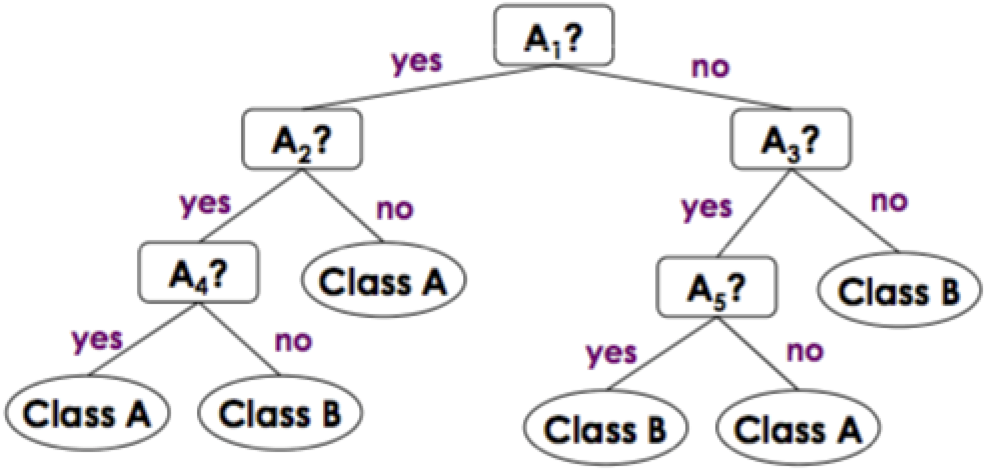
- Previous learning principles for decision trees aim at correctly classifying training observations
  - understandably, there might be no such decision tree
    - e.g. two observations with same features yet different outcomes

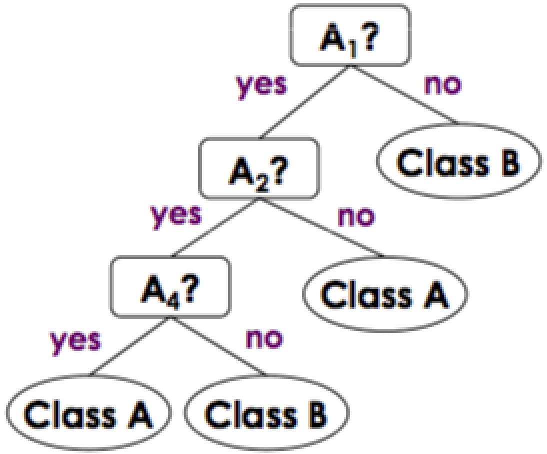
- This can produce trees that **overfit** the training data
  - **noise** in the data
  - insufficient training observations to generalize

- *Recall*: we say that a hypothesis/model **overfits** the training data when…
  - some other hypothesis/model performs worse on the training observation
    yet performs better on other observations (beyond the training set)

Overfitting: pruning

- **Avoiding overfitting**?
  - **stop growing** when data split can no longer be make with enough statistical confidence
    - e.g. impose a **minimum number of observations** on internal nodes or leafs
    - e.g. impose maximum tree **depth**
  - grow full tree then post-prune (**pruning**)
    - remove least reliable branches

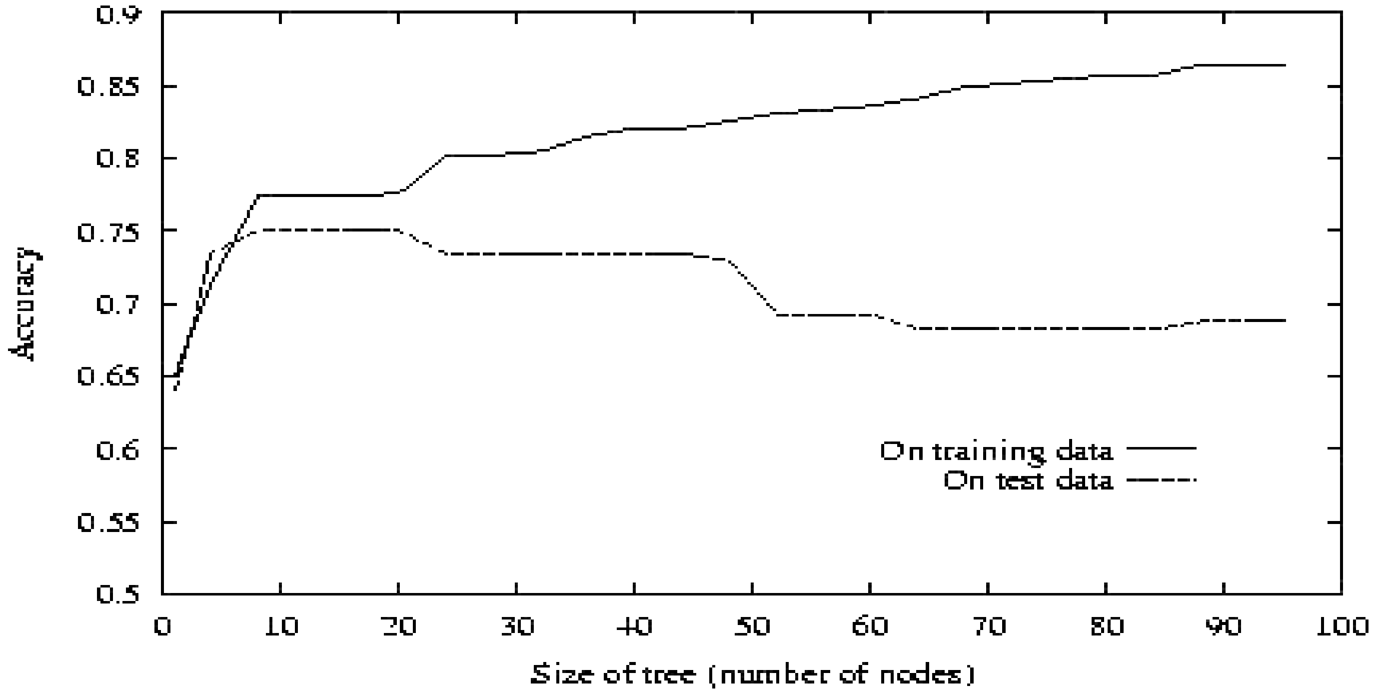




Overfitting

- How to select **best tree**?
  – measure performance over training data
  – measure performance over separate validation data

- Example: best tree size?



Outline

- **Decision trees**
  - associative learning
  - information gain
  - continuous variables
  - addressing overfitting
  - **variants: ID3, C4.5, CART**
- Advanced aspects
  - ensembles
  - pattern mining

Decision tree algorithms: ID3, C4.5, CART

- ID3 (by Quinlan) is the oldest decision tree learner
  - C4.5 and C5.0 are improved versions by the same author

- On how to **select variables**
  - ID3 and C.4.5 use entropy-based criteria to pick features
    - highest information gain in ID3 and highest gain ratio in C4.5
  - CART uses Gini impurity instead of information gain
    - binary splits are considered even for variables with +2 cardinality

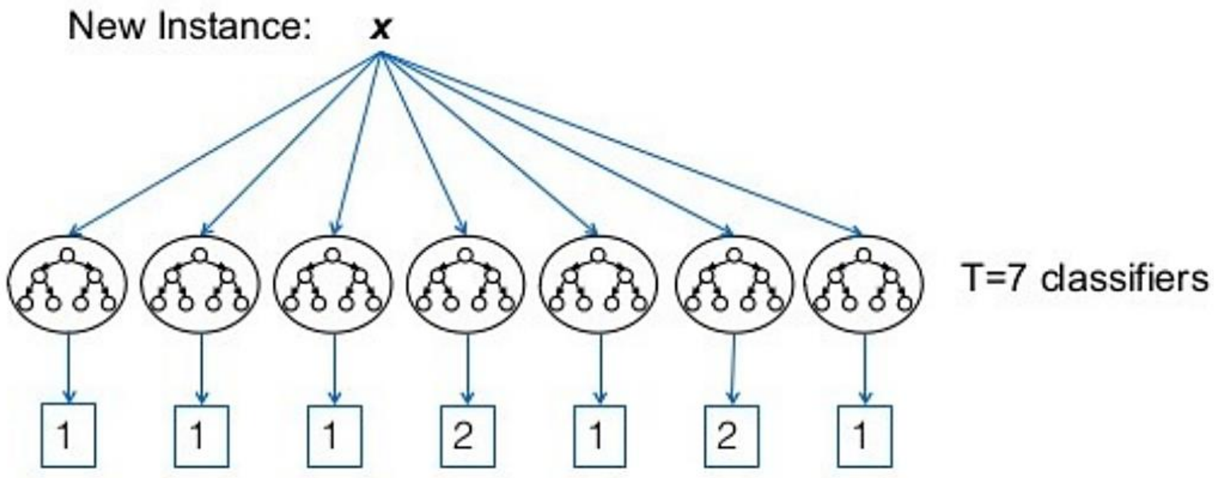
- On how to **handle continuous variables**
  - ID3 and C4.5 depend on continuous variable discretization
  - CART finds optimal splitting point on real-valued variables
    - only binary splitting supported

Outline

- Decision trees
  - associative learning
  - information gain
  - continuous variables
  - addressing overfitting
  - variants: ID3, C4.5, CART
- **Advanced aspects**
  - **ensembles**
  - pattern mining

Ensembles

- Generate many predictors and combine them to get a final prediction
  - decision can be given a simple or weighted voting step
  - simple estimators: mode (classification) or median/mean (regression)
- Generally perform better than individual predictors



Ensembles

- Important to generate diverse predictors from the available data

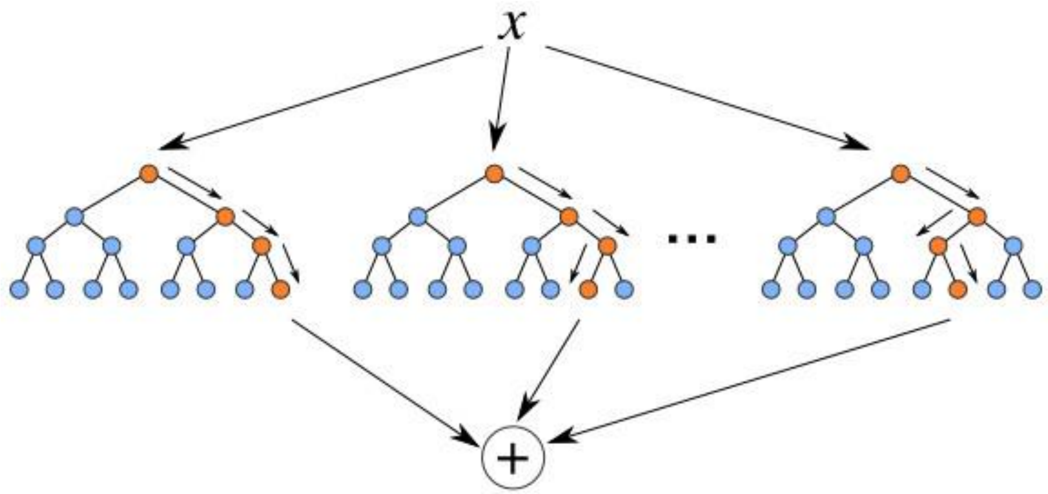
- **Principles** to generate a **diverse base of predictors**
  - use modified versions of the training data to train the predictors
    - **resampling the dataset**: multiple samples of the available data
    - select **different subsets of variables** (subspace selection)
  - introduce **changes in the learning** algorithms
    - different parameterizations
    - different learning approaches

Tree ensembles

- Generally the greater the randomization/diversification, the better the results

- **Advantages**
  - able to deal with high-dimensionality (different predictors for different subsets of variables)
  - less prone to overfitting (decision weights)
  - easy to parallelize (efficiency)

- Successful **examples**:
  - **Random Forests** (on the right)
  - **XGBoost**

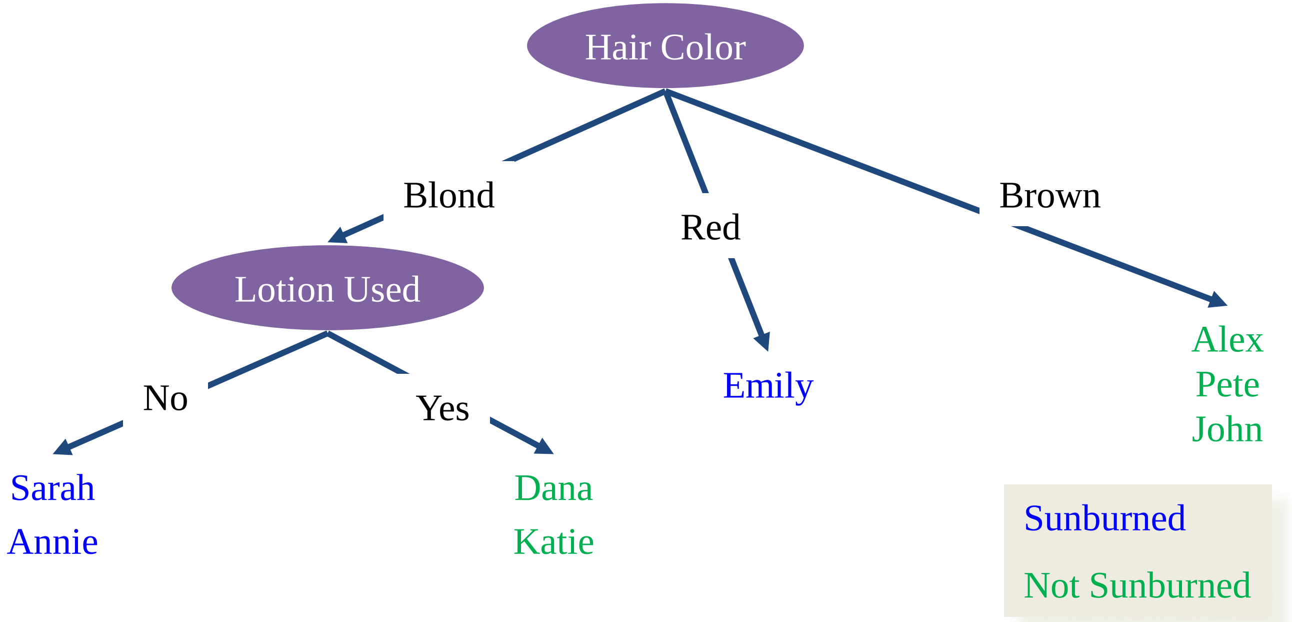


Outline

- Decision trees
  - associative learning
  - information gain
  - continuous variables
  - addressing overfitting
  - variants: ID3, C4.5, CART
- **Advanced aspects**
  - ensembles
  - **pattern mining**

From trees to association rules

Independent Attributes / Condition Attributes					Dependent Attributes / Decision Attributes
Name	**Hair**	**Height**	**Weight**	**Lotion**	**Result**
Sarah	blonde	average	light	no	sunburned (positive)
Dana	blonde	tall	average	yes	none (negative)
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none



If the person's hair is blonde
and the person uses lotion
then nothing happens

If the person's hair color is blonde
and the person uses no lotion
then the person turns red

If the person's hair color is red
then the person turns red

If the person's hair color is brown
then nothing happens

Discriminative patterns

- A decision tree path from root to leaf is an association rule

$$R\colon\ A \Rightarrow B$$

  – where A is the antecedent (set of features) and B is the consequent (set of features or outcomes)
  – if B is an outcome of interest (e.g. class), R is also termed **discriminative pattern**

- Post-manipulation
  – some rules can be reduced (check first rule)
  – unnecessary rules should be eliminated
  – default rule can be included for wider coverage

If the person uses lotion
then nothing happens

If the person's hair color is brown
then nothing happens

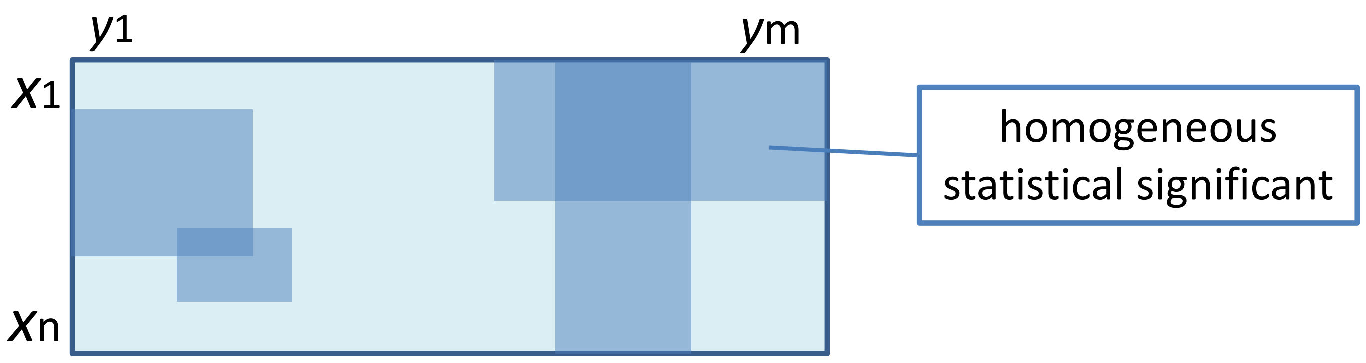
If no other rule applies
then the person turns red

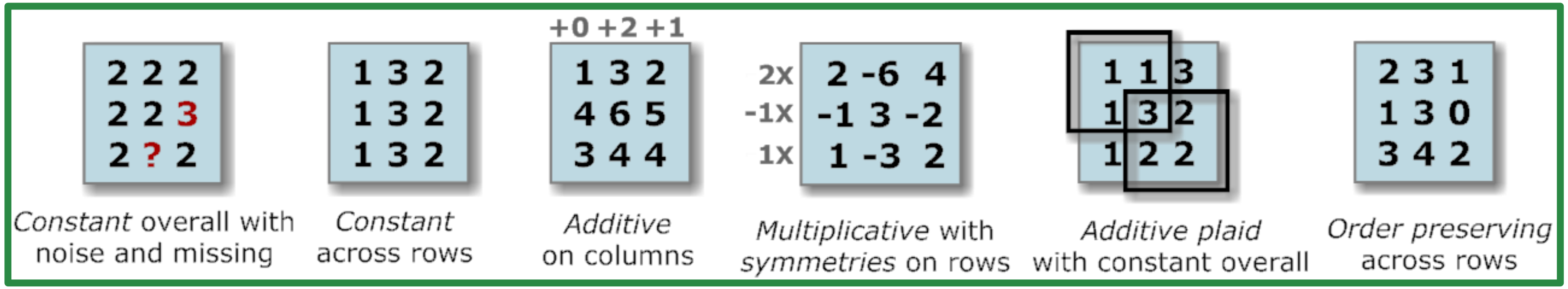
Patterns in real-valued data

- Pattern mining methods are inherently prepared to find patterns in discrete data (e.g., $\{y_1 = C, y_3 = A\} \Rightarrow c$)
  - **problem**? How to find patterns in real-valued data?
  - **solutions**:
    - data discretization
    - biclustering: patterns in real-valued data generally referred as biclusters

- Given a multivariate dataset with a set of observations X, variables Y:
  - a **bicluster** is a *subspace,* $B = (I, J)$
    - $I \subset X$ is a subset of observations and $J \subset Y$ is a subset of variables
  - the **biclustering task** aims to identify a set of biclusters $\boldsymbol{B} = \{B_1, .., B_s\}$ such that each bicluster B_i satisfies specific criteria of ***homogeneity*** and ***statistical significance***

Patterns in real-valued data





Applications

- **Social networks:** communities with shared interests, correlated activity (X=Y=individuals)
- **Text data:** content-related documents (X=documents, Y=features)
- **(e-)commerce:** browsing patterns (X=users, Y=webpage accesses)
- **Education**: performance analysis (X=students/professors, Y=topics/features)
- **Financial/trading:** profitable trading points (X=buy and sell signals, Y=stock market ratios)
- **Collaborative filtering:** groups of users with shared preferences (X=users, Y=items/actions)
- **Omic data**: biological processes and pathways (X=genes/proteins/metabolites, Y=conditions)
- **Physiological data**: patients with shared local patterns (X=signals, Y=features)
- **Clinical data**: patient groups and risk profiles (X=individuals, Y=clinical features)

Outline

- Decision trees
  - associative learning
  - information gain
  - continuous variables
  - addressing overfitting
  - variants: ID3, C4.5, CART
- Advanced aspects
  - ensembles
  - pattern mining

Thank You

rmch@tecnico.ulisboa.pt
andreas.wichert@tecnico.ulisboa.pt