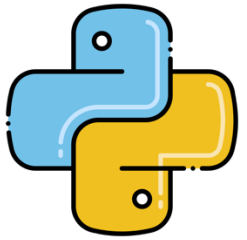
UNIVERSITÀ DEL SALENTO

Data Mining: Introduction

A gentle introduction to Data Mining Course

Goal of the Course and Exam

- Algorithms to **analyze data** and **extract knowledge**
- Use of python libraries for data mining



Project based exam

- Chose a Dataset
- Apply one or more of discussed algorithms or additional solutions
- Extract valuable **knowledge**
- Provide project code on Github (rich in documentation)
- Give a 20-minute speech to present your work and justify your choices.

Resources

Books:

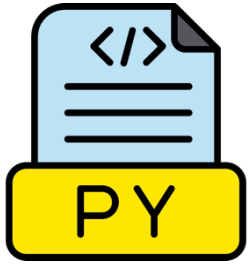
Python for Data Analysis, 3E - https://wesmckinney.com/book/

Introduction to Data Mining (Second Edition) - https://www-users.cse.umn.edu/~kumar001/dmbook/index.php

Massive Mining Data Sets http://www.mmds.org

Slides and code:

https://github.com/beppe2hd/DataMining



What is and Why do we need data mining?

After years of data mining there is still no unique answer to this question.

A tentative definition: Data mining is the use of **efficient techniques** for the **analysis** of very large collections of **data** and the extraction of **useful** and possibly **unexpected patterns** in data.

Large amounts of data can be more powerful than complex algorithms and models

Data mining is pivotal in case of:

- Need to analyze **raw data** to **extract knowledge**
- Really, really huge amounts of raw data!! (TB of data is generated by the second)
  - Mobile devices, digital photographs, web documents.
  - Facebook updates, Tweets, Blogs, User-generated content
  - Transactions, sensor data, surveillance data
  - Queries, clicks, browsing
  - Cheap storage has made possible to maintain this data

Data is power! (With great power comes great responsibility))

Today, the collected data is one of the biggest assets of an online company

- Query logs (Google)
- Chat
- Posts comments, and follows of Social Networks
- GPS data
- Transactions

Interconnected data of different types:

From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, images though cameras, queries to search engines

We need a way to harness the collective intelligence

The data world is very complex

- Multiple types of data:
  - Tables
  - Time series
  - Images
  - Graphs

- **Spatial** and **temporal** aspects

Examples

Transaction Data involves Billions of real-life customers:

- WALMART: 20M transactions per day
- AT&T 300 M calls per day
- Credit card companies: billions of transactions per day.
- The point cards allow companies to collect information about specific users

Network Data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 3.6 billion users
- Instagram: 1.5 bilion users
- Blogs: 250 million blogs worldwide

Document Data:

- Web as a document repository: estimated 50 billions of web pages
- Wikipedia: 4 million articles (and counting)
- Online news portals: steady stream of 100's of new articles every day

Environmental data:

- Soil Moisture networks: 3200 stations recording multiple soil and climate data
- European Climate Assessment and Dataset is receiving data from 101582 series of observations for 13 elements at 25000 meteorological stations

Examples

Genomic Sequences

- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- $3*10^9$ nucleotides per person → $3*10^{12}$ nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

Behavioral Data

Mobile phones today record a large amount of information about the user behavior

- GPS records position
- Camera produces images
- Communication via phone and SMS
- Text via Facebook updates
- Association with entities via check-ins

Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.

Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.

So, what is Data?

Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
- *A collection of attributes describe an object*
- Attribute is also known as variable, field, characteristic, or feature
- Object is also known as record, point, case, sample, entity, or instance

Data characteristics

- Size: Number of objects
- Dimensionality: Number of attributes
- Sparsity: Number of populated object-attribute pairs

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of Attributes

There are different types of attributes

- Categorical
  - Examples: eye color, zip codes, words, rankings (e.g, good, fair, bad), height in {tall, medium, short}
  - Nominal (no order or comparison) vs Ordinal (order but not comparable)
- Numeric
  - Examples: dates, temperature, time, length, value, count.
  - Discrete (counts) vs Continuous (temperature)
  - Special case: Binary attributes (yes/no, exists/not exists)

Types of data

Numeric Record Data

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

Such data set can be represented by an n-by-d data matrix, where there are n rows, one for each object, and d columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Categorical Data

Data that consists of a collection of records, each of which consists of a fixed set of categorical attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

Types of data

Document Data

Each document becomes a **term** vector, each term is a **component** (attribute) of the **vector**, the value of each component is the number of times the corresponding term occurs in the document.

Bag-of-words representation – no ordering

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Set Data

An example is transaction data, where each record (transaction) is a set of items.

A set of items can also be represented as a binary vector, where each attribute is an item.

A document can also be represented as a set of words (no counts)

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

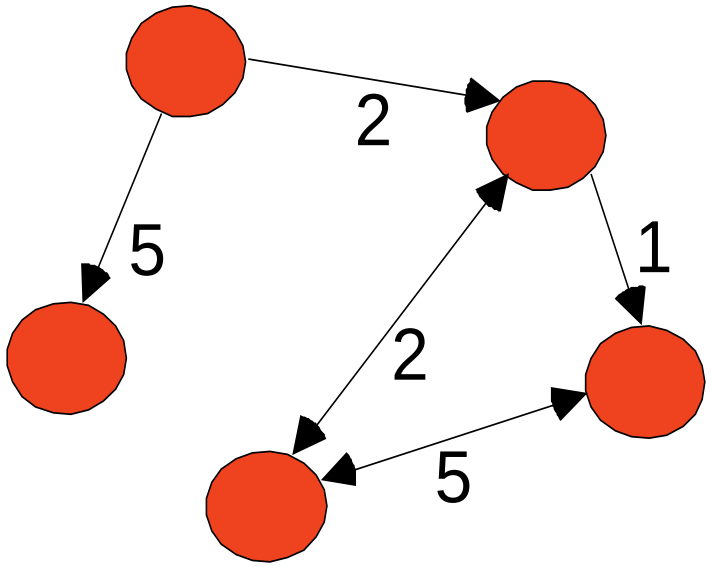
Types of data

Time Series

Sequence of ordered (over “time”) numeric values

Graph Data

Web graph and HTML Links, Social Network



What can you do with the data?

Suppose that you are the owner of a supermarket and you have collected billions of market basket data.

- What information would you extract from it ?
- How would you use it?

Possible use:

- Product Placement
- Catalog Creation
- Raccomandation

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

What can you do with the data?

Suppose you are a search engine and you have a toolbar log consisting of

- Pages browsed,
- Queries,
- Pages clicked,
- Ads clicked

each with a user id and a timestamp.

What information would you like to get our of the data?

How would you use it?

Possible use:

- Query reformulations
- Custom Ads

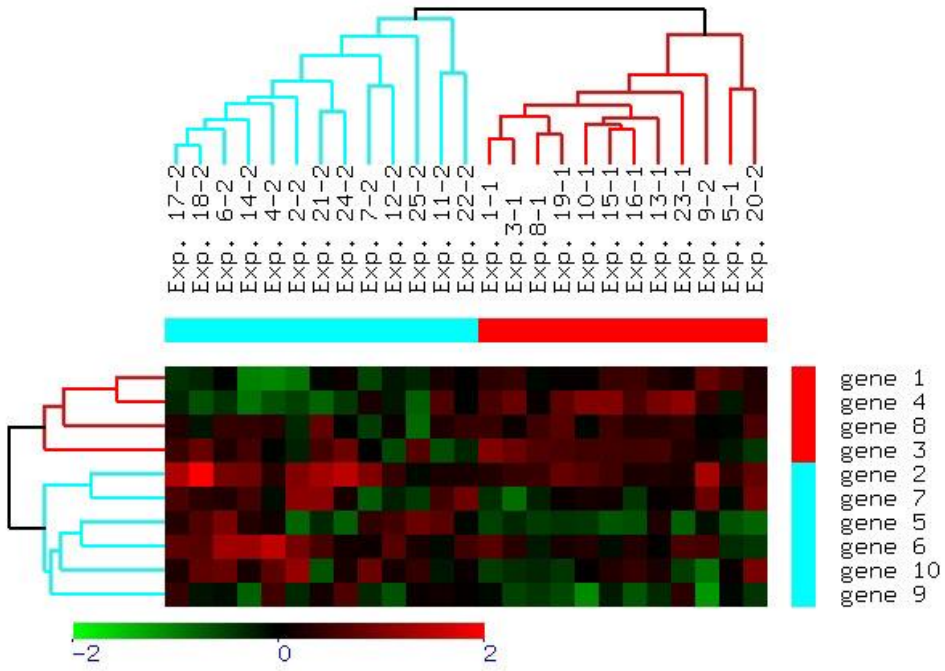
What can you do with the data?

Suppose you are biologist who has microarray expression data: thousands of genes, and their expression values over thousands of different settings (e.g. tissues).

- What information would you extract from it ?
- How would you use it?

Possible Use:

Groups of genes and tissues



What can you do with the data?

Suppose you are a stock broker and you observe the fluctuations of multiple stocks over time.

- What information would you extract from it ?
- How would you use it?

Possible Use:

Clustering of stocks

Correlation of stocks

Stock Value prediction

What can you do with the data?

You are the owner of a **social network**, and you have full access to the social graph, what kind of information do you want to get out of your **graph**?

Relevant Information:

- Who is the most important node in the graph?
- What is the shortest path between two nodes?
- How many friends two nodes have in common?
- How does information spread on the network?

What is Data Mining again?

What

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data analyst" (Hand, Mannila, Smyth)

"Data mining is the discovery of models for data" (Rajaraman, Ullman)

- We can have the following types of models
  - Models that explain the data (e.g., a single function)
  - Models that predict the future data instances.
  - Models that summarize the data
  - Models the extract the most prominent features of the data

Why

- We need the tools to analyze such data to get a better understanding of the world and advance science
- The amount and the complexity of data does not allow for manual processing of the data. We need automated techniques.

What can we do with data mining?

Course Program at a glace:

- Classification
- Neural Network
- Time Series Forecasting
- Clustering
- Recommendation systems
- Link analysis
- Frequent Itemset
- Finding Similar Items

What can we do with data mining?

Course Program at a glace:

- **Classification**
- Neural Network
- Time Series Forecasting
- Clustering
- Recommendation systems
- Link analysis
- Frequent Itemset
- Finding Similar Items

Classifiers —> IDM (3, 6)

- Introduzione al problema della classificazione
- Decision Trees
- K-nearest neighbours classifier
- Support Vector Machines
- Bayes and naive Bayes classifiers
- Ensemble classifiers
- The overfitting problem
- Class Imbalance
- Model Evaluation
- Hyperparameters
- Model Selection and Comparison

What can we do with data mining?

Course Program at a glace:

- Classification
- **Neural Network**
- Time Series Forecasting
- Clustering
- Recommendation systems
- Link analysis
- Frequent Itemset
- Finding Similar Items

Neural Network —> IDM (6)

- Introduction to neural networks.
- The Perceptron.
- Activation and Loss Functions.
- Optimization problem and gradient descent
- Stochastic Gradient Descent
- Multilayer Neural Networks

What can we do with data mining?

Course Program at a glace:

- Classification
- Neural Network
- **Time Series Forecasting**
- Clustering
- Recommendation systems
- Link analysis
- Frequent Itemset
- Finding Similar Items

Time Series Forecasting

- Naive approaches
- RNN
- LSTM
- GRU
- Evaluation procedure

What can we do with data mining?

Course Program at a glace:

- Classification
- Neural Network
- Time Series Forecasting
- **Clustering**
- Recommendation systems
- Link analysis
- Frequent Itemset
- Finding Similar Items

Clustering —> IDM (5)

- Introduzione al problema del clustering
- Curse of dimensionality
- K-means e K-means++
- Hierarchical clustering
- Density-based clustering (DBSCAN)
- Clusters Evaluation

What can we do with data mining?

Course Program at a glace:

- Classification
- Neural Network
- Time Series Forecasting
- Clustering
- **Recommendation systems**
- Link analysis
- Frequent Itemset
- Finding Similar Items

Recommendation systems —> MMD(9.1, 9.2, 9.3 9.5)

- Recommendations
- The long tail phenomenon
- Content-based raccomandation
- Collaborative filtering
- The Netflix Challenge

What can we do with data mining?

Course Program at a glace:

- Classification
- Neural Network
- Time Series Forecasting
- Clustering
- Recommendation systems
- **Link analysis**
- Frequent Itemset
- Finding Similar Items

Link analysis —> MMD(5.1, 5.4, 5.5)

- PageRank
- Link Spam
- Hubs and Authorities

What can we do with data mining?

Course Program at a glace:

- Classification
- Neural Network
- Time Series Forecasting
- Clustering
- Recommendation systems
- Link analysis
- **Frequent Itemset**
- Finding Similar Items

Frequent Itemset —> MMD(6.1,6.2,6.3) IDM(4)

- Modello market-basket
- Algoritmo A-priori
- Algoritmo PCY

What can we do with data mining?

Course Program at a glace:

- Classification
- Neural Network
- Time Series Forecasting
- Clustering
- Recommendation systems
- Link analysis
- Frequent Itemset
- **Finding Similar Items**

Finding Similar Items —> MMD(3.1,3.2,3.5)

- Document similarity
- Shingling: convertire documenti email in insiemi (k-shingles)
- Compressione mediante hashing di k-shingles
- Distances

Classification

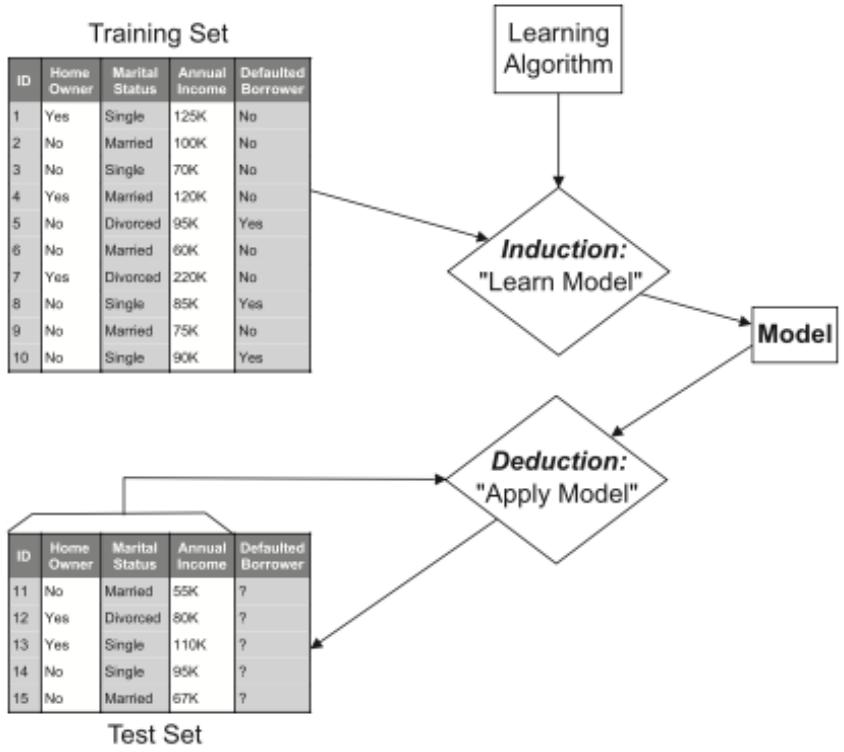
Given a collection of records (*training set*)

- Each record contains a set of *attributes*, one of the attributes is the *class*.

Find a function (*model)* getting in input the values of other attributes and providing as output the class attribute.

Goal: previously unseen records should be assigned a class as accurately as possible.

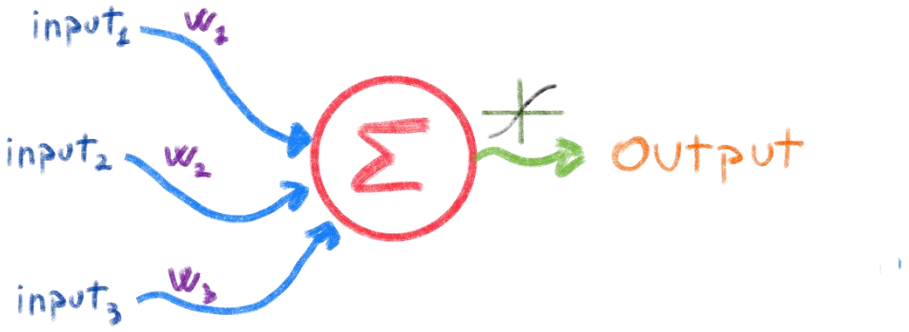
- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

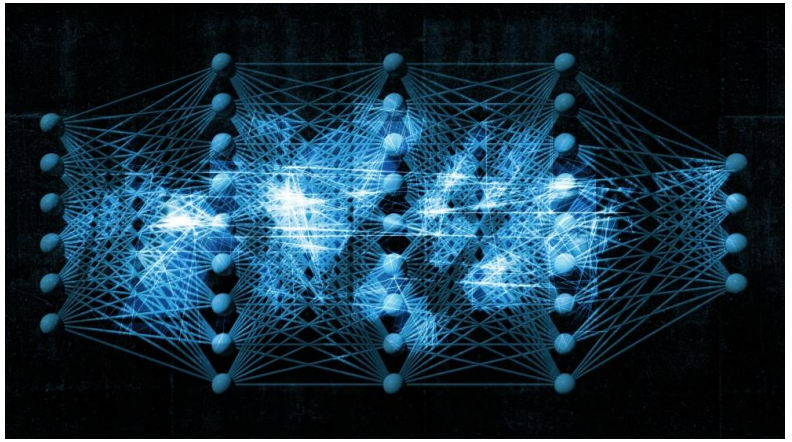


Neural Networks

Neural Networks (NN) are mostly a method

- Artificial neural networks (ANN) are powerful **classification** models that are able to learn highly complex and nonlinear decision boundaries purely from the data.
- Anyway, specific NN can be used for regression and forecasting task.
- Even generative approach are currently based on NN



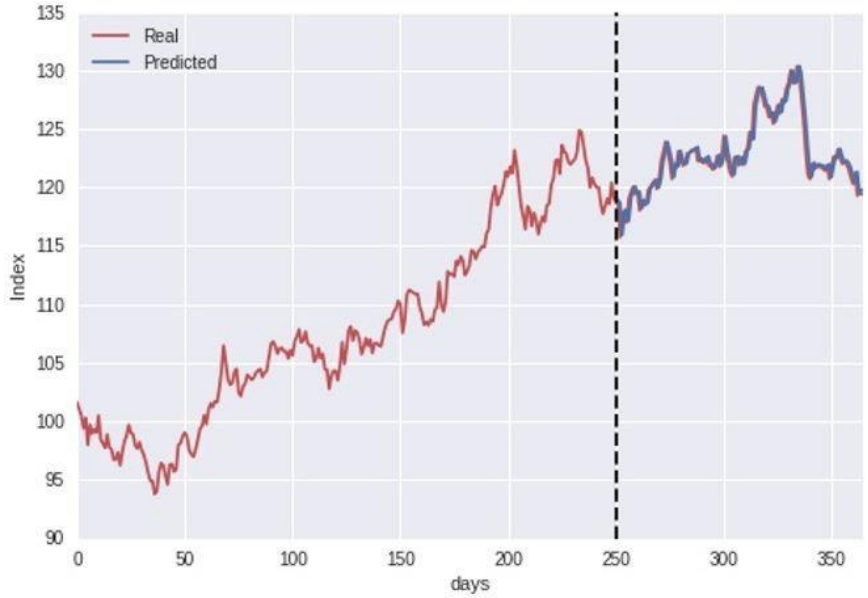


Time Series Forecasting

Time series forecasting is the process of using historical data to **predict** future values.

It involves analyzing **time-ordered data points** to identify patterns, trends, and seasonal variations, and then applying models or algorithms to *forecast upcoming data points*.

Time series forecasting is widely used in fields like *finance*, *weather forecasting*, and *demand planning*.



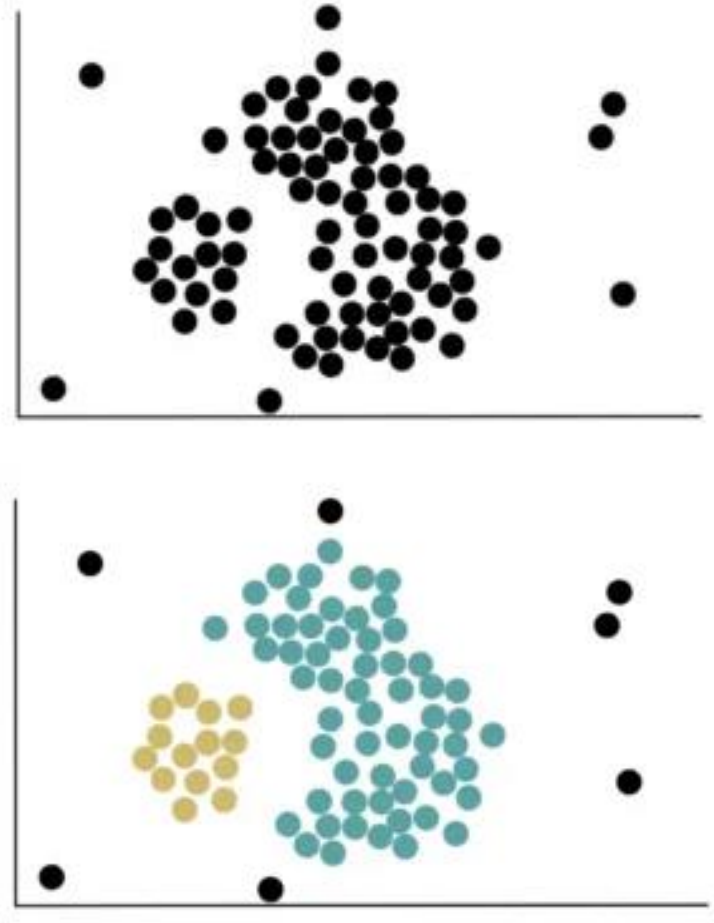
Clustering

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Similarity Measures?

- Euclidean Distance if attributes are continuous.
- Other Problem-specific Measures.



Recommendation Systems

Recommendation System involve an extensive class of Web applications that involve predicting user responses to options.

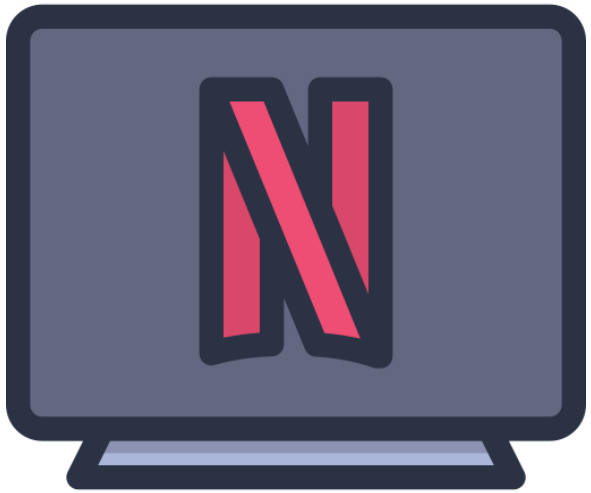
Example:

- Offering news articles to on-line newspaper readers, based on a prediction of reader interests.
- Offering customers of an on-line retailer suggestions about what they might like to buy, based on their past history of purchases and/or product searches.

Two groups of systems:

- Content based
- Collaborative filtering

The Netflix Challenge: A significant boost to research into recommendation systems was given when Netflix offered a prize of $1,000,000 to the first person or team to beat their own recommendation algorithm, called CineMatch, by 10%. After over three years of work, the prize was awarded in September, 2009.

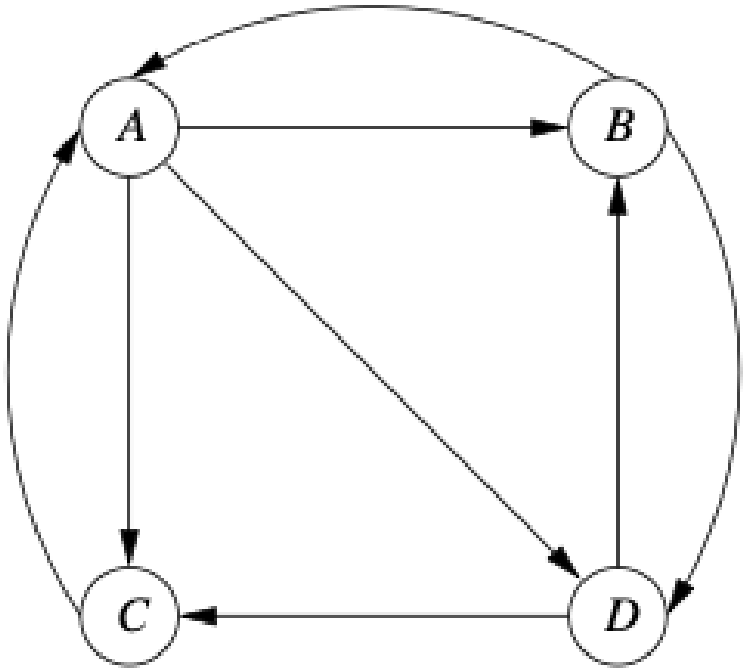


Link Analysis

One of the biggest changes in the decade following the turn of the century was the availability of efficient and accurate Web search.

The first revolution was introduced by **Google PageRank** algorithm

The war between those who want to make the **Web useful** and those who would exploit it for their own purposes is never over



Frequent Itemsets

Given a set of records each of which contain some number of items from a given collection;

- Identify sets of items (itemsets) occurring frequently together
- Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Itemsets Discovered:

{Milk,Coke}
{Diaper, Milk}

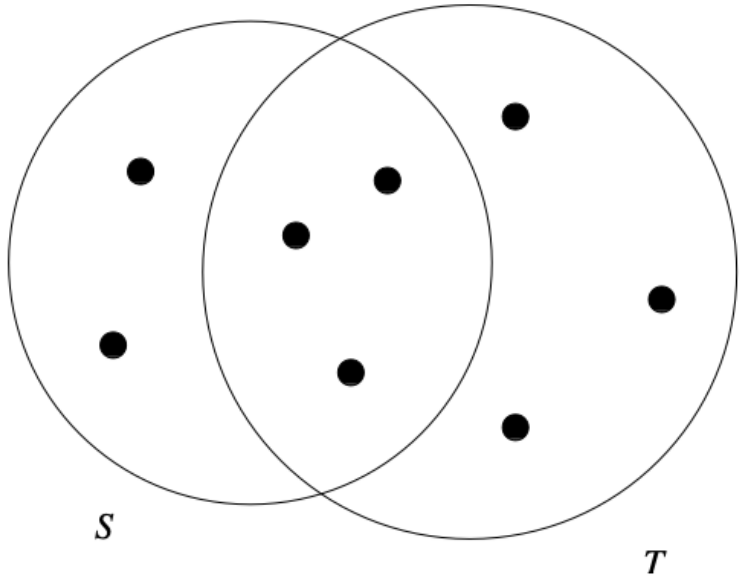
Rules Discovered:

{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Finding Similar Items

The naive approach to finding pairs of similar items requires us to look at every pair of items.

In case of large datasets, looking at all pairs of items may be prohibitive, even given an abundance of hardware resources.



Applications:

- Find near-duplicate pages in web (plagism)
- Document Similarity
- Article frome same source

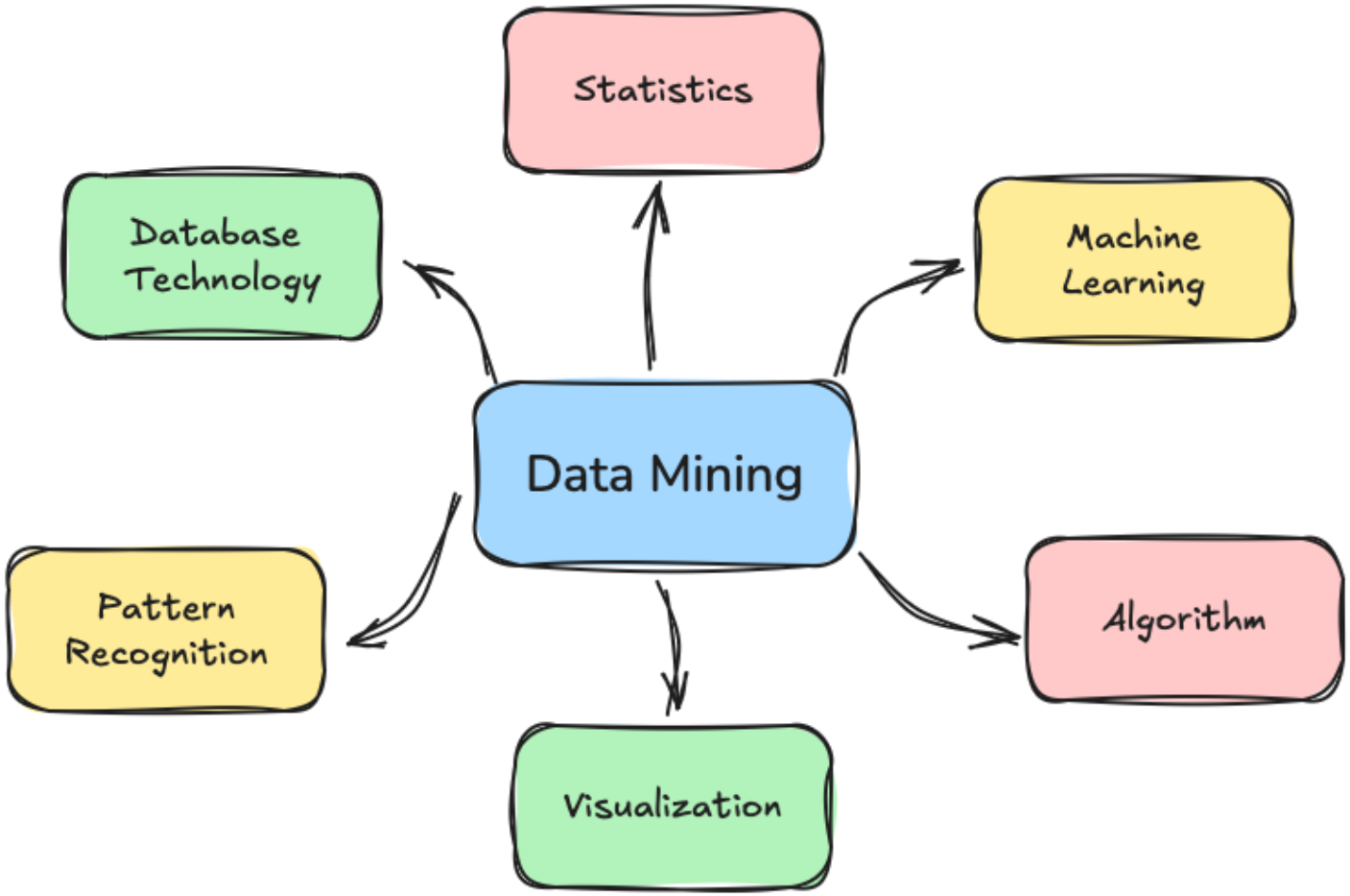
Link Analysis Ranking

Given a collection of web pages that are linked to each other, rank the pages according to importance (authoritativeness) in the graph

- Intuition: A page gains authority if it is linked to by another page.

Application: When retrieving pages, the authoritativeness is factored in the ranking.

Data Mining: Confluence of Multiple Disciplines



The data analysis pipeline

Mining is not the only step in the analysis process

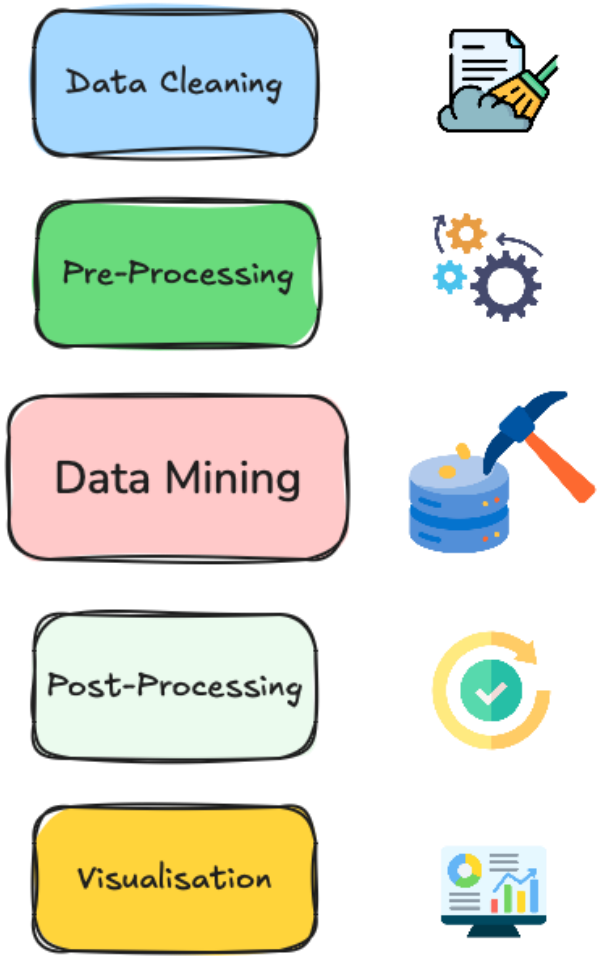
Data cleaning is required to make sense of the data

Preprocessing: Sampling, Dimensionality Reduction, Feature selection.

DATA MINING

Post-Processing: Make the data actionable and useful to the user. Statistical analysis of importance

Visualization: include all the techniques allowing to highlight the retrieved insights



It's a dirty work, but it is often the most important step for the analysis.

Preprocessing

Sampling is the main technique employed for data selection.

It is often used for both the preliminary investigation of the data and the final data analysis.

Processing the entire set of data of interest is too expensive or time consuming

Dimensionality Reduction: regards transforming data in a new space preserving the most informative data

Feature selection: Consist in exploiting only the attributes of interest avoiding the ones that are useless or just noise

Meaningfulness of Answers

A big data-mining risk is that you will “discover” patterns that are **meaningless**.

Bonferroni’s principle: If you look for *interesting patterns* in **more places** than the **available data**, you are bound to find rubbish.

Rhine Paradox

Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception.

He performed an experiment where subjects were asked to guess 10 hidden cards: red or blue.

He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right

He told these people they had ESP and called them in for another test of the same type.

Alas, he discovered that almost all of them had lost their ESP.

What did he conclude?

Rhine Paradox

Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception.

He performed an experiment where subjects were asked to guess 10 hidden cards: red or blue.

He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right

He told these people they had ESP and called them in for another test of the same type.

Alas, he discovered that almost all of them had lost their ESP.

What did he conclude?

He concluded that you shouldn't tell people they have ESP; it causes them to lose it.