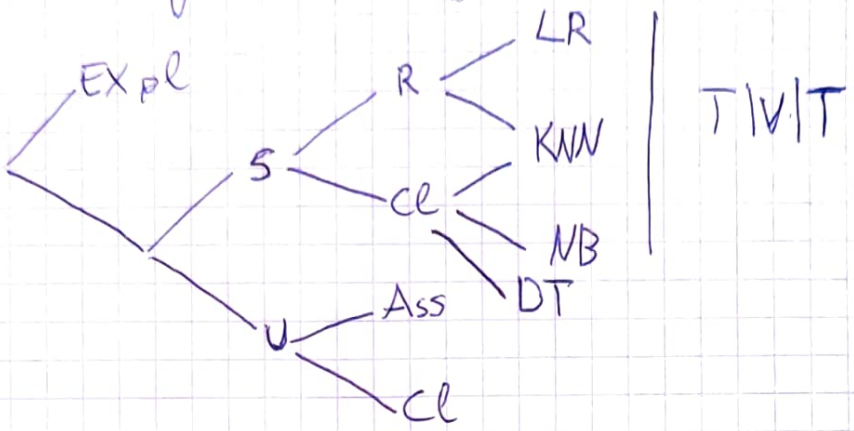
Data Mining: Data -> Information



- Unsupervised: No "ground truth"
  - Association Rules; Clustering
- Supervised: labeled "target"
  - Classification: Discrete categorical target
    - KNN, Naïve Bayes, Decision tree
  - Regression: Continuous Numeric target
    - Linear Regression

-> Supervised Evaluation -> Val Confusion Matrix

Chapter 1

Introduction

A short introduction into data mining. [jdavis] [rroels]

1.1 Definition

What is data mining?

(Semi-)automated exploration and analysis of large datasets to discover meaningfull patterns and relations.

These patterns should be:

- Valid: hold on new data with some certainty
- Novel: non obvious to the system
- Useful: should be possible to act on the itemset
- Understandable: humans should be able to interpret the patterns

Simply stated:

1. Understand the data
2. Extract knowledge from the data
3. Make predictions about the future

1.2 Popular

Why has data mining become popular?

Technology has greatly improved the last decades and more data is available.

- Technology made it possible:
  - Storage is larger and cheaper
  - Improvements in computing power
  - Improvements in the science (machine learning, data storage techniques, ..
- Datasets created this need:

- Online text sources: medline, wikipedia, ..
- Web search engines: google, yandex, baidu, dmoz, ..
- Retail transaction data: ebay, amazon, mastercard, hebbes, ..
- Goverment surveillance programs: echelon, carnivore, einstein, tempest, stingray, prism, five eyes, ..
- Commercial surveillance programs: google, facebook, twitter, microsoft, ..

1.3 Motivation

We have lots of data, with often information "hidden" in the data that is not readily evident to discover.

- Human analysts take weeks to discover useful information
- Much of the data is never analyzed at all.

"We're drowning in information, but starving for knowledge" (John Naisbett)

1.4 Useful

1.4.1 Science

- Data collected, streamed and stored
  - Remote sensors on a satellite
  - Microarrays generating gene expression data
  - Scientific simulations
- Traditional techniques infeasible for raw data
- Data mining helps scientists to
  - Classifying and segmenting data
  - From hypotheses
  - Find hidden patterns and correlations

EXAMPLES

- Medical world: determine the probability of diseases
- Email: classification of spam vs non-spam
- IDS: do these packages constitute an attack

1.4.2 Commerce

Many companies collect and store data.

- Search-engines: click data
- Stores: purchases records
- Banks: credit card transactions

- Insurance: history, transactions
- ..

There a strong competition between these companies, while computers are cheap and powerful. Data mining can help provide better, customized services:

- Better search results
- Target advertising
- Viral marketing
- Manage inventory
- Dedecting fraud
- ..

Or (simple) targeted questions:

- Who will accept a certain offer?
- What is the average expenditure for these persons?

1.5 Hybrid

Data mining draws from many disciples: database, machine learning, statistics, high-performance computing, visualization, information retrieval, ..

As a result: a mix of terminologies is used

- variable, characteristic, attribute, field: column in a dataset
- observation, record: row in a dataset
- output variables, target variables: dependent variables
- Input variables: predictor variables

1.5.1 Statistics

Data mining uses practical statistics while often ignoring the prerequisites

Tradional statistics hypothesize first,

- then collect the data to be analyzed;
- it is often model-oriented.

While data mining usually has no hypothesis,

- the focus in on data driven analysis of existing data,
- instead of strict mathematical models algorithms are used.

The ideas from statistics are very useful in data mining, particularly in evalution.

1.5.2 Machine learning

From a high-level view: these fields are very similar. Data mining focuses more on:

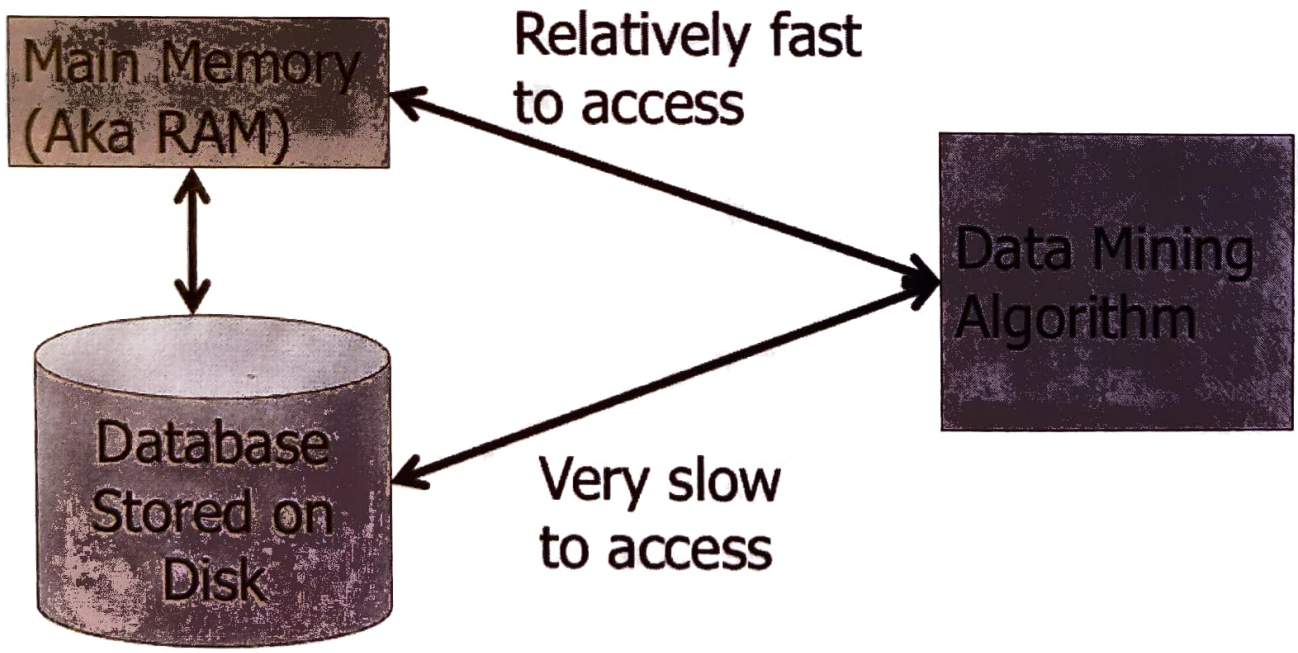
- scalability, ie, data resides in a databases
- applications
- term used more in industrial setting

Machine learning:

- more theoretical emphasis
- term more used in research and academia

1.6 Challenges

1.6.1 Scalability



The amount of storage seems inverse proportional to the cost of the type of this storage.

1.6.2 Dimensionality

Imagine instances are described by 1000 attributes, but only two are relevant to the concept. This leads to a curse of dimensionality:

- With lots of features, can end up with spurious correlations
- Nearest neighbors are easily mislead in high-dim

- Easy problems in low-dim are hard in high-dim
- Low-dim intuition doesn't apply in high-dim

Example: Points on Hypergrid

- In 1-D space: 2 NN are equidistant
- In 2-D space: 4 NN are equidistant

Spurious Correlations in Data:

- A big data mining risk is that you will "discover" patterns that are meaningless
- Bonferroni's principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap
- Another way: if more variables than examples, some variables will be correlated by chance

The danger is thus also overfitting.

1.6.2.1 Overfitting

- use training data to fit model
- problem: training data = signal + noise
- danger: modelling noise instead of signal

: overfitting

- there is always a model that does fit training data
- but: model needn't only model training data!
- as a result: model not applicable to future data
- important: At what point do you stop fitting?
- example: linear regression and regression of higher order
- the problem can also surface when too many predictor variables
  - model better fits with more var.
  - but: possible intake of variables in the model that aren't important when applying the model on future data

1.6.3 Retrospective data

Generally speaking, two types of data:

- Experimental data
- Observational data

What type of data you have influences the conclusions you can draw from it

A TRADITIONAL SCIENTIFIC EXPERIMENTAL DESIGN APPROACH:

1. Develop a hypothesis H

2. Design experiment, with controls, to test H
3. Collect data
4. Analyze results see if they confirm H

Examples: clinical trials, gene knockout experiments, etc.

These are qlso called prospective studies, there very expensive and time consuming

OBSERVATIONAL DATA

- Now we have huge observational data sets
- Examples: Web logs, customer transactions at retail stores, human genome, etc.
- Makes sense to leverage available data
  - May contain useful information
  - Very cheap to collect
- Assumptions of experimental design violated
  - How can we use such data to do science?
  - Can we do model exploration, hypothesis testing?
- Also called retrospective studies

1.6.4 Complex and heterogeneous data

Data are not simple

- Comes in different forms
- From different sources
- Collected under different conditions
- Collected with different equipment

All of these factors cause problems for analysis!

Table 1.1: Patient

PID	Gender	Birthday
P1	M	20/8/1944

Table 1.2: Drugs

PID	Date	Medication	Dose	Duration
P1	21/5/1991	RDX	700	1ms

Table 1.3: Diseases

PID	Date	Symptons	Diagnosis
P1	20/4/1994	palpitations	hypoglycemic
P1	19/3/1993	fever, aches	influenza

- Dependencies between tables
- Dependencies between rows in table

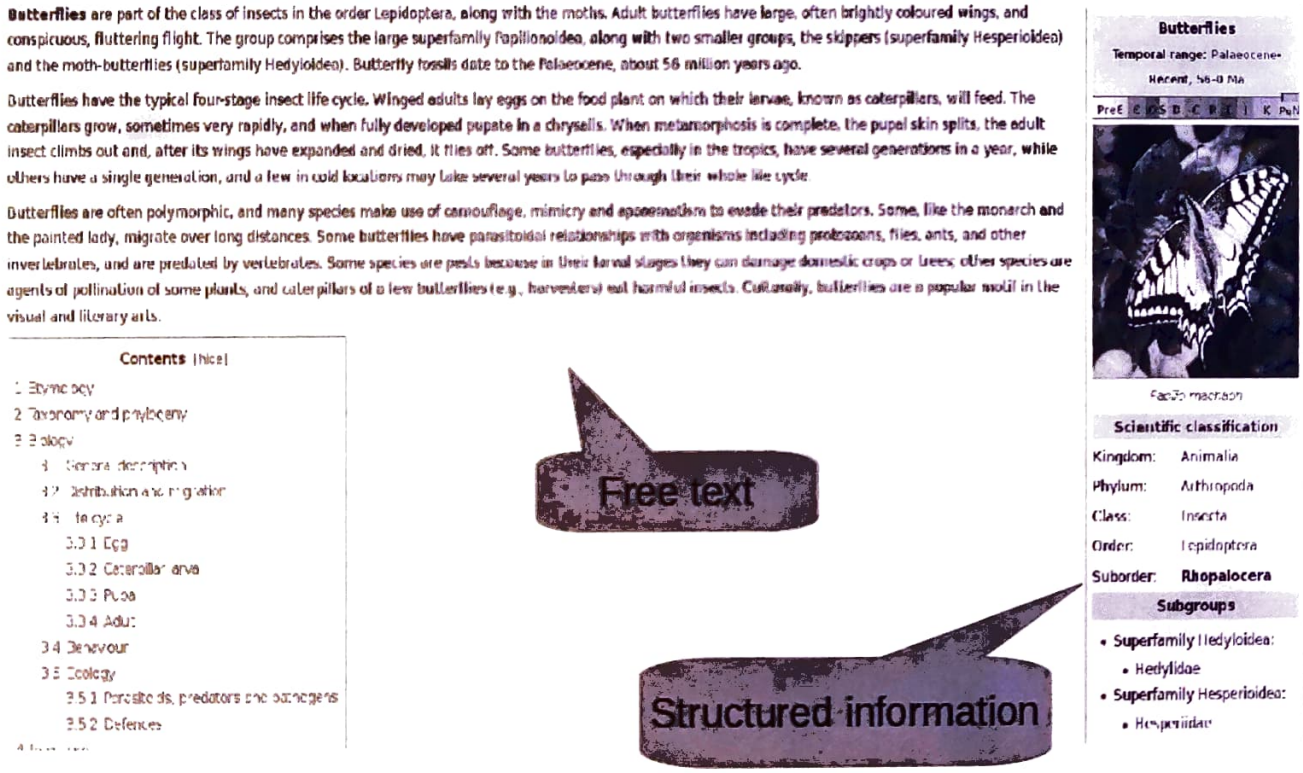


Figure 1.1: Complex Data: Semi-Structured

[bfwikipedia]

Complex Data: Unstructured "This project addresses the problem of real-world abductive inference: finding the best explanation for evidence when the latter is incomplete, noisy, possibly contradictory, and in multiple modalities (e.g., sensor networks, video, audio, text, etc.). This capability is crucial for supporting situation assessment and decision-making by military commanders in today's urban theaters of operation. Traditionally, approaches to abductive reasoning have either been based on first-order logic, by determining assumptions sufficient to deduce the observations to be explained, or based on Bayesian networks, by using probabilistic inference to compute the posterior probability of alternative explanations given a set of observations. Both of these

approaches have significant limitations. The logical approach is unable to reason under uncertainty and estimate the likelihood of alternative explanations. The Bayes-net approach is unable to handle structured representations, and therefore is incapable of effectively reasoning about situations involving multiple entities with various relations between them."

Heterogeneous Data Data are different

- Companies use different databases schemas
- Different terminology for the same concept
- Dates stored differently
- Full name vs. nick names

Gives raise to complicated problems

- Schema matching
- Ontology matching
- Entity resolution

1.6.5 Data quality

Data often missing or incomplete

- Forms have optional fields
- People are intentionally misleading

Many measurements are inexact

- How can Google measure a user's satisfaction with the search results?
- Did you display the right ads? Would someone have clicked a different ad?

Known biases in data

To improve the data quality u might have to remove, create or transform variables.

1.6.5.1 Outliers

A small amount of the total data that can really influence your inferences.

- can indicate an error in measurement/input
- can have large effects on the model
- possibly find explanation for outlier

The involvement of a domain expert is important!

1.6.5.2 Missing data (missing values)

- if few records with missing data: remove records
- but suppose 30 variables, for every record, for every variable 5% probability missing value: probability missing value in given record?
- possibly replace by e.g. average (= „imputed value‟)
  - downside: no new information for that variable
  - upside: information of other variables will not be lost!

1.6.5.3 Normalizing data

- comparing records can only be done on the same scale
- otherwise 1 variable can dominate
  - normalize
  - (value – average)/standard deviation
  - scale: "number of standard deviations of the average"
- whether necessary or not depends on the technique

1.6.5.4 How many records needed for training model?

In case of supervised learning: e.g. Delmater and Hancock for classification

Number of records = at least 6*M*N

M = number of classes

N = number of variables

Thus preferably minimise the number of variables

1.6.6 Data ownership and distribution

Data is valuable and people and companies often not interested in sharing

Can get into trouble for using existing data

- E.g., most Websites don't want you to crawl them
- Statistics from sports matches, etc.
- Images: rights often retained by photographer (or agency)

1.6.7 Privacy preservation

Many privacy issues

- Often underestimated by technology people
- Privacy breaches get much attention
  - Massachusetts health records
  - AOL search logs
- Unclear what measures need to be taken to ensure that data is not identifiable
- Can data mining be performed such that results guarantee the privacy of individuals?
- Correlations/predictors may be discriminatory

1.7.1.1 Random Sampling

- data mining: many variables and records
- restrictions on processing capacity, software

sampling: smaller data set

1.7.1.2 Oversampling of rare events

- frequent in classification: often 0, few 1
  - possibly random sample with few 1
  - little information on records class 1
  - difficult to train a good model for classification 0 and 1
- solution: oversample class 1
- can be important:
  - missing a 1 can be costly! (see attack computer network, no attack history in the sample)
  - erroneously classifying 0 as 1 is less harmful!

1.7.2 Exploratory Analysis

Know Your Data.

- First step in any data mining problem
- Inspect the data and try to get a feel for what is going on, that is, debug the data
- Clean and prepare the data (cf supra challenges)
- Getting a sense for
  - What challenges exist
  - What is possible/realistic

What Should You Look For?

- Good to look at simple statistics of
  - Number of variables
  - Size of data
  - Missing values
  - Skew
- For each attribute, look at
  - Discrete: number of possible values, are they ordered, frequency of each value, etc.
  - Numeric: mean, min, max, etc.

1.6.8 Streaming Data

Data is not static

- Stock prices
- News tickers
- Cameras
- Sensor networks

Actually have so much data that it isn't possible to permanently store all of it

- What should we store?
- How can we make use of the data we see?

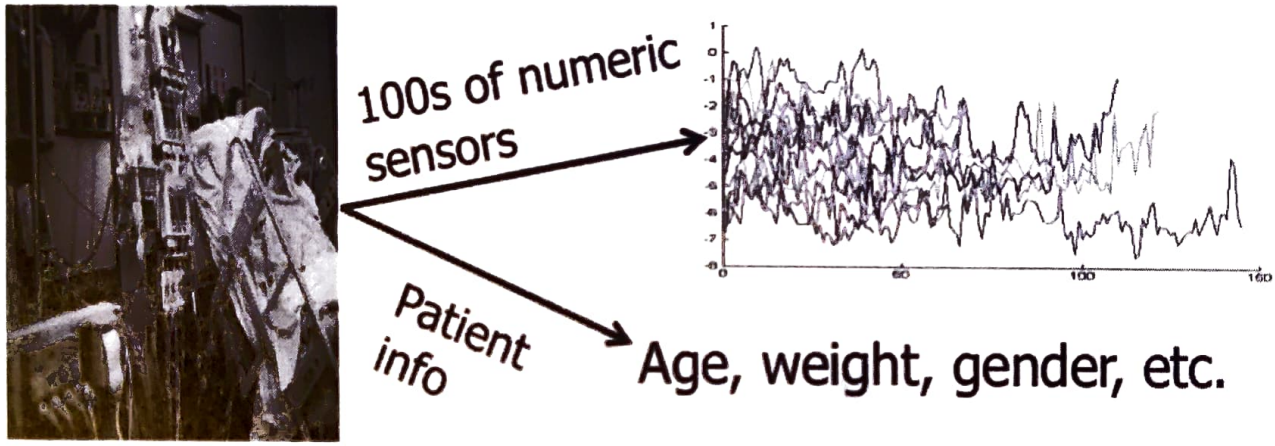


Figure 1.2: Example: Data Streams

1.7 Data mining tasks

A selected overview of some data mining tasks.

1. If possible, clearly define the purpose of the analysis.
2. After deciding upon the task, choose the appropiate technique.
3. Bringing model in production / translate into decision rules

1.7.1 Datasets

Building dataset for analysis

- Random sampling from large database
- Combine data from databases
- Internal and external data

We will often work with samples:

1.7.3 Descriptive modeling

Build model that can be

- Describe or summarize the data
- Simulate the data
- Model the process that generated the data

Techniques

- Clustering
- Density estimation/probabilistic models

1.7.3.1 Clustering

The grouping of similar data.

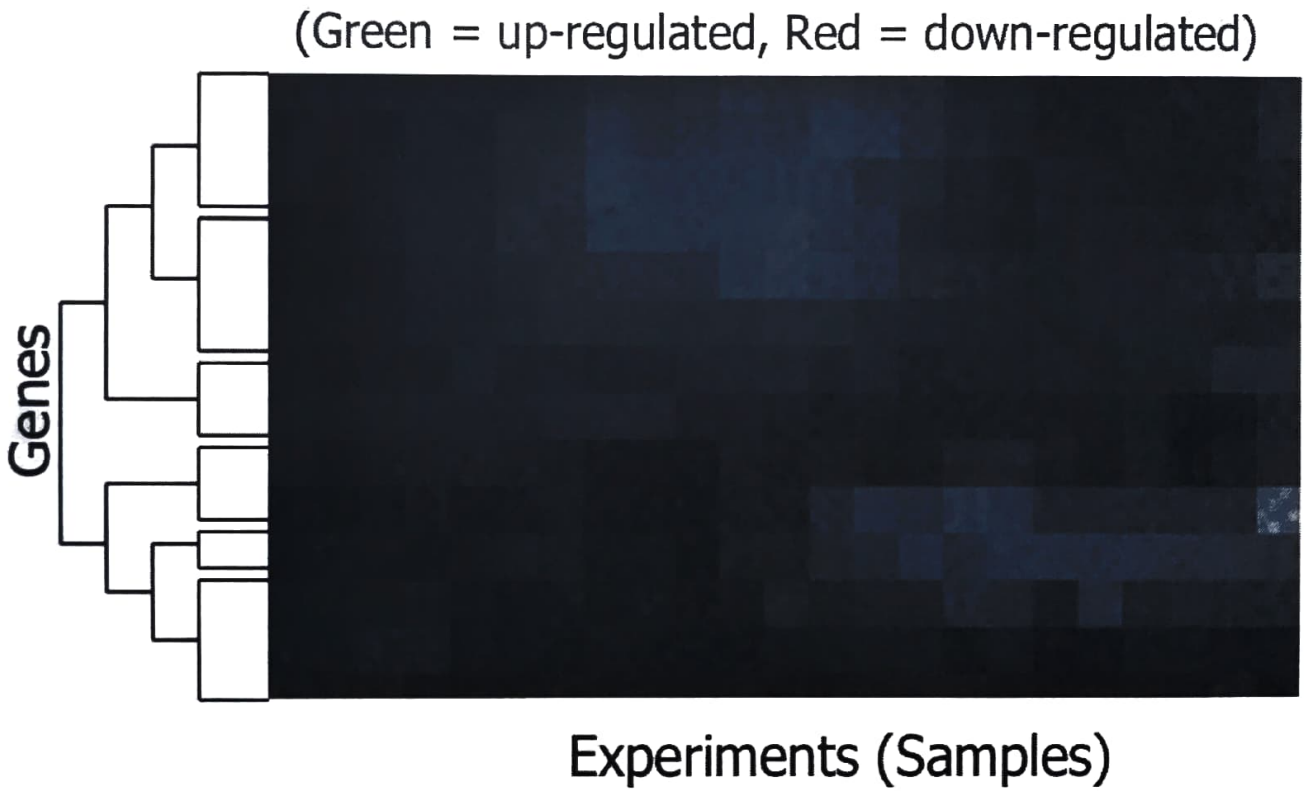


Figure 1.3: Example: Gene Expression

EXAMPLE: DOCUMENT CLUSTERING

- Web search is not great
  - System perspective: covers small coverage of Web (<16%), dead links, out of date pages
  - IR perspective: very short queries, hugedatabase, novice users
- One solution: document clustering

- User receives many (200 - 5000) documents from Web search engine
- Group documents in clusters by topic
- Present clusters as interface

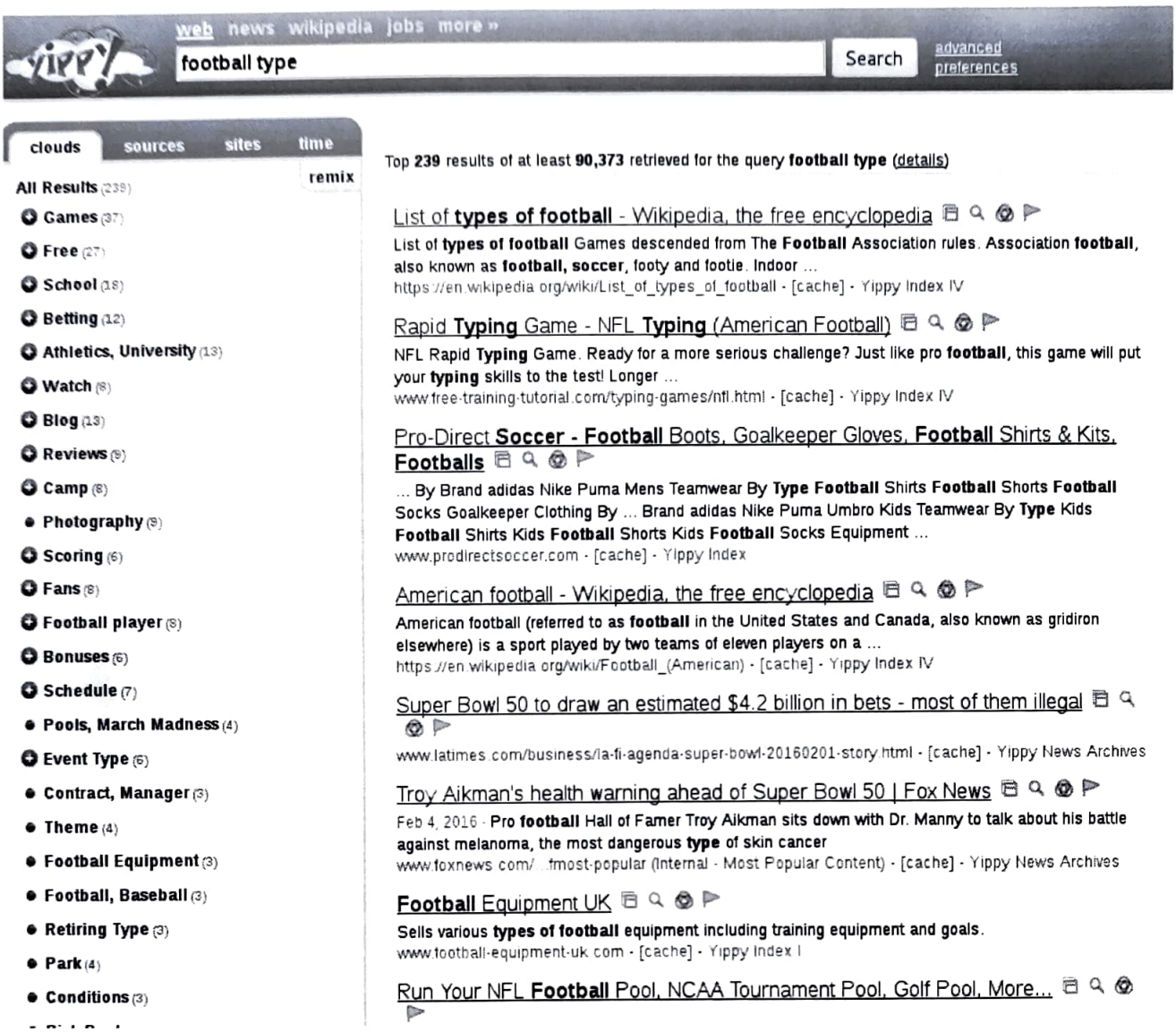
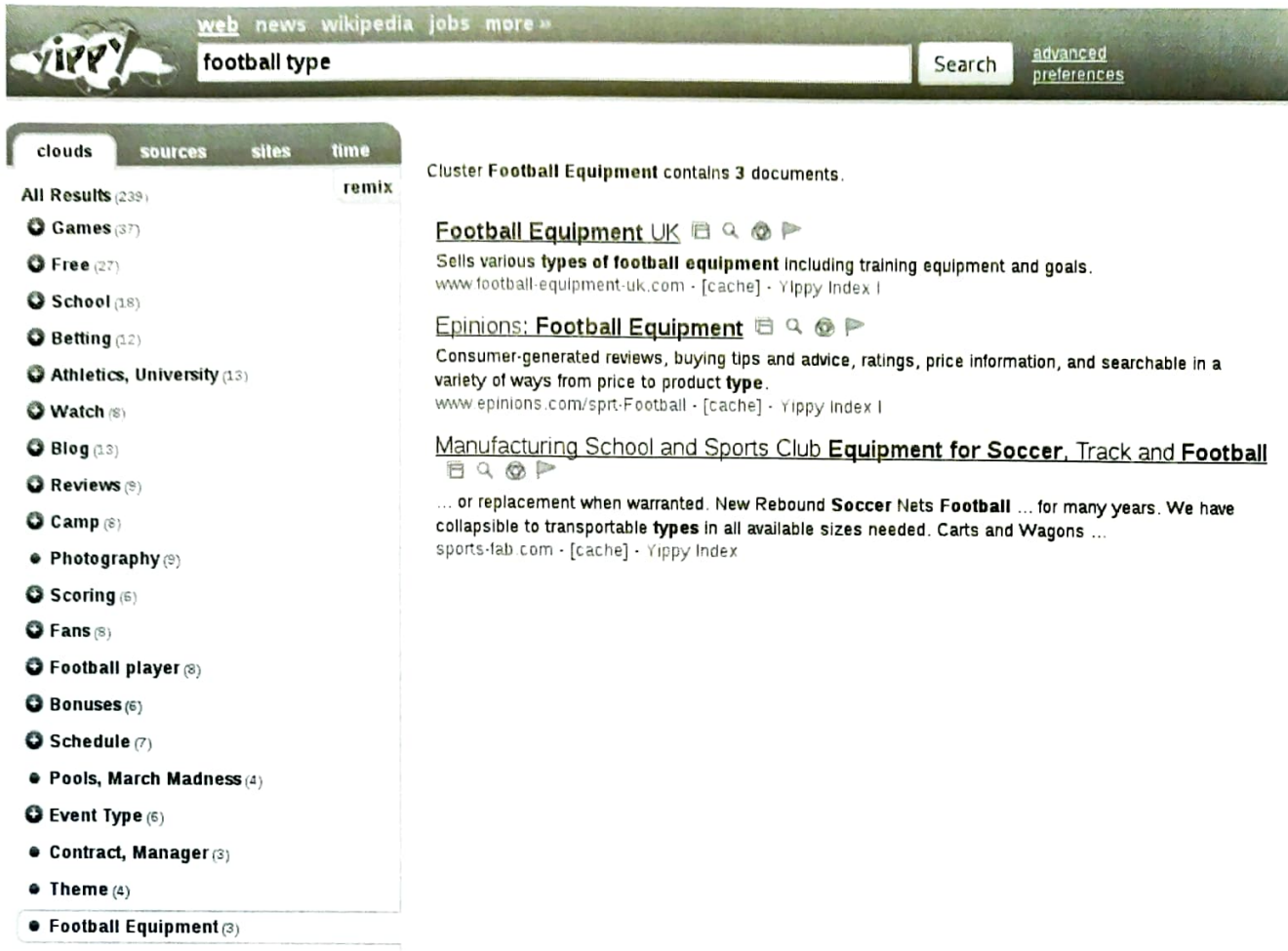


Figure 1.4: Example: Clustered websearch



1.7.3.2 Probabilistic Models

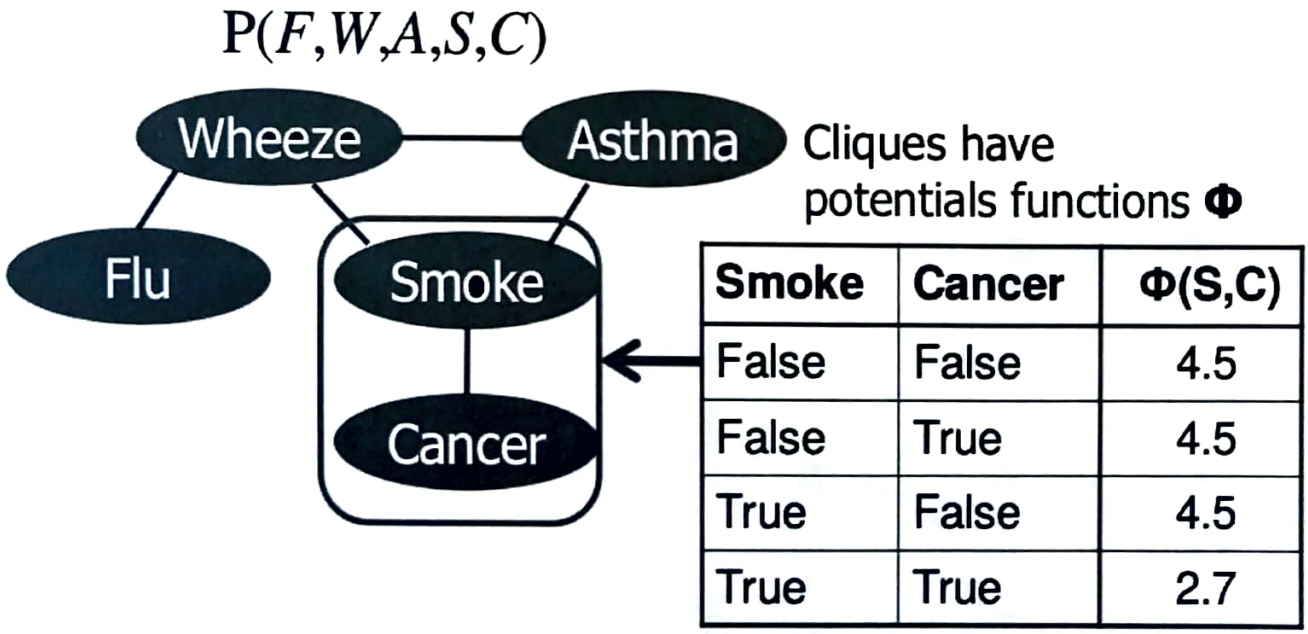


Figure 1.5: Density Estimation

Applications: Diagnosis, prediction, recommendations, and much more!

1.7.4 Predictive Modeling

- Inductive learning: Use some variables X to predict the future values of variable
  - Classification: Y is discrete, also called a class, so we try to predict the class by a fitting model on the data with know class values. Examples are: spam vs non spam, attack on the network vs no attack
  - Regression: Y is continuous (instead of discrete). Example: prediction of viewer ratings
  - Probability estimation: Prob(Y = y)
- Learn the relationship between X and Y: Approximate the function mapping configurations of $X \rightarrow Y$ Many machine learning & statistic algorithms
- Often focus on accuracy not comprehensibility

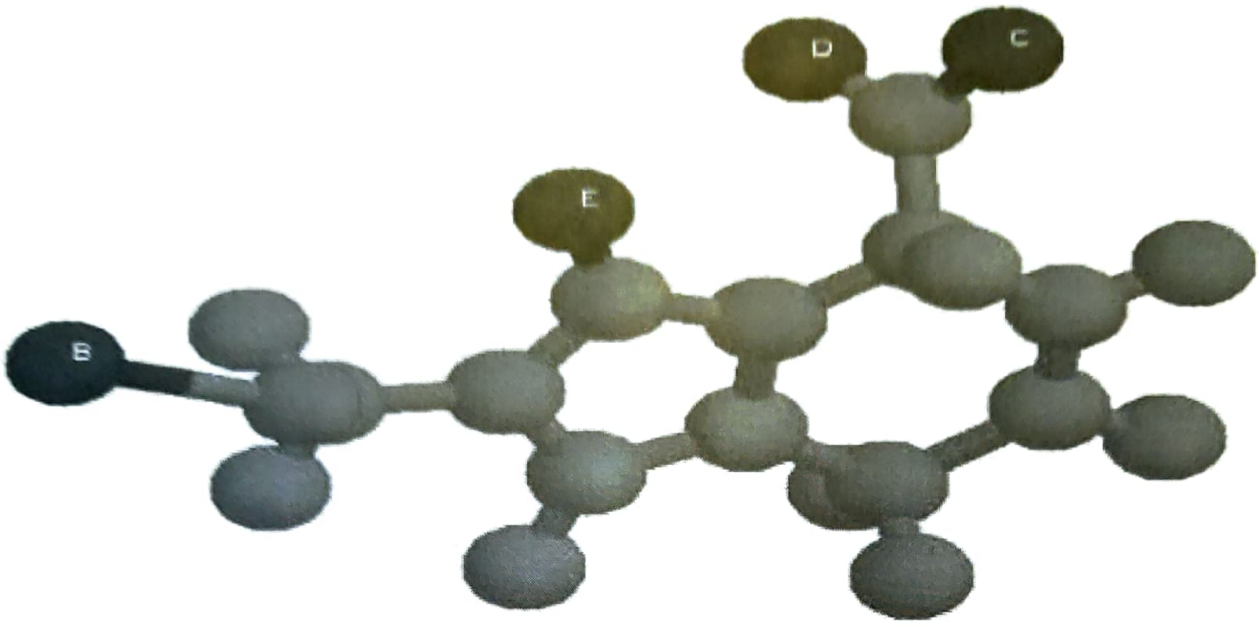
EXAMPLE: COLLABORATIVE FILTERING

- Given database of user preferences, predict preference of new user.
- Example: Predict new movies you will like based on
  - your past preferences
  - others with similar past preferences
  - their preferences for the new movies
- Example: Predict what books/CDs a person may want to buy (and suggest it, or give discounts to tempt customer)

EXAMPLE: FRAUD DETECTION

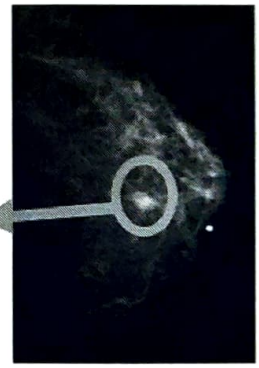
- Credit card fraud is common and costly.
- Learn model to predict the probability that each transaction is fraudulent
  - Use historical data of known fraud/non-fraud
  - Investigate high probability transactions
- Issues
  - Lots of preprocessing, feature construction
  - Tradeoff the costs between false alarms and missed detections

Example: Drug Discovery Given: 3-dimensional structures of molecules and their known binding affinities to target



Do: Learn a model to predict binding affinity for new molecules to this target

Patient	Date	Calcification Fine/Linear	...	Mass Size	Loc	Cancer
P1	5/02	Present		3mm	RU4	No



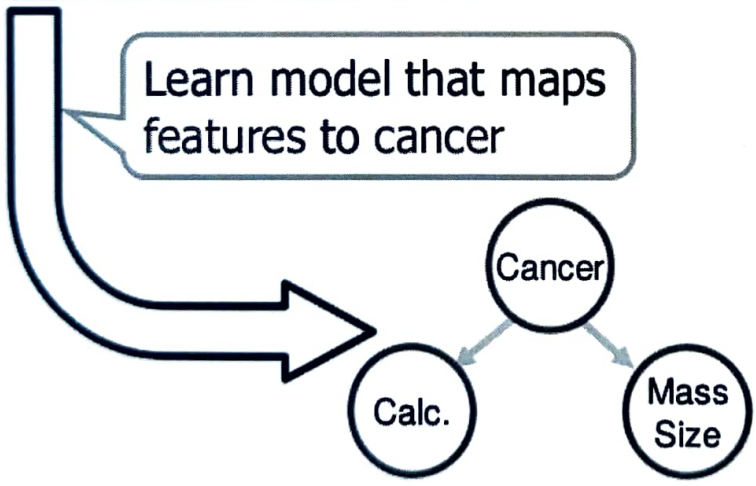


Figure 1.6: Example: Medical Diagnoses

1.7.4.1 Evaluate model

The classification accuracy is basically the most general democratic score, the closer to 1, the better. Recall is called sensitivity in the medical world.

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Figure 1.7: Table: classification scores

[sswikipedia]

1.7.5 Discovering Patterns

- Goal is to discover interesting "local" patterns in the data not characterize the data globally
- Focus on finding human-interpretable patterns that describe the data
- Techniques
  - Item-set mining
  - Pattern mining
  - Sequence mining

Example: DNA Sequences --- ADACABDABAABBDDBCADDDDBCDDBCCBBCC DADADAADABDBBDABABBCD-DDCDDABDCBBD BDBCBBABBBCBBABCBBACBBDBAACCADDADB DBBCBBCCBBBDCABDDBBADDBBBBCCD-ABBA BCDACBCABABCCBACBDABDDDADAABADCDC CDBBCDBDADDCCBBCDBAADADBCAAAADBDC ADB-DBBBCDCCBCCDCADAADACABDABAABBD DBCADDDDBCDDBCCBBCCDADADACCCDABAA BBCBDBDBADBBB-BCDADABABBDACDCDDDBB CDBBCBBCCDABCADDADBACBBBCCDBAAADD DBDDCABABCACDCBAAADCAD-DADAABBACCB ---

Example: DNA Sequences pattern --- ADACABDABAABBDDBCADDDDBCDDBC*CBBCC* DADADAADABDBBD-ABABBCDDDCDDABDCBBD BDBCBBABBBCBBABCBBACBBDBAACCADDADB DBB*CBBCC*BBBDCABDDBBADDI BCDACBCABABCCBACBDABDDDADAABADCDC CDBBCDBDADDCCBBCDBAADADBCAAAADBDC ADBDBBBCD-CCBCCDCADAADACABDABAABBD DBCADDDDBCDDBC*CBBCC*DADADACCCDABAA BBCBDBDBADBBBBC-DADABABBDACDCDDDBB CDBB*CBBCC*DABCADDADBACBBBCCDBAAADD DBDDCABABCACDCBAAADCAD-DADAABBACCB ---

EXAMPLE: NBA DATA

- NBA logs all play by play information
  - Which players are in the game
  - Shots attempts
  - Etc.
- Questions: Which lineups work well?
  - Offensive efficiency
  - Defensive efficiency
  - Etc. See: http://www.synergysportstech.com/

Table 1.4: Example: Frequent Itemsets

Items
Bread, Cheese, Wine
Chips, Salsa, Wine
Bread, Cheese, Wine
Buns, Hamburger Meat, Ketchup
Cheese, Wine
Chips, Coke, Salsa
Hamburger Meat, Ketchup
Beer, Chips, Salsa
Bread, Cheese, Wine

- Items co-purchased
  - Cheese and Wine
  - Chips and Salsa
  - Hamburger Meat and Ketchup
- Associations:

So how should we arrange our products in the supermarkt, online and offline.

1.8 Machine Learning

A comparison between data mining with machine learning.

Broadly speaking: 4 different types of learning

- Supervised learning: Inductive learning, i.e., classification, prediction
  - A training set, where the value of the target variable is know, is used to train the model. So the model learns from the training set.
  - Then we tune the modeling
  - Use the model for prediction target variable in new data_semi
  - Example: regression
- Semi-supervised learning: Inductive learning, but not all labels known
- Unsupervised learning: Given data, i.e., examples, but not the labels
  - There is no target variable to make predictions
  - The model can't learn from a training set with a know target variables
  - Example: clustering
- Reinforcement learning: Take an action, receive a reward, learn how to act

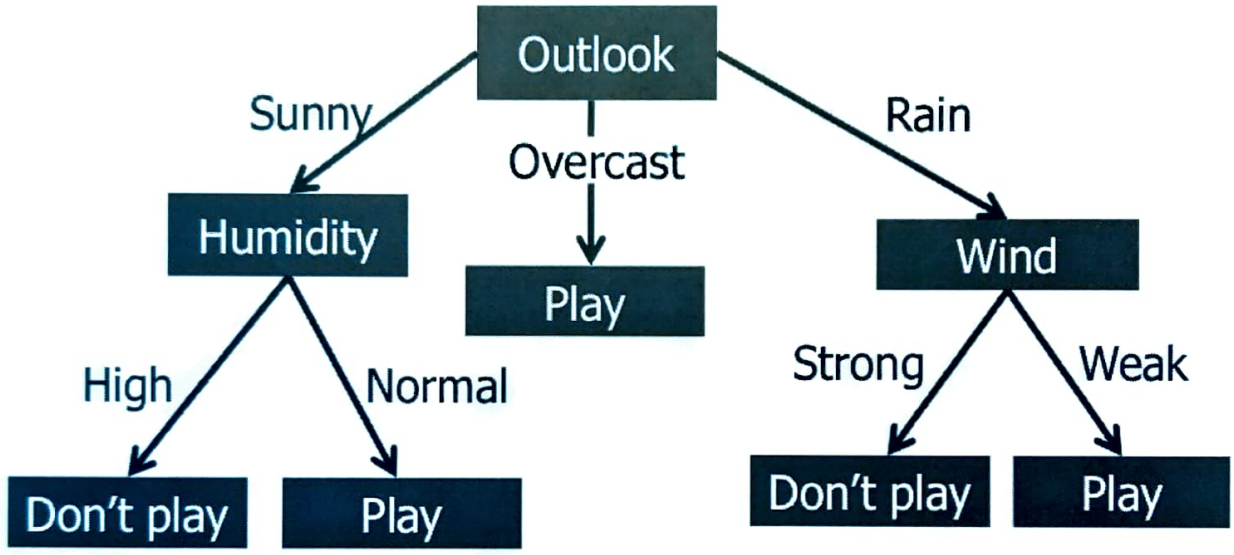
The first three learning methods are important for data mining.

The following techniques are also present in machine learning

- Decision trees

- Clustering
- Ensembles
- Rule induction
- Naïve Bayes
- Neural networks
- Instance-based learning
- ..

1.8.1 Decision trees



1.8.2 Rule Learning

Propositional rule sets Outlook = Sunny *and* Humidity = Normal **then** Play Outlook = Overcast **then** Play Outlook = Rain *and* Wind = Weak **then** Play

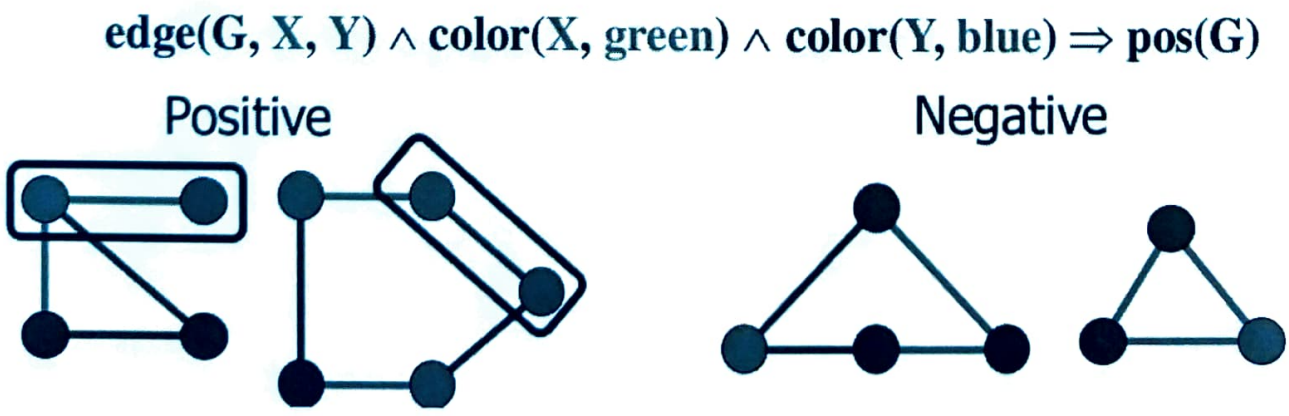
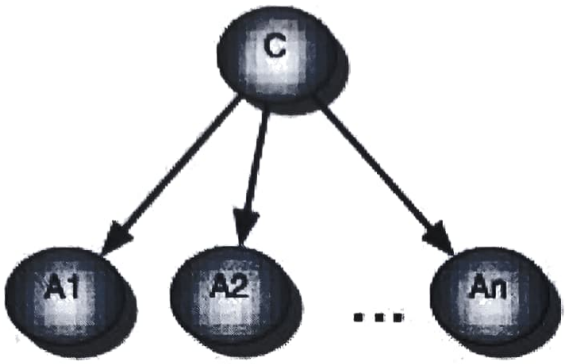


Figure 1.8: First-order rule sets

1.8.3 Nave Bayes

Example [kkersting]

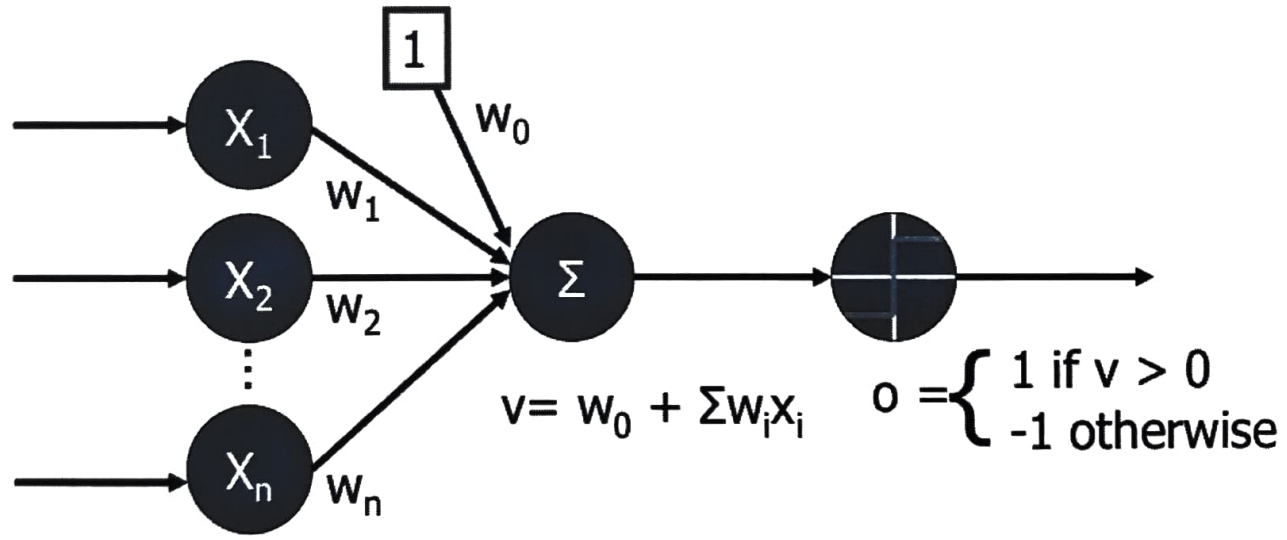


Node i stores $P(A_i|POS) and P(A_i|NEG)$

$Odds = \frac{P(A_1|Pos)*...*P(A_n|Pos)*P(Pos)}{P(A_1|Neg)*...*P(A_n|Neg)*P(Neg)}$

1.8.4 Peceptron

Example:



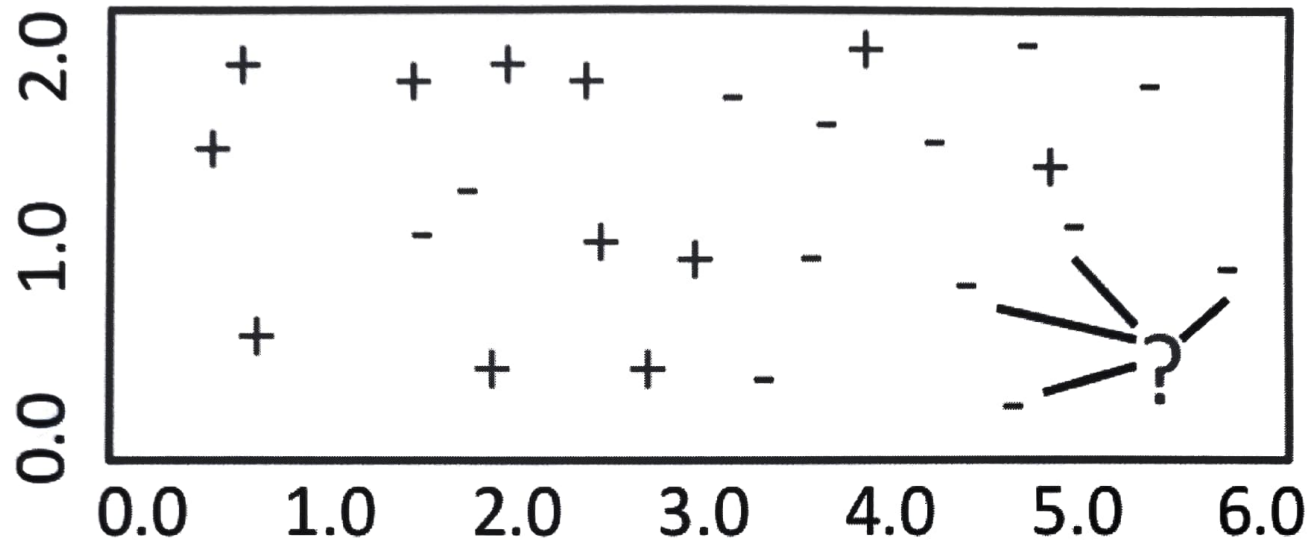
$$o(x_1,...,x_N) = \begin{cases} 1 \text{ if } w_0 + w_1x_1 + ... + w_nx_n > 0 \\ -1 \text{ otherwise} \end{cases}$$

Vector Notation

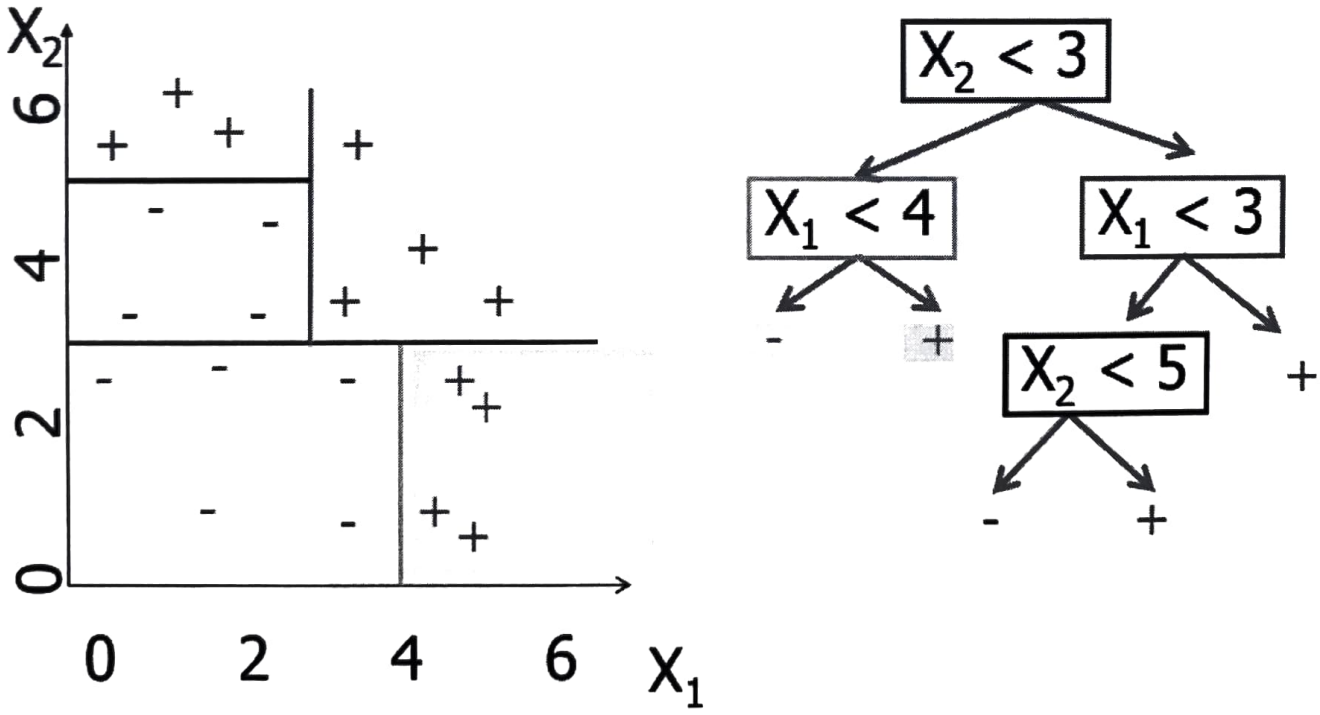
$$o(\vec{x}) = \begin{cases} 1 \text{ if } \vec{w} \cdot \vec{x} > 0 \\ -1 \text{ otherwise} \end{cases}$$

1.8.5 Instance-based Learning

- Learning ≈ memorize training examples
- Find most similar example
  - Classification: output its category
  - Regression: output its value



Label based on neighbors



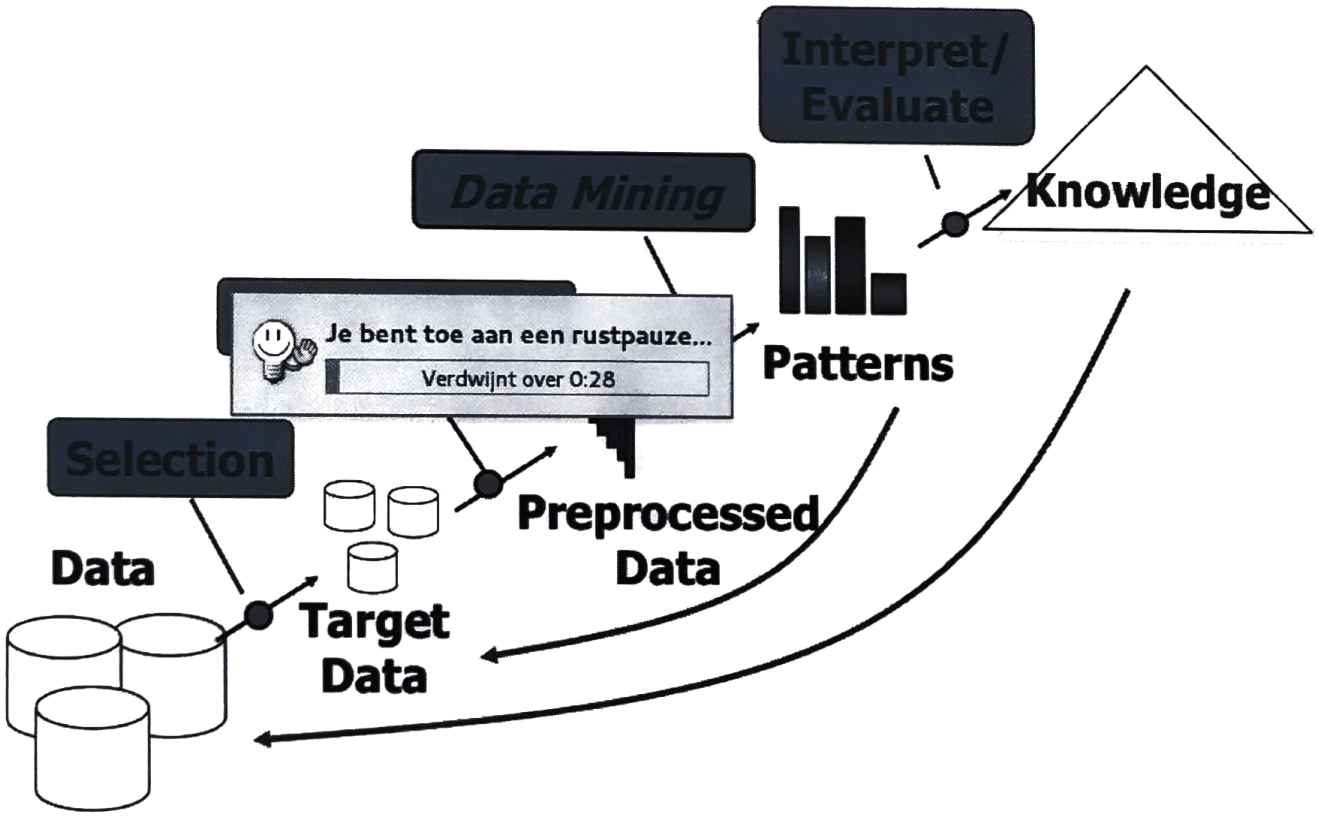
Label based on decision trees: Divide feature space into axis parallel rectangles and labels each one with one of the K classes

1.9 Concepts

- Feature construction
- Feature space
- Hypothesis space
- Forward/background feature selection
- Experimental methodology
  - Cross-validation
  - Parameter tuning
  - Accuracy, ROC, precision-recall
- ..

1.10 Data mining process

Workflow:



1.10.1 Requirements for a Data Mining System

Data mining systems should be

- Computationally sound

  - Scalability time and space complexity
  - Parallelizable, e.g., MAP-Reduce and Hadoop?
- Statistically sound
  - Are patterns meaningful?
  - Do our results generalize to new data?
- Ergonomically sound
  - Presents results in a comprehensible manner
  - Does it need 6 PhDs to run it?

1.10.2 Components of a Data Mining System

- Search
- Data management
- User interface

1.10.2.1 Representation: Data

- Feature vectors
- Relational database
- Free text
- Images
- Graphs
- Etc.

1.10.2.2 Representation: Model

- Decision trees
- Graphical models
- Rule set
- Association rules
- Graph patterns
- Sequential patterns
- Etc.

1.10.2.3 Evaluation

- Objective
  - Accuracy
  - Precision and recall
  - Cost / Utility
  - Fast
  - Etc.
- Subjective
  - Interesting
  - Novel
  - Actionable
  - Etc.

Training Set – Validation Set – Test Set Partitioning

Specifically for data mining (large datasets!)

- Training set
- use to train model
- adapt model / fit to the data
- danger: overfitting
- Validation set
- study performance of found model
- refining / tuning found model
- danger: again overfitting
- Test Set
- not always present
- compare performance best models for every technique

Partitioning at random or according to a variable Typical ratios:

- training/validation (60/40)
- training/validation/test (50/30/20)

1.10.2.4 Search

- Combinatorial optimization
  - E.g.: Greedy search
- Convex optimization
  - E.g.: Gradient descent
- Constrained search
  - E.g.: Linear programming

1.11 Summary introduction

- We live an age where large amounts of data are commonplace
- Data mining is hugely popular and hugely successful because it extracts useful information from this data
- This information comes in many forms
  - Models
  - Patterns
  - Etc.
- Data mining is challenging for many reasons

Chapter 2

Ethics

What about the ethics in using data mining? Choose one example of the previous introduction chapter and formulate according to you what would be ethical? Compare your idea with the knowledge you have from the real world.

Chapter 3

Exploratory Data Analysis

A few exercises to get started

3.1 Exercise:iris dataset -> recording 2 toledo

Use the iris dataset (File)

1. Show the data (Data Table)
   a. How many rows do we have? a150
   b. Are there missing values? aNo
2. Show the frequency table for each numeric variable (Distribution)
   a. Which numeric variable looks very promising to classify iris-setosa? $a$petal length
      i. What is the interval (ie min, max value) for this variable looking at iris-setosa? (Select rows) a[1,1.9] using the data table and sorting this column
   b. What does the probability plot show? aIt shows the probability distribution for a target variable, when looking at petal length and taking iris-setosa as target variable, it is clearly skewed, as was expected (supra)
3. Show the mean, median, 25% percentiles (Box plot)
   a. How many different petal width are there? (Select Column and then Venn diagram) a22 distinct values, u can also do this with a simple query on the data, a venn diagram is a qualitative way of representating data in sets
   b. What is the standard deviation of the petal length? a1.76
      i. Why is it less interesting to look at the overall values? What graph is more interesting? a it doesn't tell us a much as the same for each iris type apart a thus is it more interesting to group them by iris type; a which should also show us the differences better.
   c. How many rows are there for each type of iris? a50 -> w select Rows "Iris" is "Iris..."

selecteer de 3 irissen en kijk hoe lang bevat voor out: Rows = 150

3.2 Exercise:eda datasets

Use the eda dataset (eda.csv)

Looking at data quality of this dataset do an exploratory analysis and identify their interesting properties. Look at attribute distributions, relations between attributes, missing data, and so on. Try to grasp at least one pattern by using a suitable visualization. U might come back to this exercise as we progress to find more patterns. There at least 5 patterns to discover.

-> recording 3

3.3 Using your database skills

Some of the previous questions dataset could be solved more easily using your SQL skills.

1. Load the data into your own local database
    a. Verify with the appropiate queries the results obtained before

3.4 Exercise: basketball visualization

Use the shots dataset (shots.csv)

In memory of Kobe Bryant, based on this dataset. Try to visualize from where Kobe tried to score and whether or not the shot succeeded. Seeing this picture, was Kobe easy to defend?

Hints:

- U need to count how often he scored from each position
  - Create a new feature containing the X,Y coordinate (feature constructor)
  - Count how often he scored/missed (pivot table)
  - Merge the results, to get the X,Y position back (merge)

a Answer: no he scored from different positions more or less consistenly. Most shots were from beneath the basket

Chapter 4 recording 5

K Neirest Neighbours Classification

Class value unknown?: predict by using the k nearest neighbours.

When a "new" row with an unknown class value has to be classified, we look in the (training)set for examples that are closest to this "new" row. From these "close by" examples with known class values we then classify the class value of this "new" row.

This is often abriaveted as "KNN".

4.1 Questions to be answered

- How would you determine the k neirest neighbours? [rroels]
- How would you determine class value of the unknown target?

So we need to look out for:

- Metrics? (neirest)
- Number of neighbours? (k)
- Weight?
- Normelized?

→ outlier coul determin class value

class value determind by most appearing class in set

What would be the problem with a k that is too small? What would be the problem with a k that is too big?

Conclusion: using the distance to determine the weight is not a bad idea.

Other classifier: Modus (like Billy Swan, Everything's the same)

4.2 Number of Neigbhors

4.2.1 One Nearest neighbor: 1NN

Advantage: we don't need a weight function. (Why?) Graphical 2D representation: Voronoi diagram. Disavantage: one neighbor determines everything.

4.2.2 Two Nearest neighbors: 2NN

Advantage: more input Disavantage: class value might be undecided depending on the weight function.

4.2.3 All are considered Nearest neighbors: ANN

Advantage:

- even more input
- less susceptible to erroneus data in the training set

Disavantage:

- more calculations to be processed
- class might be undecided depending on the weight function
- class is fixed without a weight function

4.3 Weight function

Using the proper weight function, also called a kernel function, can be important. To illustrate this, what would happen if we used all the rows in our dataset to classify an unknown class value of a new row? Image that u use no weight function, or if you like a democratic weight function?

Examples of kernel functions:

- all are equal, there is no real weight function
- the inverse of the distances
- the inverse of the square of the distance
- based on the normal distribution

For each of these functions u can use the average using the kernel function.

What does this translate to? Let's give some examples for k=2:

- k = 2, there is no real kernel function: the prediction might be undecided
- k = 2, using the inverse of the distances: the prediction might be undecided, but is it less likely to be so
- k = 2, using the inverse of the square of the distances: the prediction might be undecided, in this case the same result as the previous example
- k = 2, based on the normal distribution: the prediction might be undecided, in this case the same result as the previous example

So choosing k wisely is also important. For example les assume there are only 2 possible class values; then k=3 is wiser than k=2.

4.4 Dimensions

As the dimensions and the size of the data might grow there are a couple of things to consider:

- VA Files: try to represent your dataset in a more compact why to reduce the size of your datasets
- Reduction of the number of dimensions: for example, do you need bmi if you know the size and the weight

4.5 Summary

The entire training set is used as the model. For each ("query") point to be classified, we search for its k nearest neighbors in the training set. The classification of the query point is some function of the labels of these k neighbors. The simplest case is when k = 1, in which case we can take the label of the query point to be the label of the nearest neighbor. [mmds]

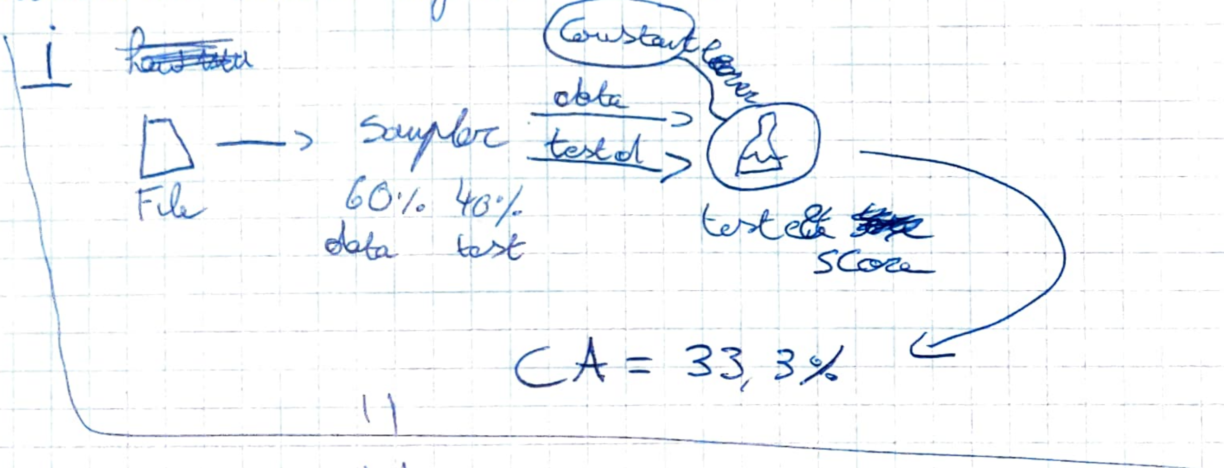
4.6 Exercise: iris dataset continued

1. Divide the data randomly in training and test dataset (60/40%) using replicable sampling and stratification (Data Sampler)
   a. Why do we use replicable sampling in these exercises? a results given are dependent of the sample, that's why we choose a fixed random seed, here called replicable deterministic sampling; with the same seed the sample should be the same. "to devide data in equal sized subsets"
   b. Why do we use stratification? What does this mean? a Stratification: try to represent the found proportions in the population in the sample. For example the there should be an equal amount of men and women in the sample; or there should be 10 times more people from Leuven than Bertem in the sample as the number of people in Leuven is 10 times bigger than Bertem. Stratification can help mediate these proportions, as they might be skewed too much otherwise. -> "take same sample every time for fair comparison"
   c. Use the modus as dumb classifier (Constant as a Learner of Test & Score, which will receive data from our sample)
      i. How much is the average classification accuracy for the test data? a 33,3%
      ii. Was this to be expected? a we have an equal number of data for each iris type in our dataset, since the data was randomly sampled using stratification, u can expect it to close to 1/3
   d. Use now KNN to classify our data (Nearest Neighbours from classification)
      i. Which distance function do u use and why? a euclidean, cm measured data for the predictor variables
      ii. Explain when to use which metric. a **euclidean**: continuous data, ok most of the time a **manhattan**: if dimensions cannot be cut through, like on manhattan island, the sum of the absolute differences for all attributes a **maximal**: or chebyshev distance, eg in dynamic logistics, how long does it take to position the crane that moves at the same speed over x and y axes a **mahalanobis**: takes into account the distributions of each feature, using the covariance matrix, can be seen as a unit-less and scale-invariant distance, so when the unit and/or scale affect your classification, u might want to preprocess your data and/or use this metric instead of euclidean
      iii. How much is the average classification accuracy for training set using 1NN? a 100% (n=1)
      iv. How much is the average classification accuracy for test set 1NN? a 96,7% (n=1)
      v. Which k gives the best classification accuracy for the test set? a several k values perform well, 7 seems the best for this specific training/validation set; in other cases we avoid multiple's of 3 and the number 2 when no weight function is used, by choosing 1 u assume that the training data has no incorrect instances
   e. Show the confusion matrix for KNN and the modus for the test set using n=2. (Confusion Matrix)
      i. How many are misclassified for the modus? a 40/60 -> good (w label)
      ii. Is is logical that this number is equal or higher than 2/3 a yes, since the majority classifier was based on the training set, and all 3 types of irises occur equally in the complete set; in (this) case of a tie, the first one is chosen
      iii. How many are misclassified for 2NN? a 3/60 with eucl uniform (2/60 using distance)
         A. Calculate the direct relation between CA (classification accuracy) and the number of correctly classified instances. a 57/60=95% (3/60 was misclassified) -> like op 2NN -> good is bad
         B. Calculate the direct relation between Recall (sensitivity or recall) and the previous question. a looking for example at iris-versicolor as a target variable, 18/20=90%, this depends on the target variable
         C. Calculate the direct relation between Precision (or positive predictive value) and the previous question. a looking for example at iris-versicolor as a target variable, 18/19=94,7%, this depends on the target variable
      iv. Which type of iris are most often misclassified? a iris-versicolor (relative and absolute in this case)
      v. Show the underlying data of the confusion matrix in a table. a don't forget to select to data in the confusion matrix

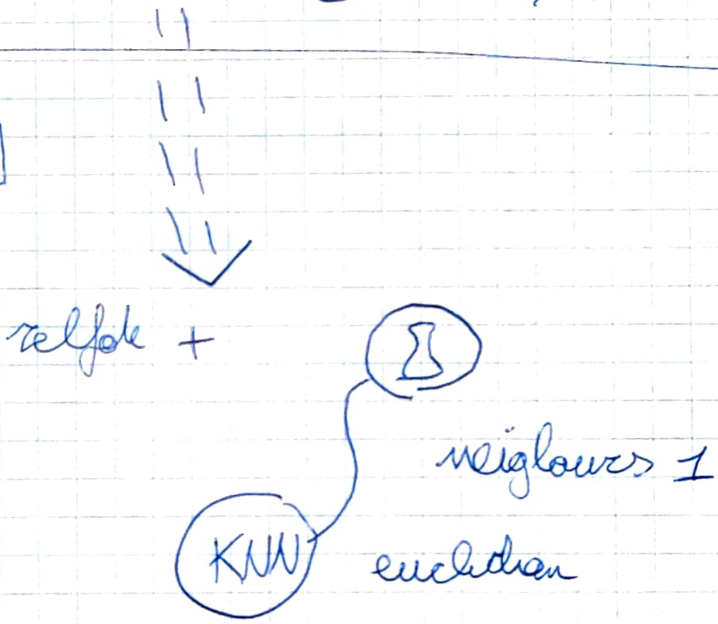
Exercise 4.6 was continued

1.c

Constant as learner of test&score ← gets data from sampler



1.d.III



1.e.III.B/C

Precision → correct type predictions / all of this/these predictions

→ Correct red pred / all red predic

Recal → all correct red pred / all actual Red

vi. Show the underlying data of the misclassified instances instances in a table and their probabilities. a select misclassified and make sure probabilities also ticked

vii. Make a plot showing this data with a color for each type or iris (Scatter Plot) a make sure the scatter plot has the information it needs, data and stats

A. Try to show plot that clearly distinguishes the 3 classes a use rank projections for optimal visualisations

B. Try to show the probability for classification according to our KNN when u hoover over the plot. a make sure the confusion matrix sends the probabilities

C. Show an interesting selection of the data on the plot as a table. a demo

f. Does the widget version of KNN use weights? a v2,7 yes; v3.3 not; u can easily check this by using K=samplesize, if you the same results as the majority classifier then the neighbors are not weigted

i. Try to write a python script using weighted KNN (python script). a u can edit the source code of the corresponding widget /orange3/Orange/widgets/classify/owknn.py by adding a combobox a or u can use Orange.classification.KNNLearner(n_neighbors=5, metric=*euclidean*, weights=*uniform*, algorithm=*auto*, preprocessors=None)

ii. Answer the same questions as above using a distance weighted KNN. a Now the ca results on the training set are 100% (as is to be expected when there no collisions) a The results differ a bit, but not much, weighted or not doesn't seem important in this case

5.1 Naive Bayes law applied

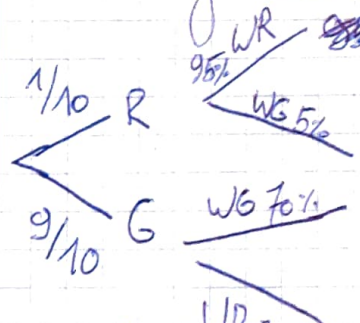
Prior probabilities

10% => red 90% => green

P(WR|R) => 95% P(WG|R) => 5%

P(WG|G) => 70% P(WR|G) => 30%

Posterior probabilities of bayslearn



naive bayes, omdraaien

$$P(R|WR) = \frac{P(R) \cdot P(WR|R)}{P(R) \cdot P(WR|R) + P(G) \cdot P(WR|G)}$$

(als witness zegt "Rood", bayslat hij Rood is)

$$= \frac{0,1 \cdot 0,95}{0,1 \cdot 0,95 + 0,9 \cdot 0,3}$$

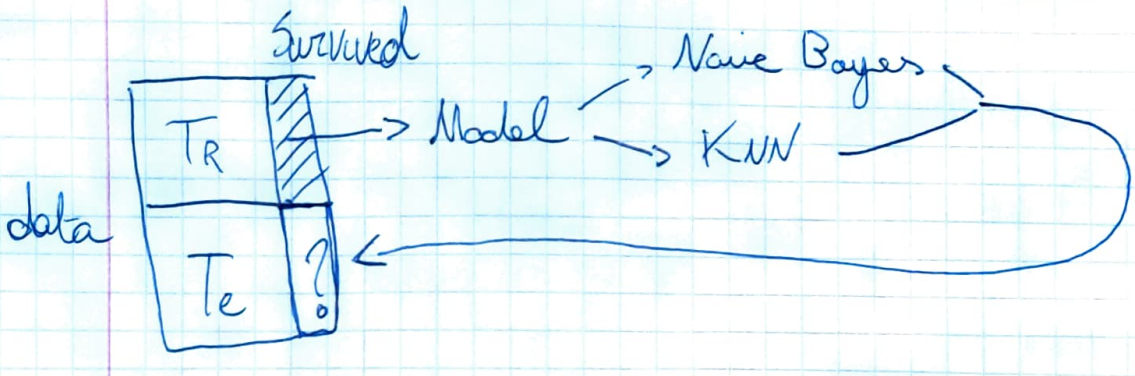
$$= 26,03\% \Rightarrow P(R|WR)$$

dus => $P(G|WR) = 100\% - 26,03\%$

= conclusie. = 74%

bayslat dat hij effectief groen is als getuige zegt dat hij rood is

Training -> Validatie -> Test



3. evaluate this classifier (test learner)
    a. on the training set a 77,1%
    b. what does sensitivity and specificity mean? Look them up. a recall is also called sensitivity or true positive rate a specificity or true negative rate (negative in this case translates to not belonging to the target class)
        i. what happens with the sensivity and the specificity if u chance the target class? a the switch, this is logical if u look at the definitions of sensitivity and specifity. a the sensitivity for target class no: 0.914 a the sensitivity for target class yes: 0.473 a the specifity for target class no: 202/427=0.473 a the specifity for target class yes: 817/894 a normally you would take target class no in the context of sens and spec
    c. on the test set
        i. how many are correctly classified in the test set? a 694 out of 880 or 78.9%
    d. on the training set a 1019 of 1321 or 77.1%
        i. what is the probability of surviving for a male adult in first class? (confusion matrix, data table, (mosiac display)) a 0%, there are no adult male survivers on our training data set
            A. is this correctly predicted by your classifier? a our class prediction says no, so it is correctly classified, remark that the predicted probability is not 0%, but 44,8%
        ii. there are many identical lines, let's take them together for readability and show them in a table (venn diagram) a here u can quickly see that are no aldult male first class survivers in the training set
4. show an overview of the influences of our predicting variables according to our classifier (nomogram)
    a. what would the hypothetical chances (according to our classifier) of surviving be for a female child crew member? U have to create this data. (predictions) a since this combo is not in the dataset, u have to create this combo, load it (file), and make the predictions using the classifier: 77%
5. show the predictions of our training set for our classifier with 3 decimals (predictions)
    a. show only the different results [2] a for the training data set u should get 31 lines
6. show a plot of status by age of the test set
    a. if u hoover over a datapoint the probability for surviving should show up a u can use a sieve diagram with test data as input, or a scatterplot
    b. try to zoom in, select and show the selected data a select a group and add a data table to the sieve diagram for output

5.2.3 compare with knn

1. add knn (100 nn) also as test learner on the test data
    a. looking at the performance scores, which one do u prefer and why? a knn has a slightly better classification accuracy (78.98%)
    b. look more in depth at the ones u already know a but has a low sensitivity (of the people who survived only 36.27% where predicted to survive), so i would prefer Naieve Bayes in this case

3. evaluate this classifier (test learner)
    a. on the training set a 77,1%
    b. what does sensitivity and specificity mean? Look them up. a recall is also called sensitivity or true positive rate a specificity or true negative rate (negative in this case translates to not belonging to the target class)
        i. what happens with the sensivity and the specificity if u chance the target class? a the switch, this is logical if u look at the definitions of sensitivity and specifity. a the sensitivity for target class no: 0.914 a the sensitivity for target class yes: 0.473 a the specifity for target class no: 202/427=0.473 a the specifity for target class yes: 817/894 a normally you would take target class no in the context of sens and spec
    c. on the test set
        i. how many are correctly classified in the test set? a 694 out of 880 or 78.9% → in confusion matrix "select correct" / total
    d. on the training set a 1019 of 1321 or 77.1%
        i. what is the probability of surviving for a male adult in first class? (confusion matrix, data table, (mosiac display)) a 0%, there are no adult male survivers on our training data set
            A. is this correctly predicted by your classifier? a our class prediction says no, so it is correctly classified, remark that the predicted probability is not 0%, but 44,8%
        ii. there are many identical lines, let's take them together for readability and show them in a table (venn diagram) a here u can quickly see that are no aldult male first class survivers in the training set

4. show an overview of the influences of our predicting variables according to our classifier (nomogram)
    a. what would the hypothetical chances (according to our classifier) of surviving be for a female child crew member? U have to create this data. (predictions) a since this combo is not in the dataset, u have to create this combo, load it (file), and make the predictions using the classifier: 77% → nomogram can "Naive Bayes" set sliders op (female, child, crew) => survived = yes 77%
5. show the predictions of our training set for our classifier with 3 decimals (predictions)
    a. show only the different results [2] a for the training data set u should get 31 lines
6. show a plot of status by age of the test set
    a. if u hoover over a datapoint the probability for surviving should show up a u can use a sieve diagram with test data as input, or a scatterplot
    b. try to zoom in, select and show the selected data a select a group and add a data table to the sieve diagram for output

5.2.3 compare with knn

Ryk voor blad papier

1. add knn (100 nn) also as test learner on the test data
    a. looking at the performance scores, which one do u prefer and why? a knn has a slightly better classification accuracy (78.98%) 79,5%
    b. look more in depth at the ones u already know a but has a low sensitivity (of the people who survived only 36.27% where predicted to survive), so i would prefer Naieve Bayes in this case

Ryk in Confusion Matrix

Rule of thumb: number of observations n in training data equals at least 5(P+2)

Chapter 6

Regression recording 9

- Constant (like Billy Swan, Everything's the same)
- Simple Linear Regression (X,Y)
- Multiple Linear Regression (X_1,..X_n,Y)

based on the breast circumference (feature) what is the fat percentage? (target) → "select" in Columns

6.1 Exercise: simple linear regression - body fat → rec 10

1. Separate the data into a training and test set (60/40). Make sure the sampling is stratified and replicable.
2. Choose the fat percentage as Y -the class to predict- and the breast circumference as X. (select columns)
3. Build a linear regression model based on the training data (linear regression with default settings)
   a. What is the regression equation? (data table) a connect a data table to linear regression to get the data, f% = -50.3 + 0.684*breast_circumference
   b. Are the intercept and X variabele statiscally significant for the model? (python script) a yes, both p-values are very close to zero (zero is reported)
   c. What is the Root Mean Squared Error? a a way to measure the differences between values predicted by a model or an estimator and the values actually observed
      i. Give the formula. a $RMSE = \sqrt{\frac{\sum_{i=1}^{i=n}(\hat{y}_i - y_i)^2}{n}}$
      ii. How much is it. a 5.756 on the training data, 6,0868 on the test data → in "test & Score" resultboer
   d. What is the Mean Squared Error? a $MSE = \frac{\sum_{i=1}^{i=n}(\hat{y}_i - y_i)^2}{n}$
      i. What is the relationship between MSE and RMSE? a $RMSE = \sqrt{MSE}$
      * ii. Calculate the MSE for your self. (Hint: Use the formula Luke > predictions, select attributes, feature constructor, data table, boxplot) a 33.133 as could be found before from test and score
      iii. Do the residuals fullfill the conditions of indepency and homoscedasticity, in other words are the residuals normally distributed (distributions) and are the residuals independent of the X values (scatter plot). a distribution of the residuals looks normal; the residuals look like a random cloud with no real pattern in it, so that's good a also the residuals look normally distributed in a distribution plot

error 1 →

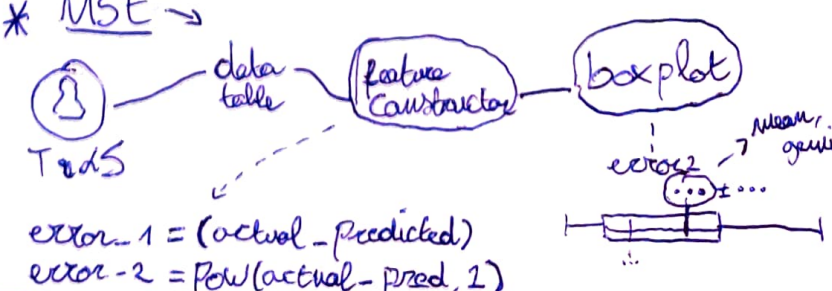
6.2 P-value of a coefficient

This p-value is the result of a hypothesis test.

- what if we ignored this coefficient?
- Does it matter? (>0,05)

A small p-value indicates that it matter.

* MSE →



Model good → look test & Score → R2 → (the higher the better)
↳ determination coefficient → how much the of the Variance of the dependand variable is described by the independant

Regression

Precision / Positive predictive Value =

$$\frac{\text{True positive}}{\text{All positive predictions}}$$

Recall / Sensitivity / true Positive Rate =

$$\frac{\text{True positive}}{\text{all actual positives}}$$

Specificity / True Negative Rate =

$$\frac{\text{True Negative}}{\text{all actual Negatives}}$$

given the set of predictor variables $x_1, x_2, \ldots x_p$
try to determin class y

Classification process

1. gather all data
2. split data in traning en test set
3. build model
4. Now use model on test set to test accuracy
5. then use model to classify

Simple linear regression

Simple: 1 predictor x, try to predict y

$Y = \beta 0 + \beta 1 * X + \varepsilon$ → noise

use traning set to estimate $\beta 0, \beta 1$

Estimate ε based on standard dev estimate in output

Multiple

$$Y = \beta 0 + \beta 1 * X1 + \ldots + \beta p \cdot Xp + \varepsilon$$

discret → KNN → to classify Variable But Now ↙

KNN regression → because its a continuous Variable

regression is about prediction and description

purpose : Description
- goodness of fit
- Calculations on training data
- Calculate R^2
  - measures how close the data is to the regression line
  - Available as output

(more interested in I) purpose prediction
- Accuracy measures
- calculations on test or Validation data
  - Available as output

Residuals

$$e_i = y_i - \hat{Y}_i$$

e_i : error

y_i : actual Value

$\hat{y}_i$: predicted Value

Mean Absolute Error all errors / # errors

Mean Absolute percentage Error

Absolute Value of $\frac{(\text{actual} - \text{predicted})}{\text{actual}}$

add them all up ← and

multiply by $\frac{100}{\#\text{numbers}}$

Mean Squared error

square all the errors, add them up, divide by n

Total Sum of Squared Errors

Square all the errors, add them up

6.2.1 Exercise: horizontal

1. Use the data horizontal.ods a create a csv file from the ods, use the import options of file for the header or use the feature constructor widget to construct continouos x and y features from the imported data
2. Separate the data into a training and test set (1/99) stratified and replicable. a thus having a training set of only 10 instances
3. Draw a scatter plot
    a. What do u see? a no real pattern
4. Build a predictor model based upon linear regression a u need the select attributes widget to select the class or target variable first, the easiest is to apply this from the beginning
    a. Show the regression equation. a y=14.7-0.001x
    b. Is the coefficient of X close to zero? a a good as zero (supra), the p-value is also insignificant for the coefficient of x, only the intercept is significant
5. Build a predictor model based upon the mean value (constant)
    a. Show the test learner results of both models. (training/validation) a they are very similar, eg -0,1 for test data, constant learner
        i. Do they differ a lot? a no, the performance scores are very close for training and validation
    b. Is the mean estimate close to the intercept estimate of the regression? (predictions) a the mean estimate is 15,3, which is close to the intercept (supra), ci of intercept 13,3-16 contains this mean estimate.

6.3 Multiple Linear Regression

Try to build a simple but feasable model.

6.3.1 Exercise: bodyfat continued, a backward selection example

1. Partition the dataset in a training set and a validation set according to the ratio of 60%/40%. Choose seed = '1' in the random generator.
    a. Apply the technique of Multiple Linear Regression to the training data in order to find a model that can predict the percentage of fat based on the all other variables.
    b. What is the linear equation? a %_fat=-33.636+age*0.117+breast_circumference*0.485+fist_circumference*-2.492+hi 0.001+length*-0.070+neck_circumference*0.162+weight*0.023 for orange 2; different sample.. a %_fat=132.529622 0.6269727+breast_circumference*-0.0741858+hip_circumference*-0.0342250+knee_circumfernce*-0.1918894+fist_ci 3.5060683+weight*0.8754334+length*-0.4994723
        i. What is the goodness of fit (RMSE,..) of the training and test set a training set: RMSE=4,552 and test set: RMSE=9,619
2. How can we simplify this model? a remove variables (infra)
    a. Now, systematically remove the variables where the p-value of the coefficients is higher than 0.05. This means, first drop the variable with the highest p-value and apply Multiple Linear Regression again. Repeat this procedure until the p-value of all coefficients are less than 0.05. Does the root mean squared error of the validatie set change much over the successive steps? (select columns with python script, use impute to remove missing values) a first remove hip_circumfernce, then remove breast_circumference, knee circumference, neck circumference; with each step there are very small changes in RMSE. a our model now only consists of 4 predicting variables

Nog Wat Voog

b. Interpret the coefficients of your model, what do the mean? a the intercept is *biased*, non existing people should have a zero %fat, not a negative one; remember inter and extra polation in statistics. a example of interpretation: per unit of length increase there is a .44 decrease in %_fat a you should also note that the model consisting of only breast_circumference is better and simpler.

3. Create a column containing the residuals

   a. Draw on a Box Plot and a Histogram based on the column residuals. Can you agree with the assumption that the dispersion of the variable fat percentage follows a normal distribution? a Histogram doesnt seem to deviate a lot from the normal distribution, Boxplot confirms this, mean and median don't differ a lot

   b. Check for independence of the residuals vs the signifant predicting variables (several scatterplots) a check the scatterplot for each combination of x-variables and the residuals, they all look random, thus independent

Chapter 7

KNN Regression

The same principle as KNN Classification * predict: take the average value

7.1 Exercise: bodyfat continued, comparison of regression and knn

1. Use knn regression as a prediction method
    a. Compare the performance scores on the test set of the following models:
        i. Simple linear Regression (breastcircumference predicts body fat) a MSE=37.034 RMSE=6.086 MAE=5.034 R2=0.503
        ii. Multiple linear regression (all variables predict body fat) a MSE=92.525 RMSE=9.619 MAE=5.789 R2=-0.241
        iii. Statistically significant selected mulitple linear regression (the trimmed down model) a breast_circumference, fist_circumference, hip_circumference, length, a or subset with breast_circumference included a MSE=41.353 RMSE=6.431 MAE=4.670 R2=0.446
        iv. 5NN regression using all the variables a MSE=39.766 RMSE=6.306 MAE=5.071 R2=0.467
        v. 5NN regression using only the same variables as with the statistically significant selected mulitple linear regression. a MSE=35.252 RMSE=5.937 MAE=4.816 R2=0.527
    b. Which predictive model is the best according to you? a purely looking at the performance scores: 5NN with the 4 features a but i would prefer the linear regression model based upon breast_circumference, as i also performs well, and is easy to interpret.

7.2 Exercise: Forecasting daily demand orders

1. Use the *Daily_Demand_Forecasting_Orders* dataset. We will try to predict the *Target* or *Total orders* for a day, this is the total amount of orders ordered for that day.
    a. Explore the data. Can the total amount of orders be a double or float? a no
    b. There is dot-separator indicating a thousand; most programs though see this a a decimal point. First fix this problem. a eg change in a spreadsheet or use search and replace
2. Divide the data in a stratified replicable training and test dataset (60/40%) and use the training dataset for following exploration.
    a. What is the average amount of orders? a 309146.9
    b. What does the distribution of the target look like, when looking a a histogram with 10 classess? a u can do this with the distribution widget, it skewed to the left; thus not symmetrical.

c. What is the most informative projection for the scatterplot? a Banking order (2) and Order Type B

3. We'll to try build a model with the techniques that we have seen so far.

   a. Using linear regression and 6nn with a weighted euclidean distance, how do the models performe on the training data set? a perfect, which is strange for linear regression

   b. Is the performance score of 6nn surprising on the training set? a no, the scores are weighted and the point has to predict itself

      i. Is there a difference when no weights are used? Use again the euclidean distance but now with a uniform distance. a Yes, this is also not surprising, as we are using 6nn and not 1nn

      ii. Is the performance score for the linear regression suprising? a Yes

         A. What could be the reason for this? a We have a many features and not so much data, but the fit shouldn't be perfect, let's investigate more.

      iii. Make scatter plot of the predicted values in the test data. Show 6nn prediction versus linear regression prediction and give the correlation. a 0.95 is a high correlation, but RMSE and others seem high

   c. We continu with distance weigthed 6nn from this point. Let's investigate more. There must something going on. A perfect correlation is seldom observed and usely points to rule or alike.

      i. Try to to find which predictor variables are predicting the target perfectly. a The three order types

         A. What does the regression model look like? Does this make sense? a yes, target is the sum of those 3

   d. Why isn't *urgent* combined with *non_urgent* a perfect match for the target? a in the concrete orders they are nullable or we are missing a feature ()in the dataset) a combined with fiscal sector orders the match is also perfect

      i. Try out a linear regression with just those 2 features, evaluate and look at the regression model. a better on the test data (r2=0.966), than on the training data a t = 0.941 non_urgent + 1.002 urgent + 20158.9

      ii. U would expect that we could should combine them, but apperently the coefficients are slightly different from 1 with a constant in the model. Show the predictions on the test data set and build your own model handmade model without the intercept. What is the equation of this model? a t = 0.941 non_urgent + 1.002 urgent , just the same, as asked, but without the intercept

         A. Calculate the predications according to this handmade model (feature constructor) on the test data set. a adding a new column with feature constructor, using the above handmade equation

         B. Calculate the residual or error for each observation of the test data for this model. a actual value - predicted value with feature constructor

         C. What is the average residual or error of our handmade model? a boxplot, showing the same value as the intercept before

         D. Have you seen the value before? Where? What does this mean? a on average there are 20158.9 orders a day that are not marked as urgent or not urgent.

   e. Which features are left over to investigate on? a the two timing features

      i. What *day of the week* 1 mean? a sunday

      ii. Let's try to buil a model using *week of the month* and *day of the week*. Use all the regression techniques that we have seen so far. Summarize your results. a linear regression seems the work beter than 6nn and constant a non of them have a great performance score

         A. What can we learn from this? Look at the regression model, also check if the coefficients are statistically significant or not. a only dow is left over, as the week progresses, the amount of orders declines with an average of 26236 orders per day. An additional interpretation could be that order stack up in the weekend.

         B. What does this mean for the labour force? Make a scatter plot using the predicted values of your model and the day of the week. a It's asymmetrical, we need more workers on monday then the next day, and so on.

         C. What does this mean for storage? a if we have to import materials our selves, this is better done towards the end of the week, as we are likely to have more space. This is under the asumption that orders are shipped on the same day, we don't know this.

Evaluation

Model Evaluation

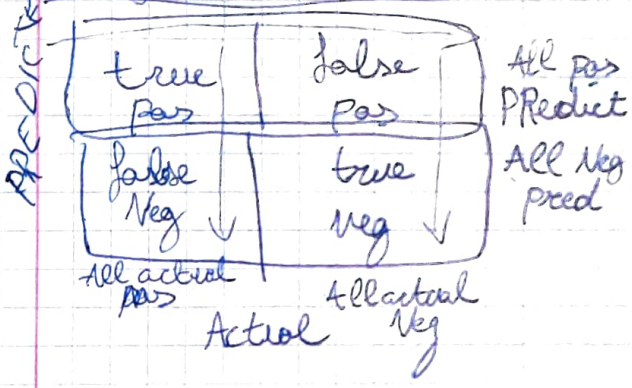
does model fit ? —> business question ?
—> existing or unseen data ?

types of validity

- Apparent (own sample) -> training set
- internal (own population) -> Validation set
- External (other population) -> test set

- Business Relevance ?
- Operational Efficiency and Economic cost ?
- Regulatory Compliance ?
- ~~the~~ ~~Generator~~ • Statistical performance ?

Confusion Matrix



Accuracy =>

$$\frac{\text{true pos} + \text{true Neg}}{\text{Total}}$$

Precision / Pos Pred Value =

$$\frac{\text{true pos}}{\text{All pos Pred}}$$

Recall / Sensitivity / true positive Rate:

$$\frac{\text{TRUE positie}}{\text{All actual positives}}$$

specificity / true Neg Rat =

$$\frac{\text{True Neg}}{\text{All actual Neg}}$$

Regression to classification

using threshold: if pred > threshold, predict class
else NOT predict Class

ex. 0,5 of 0,7

↓ ↓

Make Confusion Matrixen

Vergelijken en kijk na

Validating a model

Overfitting

use train, validation, test set

Cross Validation

K-FOLD Vb 3-fold

- Split data in three groups
  - and build 3 models
    - 1 (blue) is built using split 1 en 2 as training set
      ~~split~~ split 3 for validation
    - 2 (green) slit 2,3 for train, 1 for validatie
    - 3 (orange) split 1, 3 for train, 2 for Validatie

(using whole set, but also able to validate each time)
(Not leaving out data)

- Choose best Model
- evaluate "best" on orgual ~~training~~ test

Leave one out

- K = numb of instances
- leave one instance for validation and train on remaning training set
- Repeat for all instances

Chapter 8

Foreign Data Wrappers

A foreign data wrapper is a way to connect to external data. U can find this in different software products.

A short introduction using PostgreSQL and an FDW.

8.1 Google vs Duckduckgo search

The question is: does a search engine -like Google or Duckduckgo- sometimes give different search results given the same query string? Some called this tailored, others biased or misleading.

Given a common search keyword.

Steps:

1. Install www_fdw https://github.com/cyga/www_fdw (u need PostgreSQL to be installed)
    a. Check the wiki of the project for steps u can take for Google search
    b. Check https://duckduckgo.com/api for duckduckgo api access
2. Store the results
3. Compare with others
4. Conclusion of the question?

8.2 Wikipedia

The questions are (given a common search keyword):

- give a distribution of the 10 most common words used in the description of the page on the english wikipedia and your native language.
- are these distributions similar?
- which language uses the most different words (in other words, the highest amount of different words)?

Steps:

1. Check the wiki of https://github.com/cyga/www_fdw howto create a foreign table pointing to wikipedia
2. Store the results of 100 articles in the two languages

a. Possible tips
   - there is a function to parse text to an array of words (string_to_array)
   - there is a function to convert an array to a table (unnest)

3. Show the distributions of the 10 most common words in both languages
4. How many different words does each language use in this case?

8.3 Analyzing social media

There are companies that aggregate the open available social media for different reasons, for example for trend watching.

One of them is socialmedia.com

Given a common keyword(s) search:

- are the social media correlated?
- which socialmention category has the most entries (blogs, microblogs, events, comments, news ..)?
- is the keyword stagnant, declining, inclining over time?
- is the author related to the keyword (thus promoting the keyword)?

Steps:

1. Check http://code.google.com/p/socialmention-api/wiki/APIDocumentation for the structure of your foreign table.
2. Notice: the number of queries is limited to 100/day/user

8.4 Importing data into orange

- using the native structure: create a tab delimited file with a predefined structure
  - u can easily view this structure when u open an documentation dataset with a simple texteditor
    * domains
      1. d: discrete
      2. c: continous
      3. s: strings
    * type
      1. class: define a class variable
      2. meta: meta variable
      3. string
  - from a database
    * posgresql steps, a psql example:

```
\?
\h COPY
\o mytabdelimitedfile.tab
COPY
        (SELECT * FROM SPELERS)
        TO stdout
        WITH (FORMAT csv, HEADER true, DELIMITER E'\t');
\o
```

- connect to a database
  - postgresql example:
    - python script widget
    - u need to use psycopg2
  - don't forget the datastructure

```
import psycopg2
..
column_names = [desc[0] for desc in cursor.description]
..
```

Associations

Remember:

Supervised: learning depends on known target value -> ground truth

Unsupervised: ~~unsupervised~~ has no target — for -> underlying patterns in data

- Assos. are unsupervised
- In set of transactions, there are items that occur together in many trans. -> these co-occurences are Assos.
- We look for things that happen together often/regularly/freq. in the dataset.

- Given set of transactions:
  - find if-then rules based on items occurring together
    - e.x. {Diapers} => beer
    - {Milk, bread} => eggs

Co-occurence does not ~~~~ imply causality ?

↓

principle that there is a cause for everything.

- Types of Assos
  - Boolean
    - bread ^ Milk => Diapers
  - Quantitative
    - age [30, 39] ^ income [1500, 400] => buys PC
  - Single attribute
    - Beer ^ chips => Sausage
  - Multiple Att
    - age [18, 25] ^ income < 1500 => student
  - Multi relational
    - buys(x, PC) ^ friends(x, y) => buys(y, PC)

relation: "person x bought pc"

Breaking it down

- Single level
  - Beer => ~~~~ diapers
- Multi-level
  - Stella => Bavi diapers
  - Chauffe => Pampers

Apriori Algorithm

1. find all freq 1-itemsets
2. Combine all freq 1 itemset to find candidate 2-itemsets
3. Check candidate 2-itemsets and eliminate those below support threshold
4. Combine frequent 2-itemset with matching items to identify candidate 3-itemsets
5. Eliminate any candidate 3-itemsets below the support threshold
6. Continue ...

Voorbeeld in lecture video ↑

* Interest is • 0 : A has no influence on B
• >0 : B more likely to occur if A occurs } checking if its not
• <0 : B less likely to occur if A occurs } coincidentally

Chapter 9

Association Rule Mining

Can we use a set of transactions to find rules that for a given itemset predict the occurrence of another itemset? We are looking for co-occurences, we cannot say anything about causality.

Terminology:

- Transaction (or basket) -> list of items that occur together : purchase, webpage, ...
- Itemset -> list of 1 or more items that occur together
- Item -> thing/concept that happens within transaction
- Support -> fraction of transactions that contain particular itemset -> bread -> 4/5
- Confidence -> conditional probability => $(A \Rightarrow B) = P(B|A) = \frac{support\{A,B\}}{support\{A\}}$
- Interest (or leverage) -> indicates the influence A has on B -> $interest(A \Rightarrow B) = confidence(B|A) - support(B)$
- Coverage
- Strenght
- Lift
- Rule

9.1 Exercise: Titanic dataset revisited -> rec 15

We already gained some insights in this dataset in the previous exercises. This time we try to come up with some rules that hopefully complement our previous findings.

Use the whole dataset, as a exception we will not be working with training and validation in this example. U should however use them as before in new situations.

1. Get the itemsets with a minimal support of 60% (itemsets)
   a. Show these itemsets (itemsets explorer) <- m.
   b. What are the 1-itemsets that are frequent? a age=adult, sex=male, survived=no a u need to make sure all variables are used, eg with select columns, otherwise the target is ignored for example
   c. What are the 2-itemsets that are frequent? a age and sex; age and survived; sex and survived a with the same filters as the above 1item sets
   d. What is the only 3-itemset that is frequent? a age, sex and survived. If this 3-itemset is frequent, then all sub 2-item sets and all sub 1-item sets must be frequent; if they were not, then this 3-itemset could never be frequent. This is the principle of monotonicity. a with the same filters as before, adult, male not surviving

check in

   e. How many people did not survive? a 1490
      i. Try to check this number in another way. a itemset explorer or attribute statistics

2. Try to find the association rules with a minimal support of 40% and a minimal confidence of 20% (association rules) a u will find 14 rules

3. To your opinion, which association rules are the strongest?

   a. With a high support or a low support? a a higher support means there are more instances supporting this rule, ie it occurs a lot; frequency
   b. With a high confidence or a low confidence? a a higher confidence means that its more likely that the consequent appears given the antecedent, its a conditional probability
   c. With a positive leverage or a negatieve leverage (interest)? a the leverage or interest compares the confidence with the support, thus a positive leverage means that the antecedent may "encourage" the consequent, and a negative leverage the opposite (discourage)
   d. With a high lift or a low lift? a a lift that's higher than 1 suggests a positive correlation, similar to a positive leverage, a lift that's lower than 1 suggests a negative correlation
      i. What are the reference points for lift and leverage? a 1 for lift, 0 for leverage
   e. To what use is strength? a it compares the frequency of the consequent with the antecedent, a value that's lower than 1 means that the antecedent occurs more than the consequent
      i. Try to calculate the strength in another way. a using the boxplot statistics, for example the strengt of male → not surviving is 0.861=1490/1731
   f. To what use is coverage? a coverage is like support, but limited to the antecedent; so what the probability that the antecedent occurs

4. Calculate the conditional probability of a male not surviving, ie given a passenger is male, what are the chances of not surviving? a using distributions, 1364/1731 = 0.788

   a. Compare with the confidence: given a male passenger, how much do not survive? a 0.788 as the above is the definition of the confidence

5. Look at the rules more in depth (association rules)

   a. Which 2 rules would u choose? a adult males are not likely to survive; and in short males (or adults) are not likely to survive
   b. Does this complement the findings before with Naive Bayesian Classification? a u can compare this easily with the nomogram, both sex and age confirm this finding; the status variable doesn't come into play here since we started with minimal support of 40% and a minimal confidence of 20% for building the assocation rules.
   c. Conclusions? a being a male or adult or both clearly wasn't a good thing on the titanic..

9.2 Exercise: small market-basket

Use the dataset market-basket

1. Which 1-itemsets are the most frequent (minimal support of 60%)? a bread, milk, diapers and beer
2. Which 3-itemsets are the most frequent (minimal support of 60%)? a none

   a. Can you find a item in the above found 3-items set, that isn't frequent as a 1-itemset? a as there are none,.., but if you look at the 2 itemset, then every single item a in the 2 itemset is frequent

3. Given a minimal support of 40% and minimal confidence of 90%.

   a. What is the rule with the highest lift? a buying cola implies also buying milk and daipers
   b. What is the rule with the second highest leverage? aa buying beer implies buying daipers

9.3 Exercise: large online retail

Use the Online_Retail dataset. This dataset contains a row for every single item that is bought.

1. Show the number of transactions per country.
2. Can u use this dataset as such?
3. Find a way to group the items that were bought in the same transaction. U have several options, use your force Luke.
    a. Find the most frequent 2-itemset in Belgium.
    b. Find the rule with the highest leverage in Belgium.

Clustering

Supervised: -> data -> set of pairs $\langle x, y \rangle$, where $y = f(x)$
goal: Approximate f

Unsupervised: -> data is just x
goal -> find structure in data
challenge: ground truth is often missing

uses
- Visualization
- Data compression
- Density estimation
- preprocessing step for supervised learning
- partition data
- Novelty detection

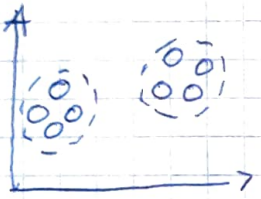
WHy?
- in many cases -> No class label
- Humans: how do we form categories of objects?
- humans are good at creating groups / categories / clusters
- image analysis finding groups in data is very from data useful

WHAT?
- Cluster: a collection of data objects
  - similar in same cluster
  - dissemilar to the group objects
- Cluster Analysis: grouping objects into clusters
- is unsupervised
- Clusterings are usually not right or wrong

How?

- Clustering is grouping similar items together
  - to establish prototypes or detect outliers
  - simplify data for further analyses / learning
  - visualize data
  - As a stand-alone tool (get insight in data distribution)
  - As a preprocessing step for other algorithms

eg



Uses in
- Marketing
- land use
- insurance
- urban planning
- seismology

What is good Clustering?

good clustering method will produce
- Cluster with
  - high intra-class similarity
  - ~~high inter~~
  - low inter-class similarity
- precise definition of clustering quality is difficult
  - Application-dependant
  - ultimately subjective

- The Clustering problem
  - let $x = (x_1, x_2, \ldots, x_d)$ be a d-dimensional feature vector
  - let D be a set of x vectors,
    $D = \{x_1, x_2, \ldots, x_N\}$
  - given data D, group the N vectors into K groups such that the grouping is optimal

Basic Concepts: Distances / similarities

- Clustering Methods use a distance (similarity) measure to assess the distance between
  - a pair of instances
  - a cluster and an instance
  - a pair of clusters
- given an distance value, we convert it to a similarity value $Sim(i,j) = 1/[1 + dist(i,j)]$
- Not always straight forward to go the other way
- ...

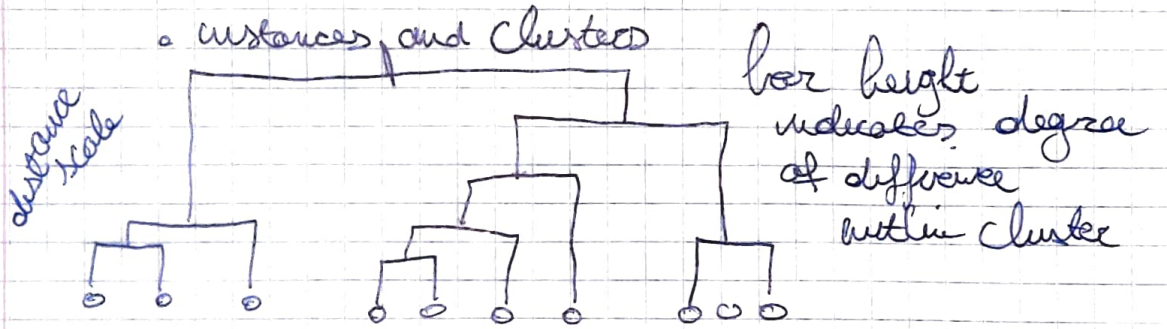
Euclidian distance ($p=2$)

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{p1} - x_{p2}|^2)}$$

We see only hierarchical
and
partitioning

Hierarchical Clustering: Dendogram

- Can do top down (divisive) or bottom-up (agglomerative)
- In either case we maintain a matrix of distance (or similarity) scores for all pairs of
  - instances
  - Clusters
  - instances and Clusters



Chapter 10

Clustering

(unsupervised)

This is an unsupervised data mining technique.

We want:

- Low inter-class similarity
- High intra-class similarity

10.1 Exercise: store dataset

Which items would you recommend to a customer? Or which items are often bought together?

1. Which tecnhique already seens comes to mind? a assocation rule mining
2. Use the *Online Retail* dataset of the datasets widget.
    a. How many rows are there? a 914
    b. What do they look like? a each transaction is a row
        i. Can we calculate the distinces between *what* in this case? *What* is interesting to calculate the distances between? (see start exercise). a which items are often bought together, we could also calculate the distances between columns as they contain as features the items.
            A. U can tranpose the data in meaningfull way. What does the data look like now? (transpose) a the rows are now the items, we want to calculate the distances between the items, thus in this form between the rows
    c. Which distance metrics don't mind missing values? (distances) a euclidean, manhattan, cosine, jacard
        i. Use the cosine metric to continue
            A. When using single linkage what do u notice? a a chaining effect, a small cluster to which items are added step by step
            B. Which 4 items would you recommend based on complete linkage to someone (ready to) buying a Mug? U can use annotation to show extra information on the cluster. a T-light holder, Decoration, Sign, Bottle
3. How can you show the distances in a matrix? a Distance matrix
    a. How do you show this graphically? a Distance map

10.2 Exercise: zoo dataset

1. Calculate the distances for the attributes/columns (Distances)
   a. In order to be able to calculate the distance you have to transform your colums (Continuize)
      i. What would be an appriopate coding conversion for the discrete attributes (hair, legs, ..; multinomal attributes) to make the interpretation of the clusters easier in this case? Take a value range of 0 to 1. a Multinomal attributes: most frequent value as base
   b. Which clustering technique are u going to use and why ? (hierarchical clustering) a average linkage, as it is less susceptible to chaining than single linkage; complete linkage can also be useful, u have to look at the clusters they generate
      i. Try to find a good distance metric for this use case. a there are several good metrics in this case
         A. Look up what these distance metrics mean. a Check the orange online documentation and links, examine the diffences between these metrics
         B. Change the distance metric and look at the effect it has on the formed clusters. Use the dendogram and a cutoff line for the distance to see the clusters that hopefully make sense. Make sure the use as annotation *Attribute names*. Give the cluster that make sense in your opinion and shortly discuss the different metrics for this case. a There are not many animals with 5 or 8 legs, so we might want to minimise their influence a Some clusters that make sense: (catsized, hair, milk and no eggs), (aquatic, fins and no legs or breathes), (two legs, feathers, airborne and not toothed) and (6 legs with no tail or backbone) a euclidean: not a bad job, 5 and 8 legs get grouped together a manhattan: similar to euclidean a cosine: similar to manhattan, but ignores 5 or 8 legs, on higher level, birds are not predators, and so .. a jacard: similar to cosine a spearman: not bad, but early cluster of venomous and 8 legs seems less interesting a spearman absolute: not bad, but toothed seems a bit in an odd place a pearson metric will behave the same as the spearman in our setting since alle variables have been coded with 0 or 1 a since the interpretation of the formations of these clusters is personal (not having studied biology), i would choose cosine in this (reasons: look at the clusters, look at the toothed, 8 legs and (not) venomous and the behaviour of categories that don't occur a lot)
      ii. How many clusters are there on the lowest level using average linking and the cosine distance? a 20, since there 20 columns, features with some differences
   c. Which attribute is left alone for every form of linkage using the cosine distance, u can chance the annotation to default? a 5 legs, the lonely starfish
2. Calculate the distances for the rows/examples.
   a. Perform clustering with average linkage.
      i. Use the cutoff line to find a cluster containing nothing but mammals, u can use type for the annotation.
      ii. Which distance metrics perform best for this task? a euclidean an manhattan have a reptile in the cluster, the other metrics seem to perform well, some other clusters are mixed at this cutoff value. a For seperating mammals in 1 cluster, for example jacard with average linking works.
      iii. Do the top 10 clusters for spearman make sense? a 10, yes, (almost) perfect clusters: amphibian and birds; approximately: the other types. Reptiles are more difficult. U can also investigate with other linkage forms.
      iv. Visualize the data in another way (linear projection)
         A. Try out different projections.
         B. Does it make sense that the birds are on one side of the graph, while the mammals on the other side when u use airborne, feathers and milk? a yes, feathers/eggs vs milk
         C. Same question, does it make sense for the other types of animals? a depends on the type and the projections, in this projection: amphibian, fish, some insects, invertibrates and reptiles are not clearly seperated
3. Try now with k-means and silhoutte scoring to obtain the best k value, and use this k-value (k-means clustering). We don't want more clusters than types of animals. a k should be between 4 and 6, with a preference for 6, u can increase the number of simulations (reruns and interations). If you want to minimise on the number of clusters, 4 is good choice.
   a. Make a scatterplot feather by cluster, use 4 clusters.

i. Which so classified bird doesn't have feathers? a a tortoise

b. Change the type meta column (eg bird, mammal..) to a regular feature a select columns

c. Look at the distribution of them for the variable cluster grouped by type. a for each cluster we get to see if they correspond to a one of more types

i. Which type dominates an entire cluster? a there is an entire cluster of only mammals

ii. What else can you see? a all the birds are in the same cluster; insects and invertibrates seem related.

Chapter 11

Loading Data from a database

We can load data from a datase into our datamining framework.

Some remarks:

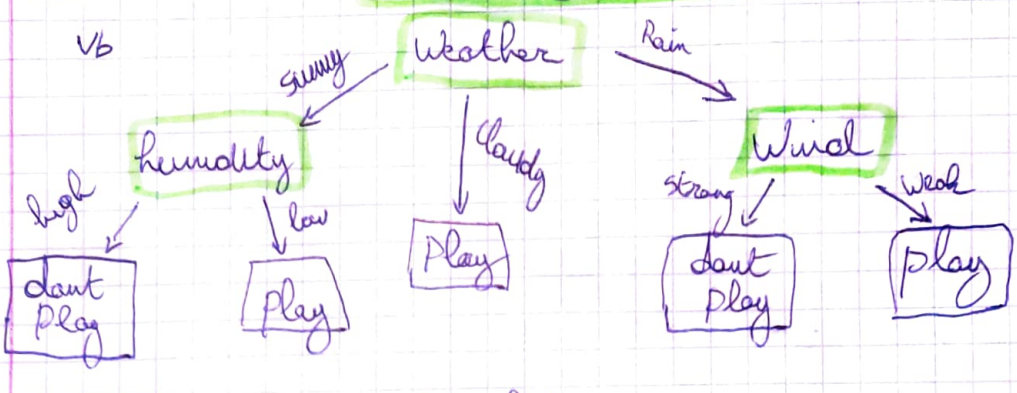
- What is the size of the dataset?
  - Small (fits in memory): load the whole dataset
  - Too big to fit into memory:
    * Take a representative sample
    * Or work in batch (or parallel depending if you have the hardware)

If you need to join data, or other querie like operations, this is probably best handled by your database and not your datamining software. So the data preparation starts there.

11.1 Exercise: weather dataset

Try to investigate what happened in Denmark some years ago. U can use queries on your database and/or datamining techniques (SQL Table).

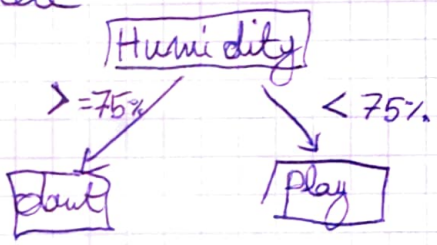
Decision trees for classification



- the ▭ => Nodes => Attributes
- the "→" => Attribute Value = ARCS
- the ▭ => bottom nodes => leaves = classification
  - L> leaves tell you ~~which~~ which class

- Discrete Attribute : each possible attribute
- Value is A path : arc in decision tree

IF
- Continuous Attribute : threshold (cutoff points)
  - L> Convert to discrete

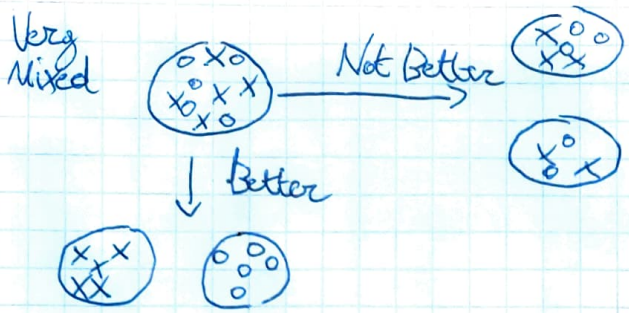


Building dec tree

Nodes :

Root : first split
Internal : split based on value
leaf : class label

- Start with Root
- add internal nodes to further split the data
  - each split divides data into groups
- how to choose split ?
  - information Gain
- goal best tree possible (???)
  - More pure leaves



Entropy measures the homogeneity of a group

Ent of 0: pure group, all one class

Ent of 1: complexity mixed, 50/50 split of classes

Formula in Decision tree pdf

- information ~~Gain~~ Gain

- $Gain(T, A) = Entropy(T) - Entropy(T, A)$

Ex: $Gain(playTennis, Weather) = E(playTennis) - E(playTennis, Weather)$

$= 0,94 - 0,693 = 0,247$

Choose split that gives the best information gain.

- Basic decision tree Algorithm (ID3)

- in this example => play Tennis ? => {yes} or {No}

Its our target Variable.

Chapter 12

Decision trees

We can use decision trees as a supervised learning method

- Classification trees
  - Attribute Selection
  - Binarization
  - Pruning
- Random Forest
  - Forest = many trees
  - Democratic Vote

12.1 Excercise: car evaluation

When designing and producing a car, it is important what to know whether or not it probably will sell. We will call this car acceptance. At least we want the cars to be acc (acceptable)

Description: CAR acceptability

- PRICE overall price
  - buying buying price
  - maint price of the maintenance
- TECH technical characteristics
  - COMFORT comfort
    - doors number of doors
    - persons capacity in terms of persons to carry
    - lug_boot the size of luggage boot
  - safety estimated safety of the car

Questions

1. Use the car evaluation dataset (datasets)
2. Use a replicable stratified training and test sample of 60/40%.

3. Using the training dataset

   a. What is the target class to predict and what are the 4 values? a car: unacc, acc, good, vgood

   b. Sort the attributes by information gain (rank). What is the most important feature according to this criterium? a safety

      i. Which attributes explain indivually 50% of the entropy present. a safety, persons, buying

      ii. What is information gain in one sentence? a the expected amount of information as a reduction of entropy, which is a measure for chaos or disorder

   c. Use a decision (tree) to build a model, a node can have more than 2 branches. How much is the classification accuracy (on the the training data or test data?)? a 0,891

      i. What does the decision tree look like? (tree viewer) What is the first criterion being used? a large, zoom and navigation needed, or limit the levels a safety

      ii. Did you expect this to be the first criterium? Why? a yes, the most information gain (supra)

      iii. How likely is a car to be acceptable or better? a 30% -> Kans acc + K.good + K.Vgood = 29,00 %

         A. If the safety is low, what does this mean for the car? a 100% unacceptable -> Buchlaan in tree.

         B. If the safety is high, how many cars are unacceptable? a 172 are unacceptable -> Root -> high -> 172 unacc

         C. What does the 345 mean? a 345 is the total number of cars with a high safety in our training set

         D. Given a medium safety, with a maximum of two persons in the car. Is this car likely to succeed on the market? a also 100% unacceptible according to the model -> zie in tree

      iv. U can also try out a pythagorean tree. The big square is the training dataset. Why is there a large square in the first *split* colored as unacc? a if you hover of this picture, these are the cars with low safety (cf supra)

   d. If you compare this tree to a binary decision tree, which one seem to perform better? a binary tree

   e. Continuing using the decision tree (not binary)

      i. In absolute values, which combination of (actual, predicted) is misclassified the most? Actual being the actual measured value, predicted is the value according to the model. a (acc predicted, unacc actual)

         A. Is this the same for the binary tree? a yes

      ii. In relative values to predicted, which combination of (actual, predicted) is misclassified the most? a (good pred, acc actual)

         A. Is this the same for the binary tree? a no, good predicted, unacc actual

      iii. Why is it logical that unacceptible (actual) is part of a combination in absolute values? In other words: why do we see so often unacceptible? a 70% of the training set is unacc

   f. Compare with other classification techniques that you have seen. Which one seems to performs best using CA? a Using CA, the bin decision tree seems best a compared to nb, weighted 9NN euclidean and decision tree

12.2 Exercise: zoo dataset revisited

U can use the whole dataset as a training dataset, as a exception we will not be working with training and validation in this example. We are still waiting for test dataset from the farm of Uncle Jef.. U should however use them as before in new situations.

1. Build a binary classification tree based upon the information gain criterion (tree)

   a. Look at this classification tree

      i. What is the probability for fish given no milk, no feathers and a backbone a 0.591

         A. If u sort by information gain, which criterion should be the first split? (rank) a legs

         B. What is the first split in our tree? a milk
         [Error: orderedlist too deeply nested]

         C. Show a scatter plot of milk by feathers of our data.

            - What do u see above milk 1? a mammal
            - What do u see above milk 0 and featers 1? a bird

b. Show a graph of our classification tree.

   i. How many leaf nodes contain more than 1 type of animal (which each leaf nodehaving at least 2 instances)? a 1

   ii. Which animals are misclassified by this tree? U can select a leaf and use predictions. a seasnake

      A. How many animals have 5 legs? a 1

      B. Is this an error? a no, starfish

      C. What is overfitting when we talk about model building? a A model can be too detailed or too specificly adapted to the training set, this can quickly happen if the number of attributes or features is relatively high compared to the number of instances or rows. a in this case the fitting for the lonely starfish; it is different though and can easily be identified in this way

2. Build a random forest (random forest) with a random seed of 1

   a. Show the first tree that was randomly generated a 13 nodes, 7 leaves, 100% for bird and mammal; a meaningless seperation for fish

3. Evaluate the classification for the classification tree vs the random forest. a If you choose by CA: our tree is slihtly better (0.9703) then for the rf; but overall both are very close. Another way of choosing is insects vs invertibrates. Also, we did not use training and validation set!

Chapter 13

Logistic regression

Logistic regression is a robust classification technique.

13.1 Cars (continued)

1. Logistic regression is a technique that comes from categorical data analysis.
   a. What are odds (and odds ratio)? a https://en.wikipedia.org/wiki/Odds a https://en.wikipedia.org/wiki/Odds_ratio
   b. How can interpretate the coefficients of the logistic regression? Do they have a practical meaning? a https://en.wikipedia.org/wiki/Logistic_regression a Binairy case: https://en.wikipedia.org/wiki/Logit
2. Use the car evaluation dataset from datasets. Make sure the target variable is car.
   a. Recap: What are the possible values for car and their meaning? a unacceptibla, acceptible, good, very good
3. Apply 60/40 replicable stratified sampling.
4. Use logistic regression to build a model, always with C=1.
   a. What does the model look like? a use coefficient>data to data table
      i. How many equations are there? a 4
         A. What does each equation mean? a log odds for that outcome value
   b. Do the coefficients differ a lot when using L1 or L2? a yes, with L2 no zeros, with L1 1 per group is zero
      i. Use a feature constructor to calculate the odds ratios for *unacc*
         A. What is the odds ratio for *low safety* in this case (using L2)? a 122,5 = exp(4,808)
         B. What does this odds ratio mean? a odds low vs odds not low for unacc
   c. Using L1
      i. Show the log odds ratios with a nomogram
         A. Which features have the largest influence on our model? Have we obtained this information before? a safety and doors, cd ex decision trees
         B. With target class *unacc* the log odds ratio for unacc is about 11,5. How can you *calcute* this number from the regresssion equation? (Tip: look at the position of *med* and *high* and the regression coefficients) a 9,4..+2,1.. = 11,5..
5. Which classification technique seems to perform best, based on the classification accuracy? (nb, 9nn (eucl, weigth), tree (bin an not)) a bin tree
6. Use the predictions widget to get the individual predictions by logistic regression.

Logistic regression

- Resembles Regression: Continuous numeric output
- But Really; Classification: Numeric output gives a probability of belonging to a discrete class

↳ instead of finding continuous y (linear regression) you get a discrete binary label, yes/no, 1/0, etc

(but instead of 1 or 0, we predict the probability that $y=1$)

a. Use the confusion matrix to select interesting misclassifications. How many are classified as good while actually being unacc? a 2
   i. Show these points on a scatterplot for buying and maint. What combination of buying and maint contains these two misclassifications? a buying and maint low
   ii. What are th log odds values for good vs not good for these two instances? a 0.570 and 0.517
   iii. Is it logical that this is the largest value of the 4 log odds for these instances? a yes, they were predicted as good

13.2 Cell data

1. Data preparation
   a. Use the complete merged dataset to start from (db and csv), merge to the two sources into one database table.
2. Depending on the available amount of memory, divide into a training and validation dataset or take two samples to act for the training and validation dataset.
   a. On what dataset does it make sense to make the predictions?
   b. Use logistic regression to build a model to predict the country of the measured signal(mcc). (predictions)
      i. Give the regression equation for Belgium.
      ii. Give the practical meaning of at least one paramater in the model. a cf link supra, odds factor exp explanation
   c. Compare with the results from the classification tree.
      i. What seem to be 2 most important factors according to our tree?
   d. Evaluate logistic regression and the classification tree.
      i. U can try to tweak the settings of the classifiers.
      ii. Which one is the best according to you?
      iii. What does *F1* mean?
   e. Use a scatter plot to identify the misclassified rows in the general picture.
      i. This plot needs the dataset and the misclassified rows as input.
      ii. Does it make that these items where misclassified?
3. The same exercise trying to predict cid (cellid) a U should notice that this is probably a bad idea in this use case as the cellids are not unique.

Chapter 15

Support Vector Machines

If stuck try another perspective. [svandenbroucke] Try to maximaze the seperation.

15.1 Iris

We take a training sample 60/40 as before.

1. SVM background.
    a. Why do we need a kernel function?
    b. What does the normal SVM mean?
    c. What does nu-SVM mean?
    d. What is a recommended value for g of the kernel function?
2. Compare SVM with other classification techniques you know.
    a. Which one do you prefer for the Iris dataset?

Chapter 16

Neural network

In this case we use a multi-layer perceptron (MLP) algorithm with backpropagation.

16.1 Svm exercise continued (Iris)

1. How can you define the multiple layers? (neural network)
2. Try to find the best neural net.
    a. Compare with the other classification techniques you know.
    b. Which one do you prefer for the Iris dataset?

Chapter 17

Model Evaluation

Not only technical but they are real world aspects to take into mind.

17.1 Titanic

1. What does the calibration plot show and when do you have a good model according to this evalution tool?
2. Use the titanic dataset.
    a. Use data sampling to setup a training, validation and test set.
    b. Try all the classifiers u have seen and evaluate them extensively. (roc analysis, lift curve, calibration plot)
    c. What would it mean if we take precision as the most important?
    d. What would it mean if we take recall as the most important?
    e. Which evalution measurement would you take?
    f. What seems to be best cutoff value for your best model and why?
        i. Explain deeper using roc and lift.
        ii. Why is this dependent on the target class?
        iii. Show a table of the probabilities for each model.

Chapter 18

Visualisation

Intuitively we interpret charts linearly. Visualisation should be an assistance for interpretation, unnecessary elements should be avoided.

18.1 Iris

1. Do you see a pattern for Iris Setosa (parallel coordinates) a Differs from the other 2, having a large sepal width, but small petals
2. How does a scatter map differ from a scatter plot? a It might be easier to visualise the density (linear), a map is less precise though.
3. Show the class density with a linear projection of a good projection. a U can use rank projections from the scatter plot
4. Use a heat map to select the visually two most similar iris types. a Versicolor and Virginica, optionally use clustering to order rows and columns.