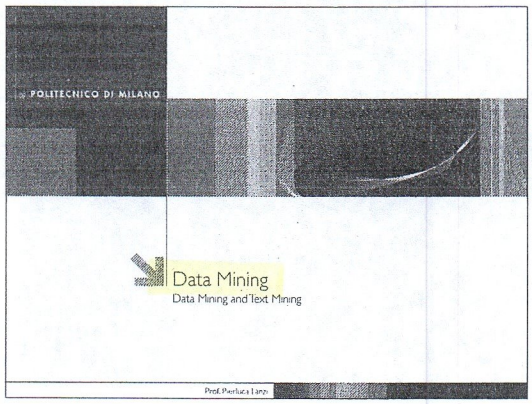


Motivations

Why?

"Necessity is the mother of invention"

Explosive Growth of Data

Pressing need for the automated analysis of massive data

Emerged in the late 1980s

Major developments in the mid 1990s

Several names over the years

Based on: Statistics, Machine Learning, Database technology

Evolution of Technology

4

- 1960s: data collection, database creation, & network DBMS
- 1970s: relational data model, relational DBMS implementation
- 1980s: RDBMS, advanced data models (extended-relational, OO, deductive, etc.); application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s: data mining, data warehousing, multimedia databases, and Web databases
- 2000s: stream data management and mining, web technology (XML, data integration), global information systems
- 2010s: social networks, NoSQL, unstructured data, etc.
- 2020s: …

Examples

5

- **Customer attrition**
  - Given customer information for the past months
  - Predict who is likely to attrite next month, or estimate customer value
- **Credit assessment**
  - Given a loan application
  - Predict whether the bank should approve the loan
- **Customer segmentation**
  - Given several information about the customers
  - Identify interesting groups among them
- **Community detection**
  - Who is discussing what?
- **Sentiment Analysis**
  - …

Big Data?

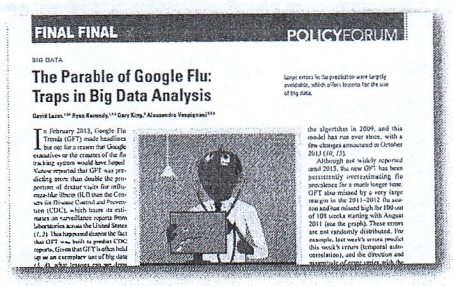
Massive datasets

Gigantic quantities of data are being collected in every domain imaginable: science, agriculture, healthcare, transport, sport …

Exponential increases in computing power has led to ever decreasing costs of instrumentalization and storage

Big Data

The quantity of data of quantity is so large that
(1) it facilitates predictions not previously possible, and
(2) it necessitates special processes to deal with
Quantity could be total storage, arrival rate, …



FINAL FINAL POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[5] Alessandro Vespignani[3,4,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. Nature reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) that the Centers for Disease Control and Prevention (CDC), which tracks reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data

the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (10, 11).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's report predicted this week's return (temporal autocorrelation), and the direction and

http://simplystatistics.org/2014/05/07/why-big-data-is-in-trouble-they-forgot-about-applied-statistics/

http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/?_r=0

http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

Machine Learning

"A computer program is said to learn from experience E with respect to some class of task T and a performance measure P, if its performance at tasks in T, as measured by P, improves because of experience E."

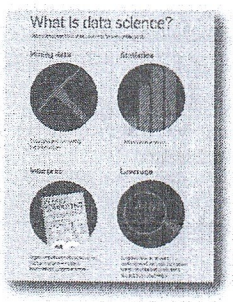
Machine Learning Paradigms

- Suppose we have the experience E encoded as a dataset,

$$D = x_1, x_2, x_3, \ldots, x_N$$

- Supervised Learning
  - Given the desired outputs $t_1, t_2, t_3, \ldots, t_N$ learns to produce the correct output given a new set of input
- Unsupervised learning
  - Exploits regularities in D to build a representation to be used for reasoning or prediction
- Reinforcement learning
  - Producing actions $a_1, a_2, a_3, \ldots, a_N$ which affect the environment, and receiving rewards $r_1, r_2, r_3, \ldots, r_N$ learn to act in order to maximize rewards in the long term

Data Science?



http://wikibon.org/blog/role-of-the-data-scientist/

- Missing data
- Statistic
- interpret
- leverage

Data Mining?

Data Mining

The non-trivial process of identifying (1) valid, (2) novel, (3) potentially useful, and (4) understandable patterns in data.

An Example Using Contact Lens Data

17

features

target variable

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	Hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

goal: given the features assign the right target

An example of possible pattern

if astigmatism = yes
and tear production rate = normal
and spectacle prescription = myope
then recommendation = hard

Is it a "Good" Pattern?

19

- **Is it valid?**
  - The pattern has to be valid with respect to a certainty level (rule true for the 86%)
- **Is it novel?**
  - Is the relation between astigmatism and hard contact lenses already well-known?
- **Is it useful? Is it actionable?**
  - The pattern should provide information useful to the bank for assessing credit risk
- **Is it understandable?**

What is the General Idea?

20

- Build computer programs that navigate through databases automatically, seeking regularities or patterns
- However …
  - Most patterns are uninteresting
  - Most patterns are spurious, inexact, or contingent on accidental coincidences in the particular dataset used
  - Real data is imperfect: Some parts will be garbled, and some will be missing
- Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful

Descriptive vs. Predictive

Are the models built for gaining insight?
(about what already happened)

→ descriptive data mining task

Or are they built for accurate prediction?
(about what might happen)

→ predictive data mining task

I want to analyze who are the customers that shop at my store more than twice a week

→ descriptive

I need a model to help me investing in the stock market!

→ predictive

...

There is a 3rd term :

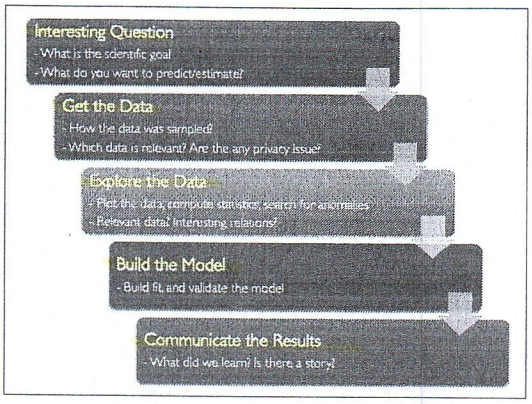
"Prescriptive"

use descriptive and predictive mining to recommend a course of action

using descriptive and predictive data mining techniques we have to give some recommendations on what to do next

---

The Process

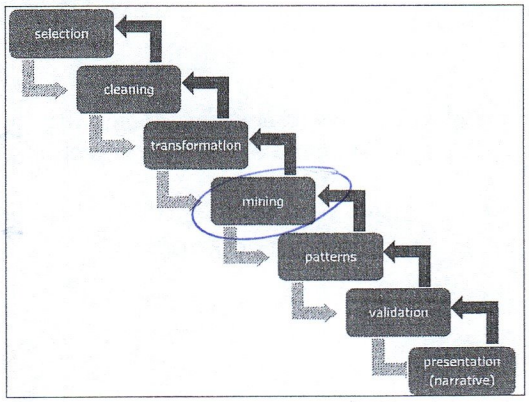


! When we present the results:

What did we learn?

Is there a story?

Can we build a narrative?



What Are the Main Steps?

- **Selection**
  - What are data we actually need to answer the posed question?
- **Cleaning**
  - Are there any errors or inconsistencies in the data we need to eliminate?
- **Transformation**
  - Some variables might be eliminated because equivalent to others
  - Some variables might be elaborated to create new variables (e.g. birthday to age, daily measures into weekly/monthly measures, log?)
- **Mining**
  - Select the mining approach: classification, regression, association, etc.
  - Choose and apply the mining algorithm(s)
- **Validation**
  - Are the patterns we discovered sound? According to what criteria?
  - Are the criteria sound? Can we explain the result?
- **Presentation & Narrative**
  - What did we learn? Is there a story to tell? A take-home message?

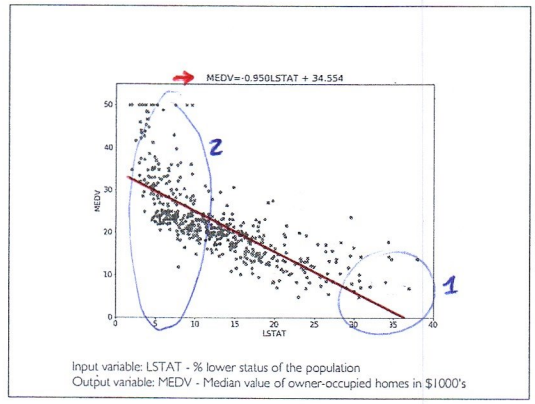
Data Mining Tasks

Prediction & Regression

Boston Housing Dataset

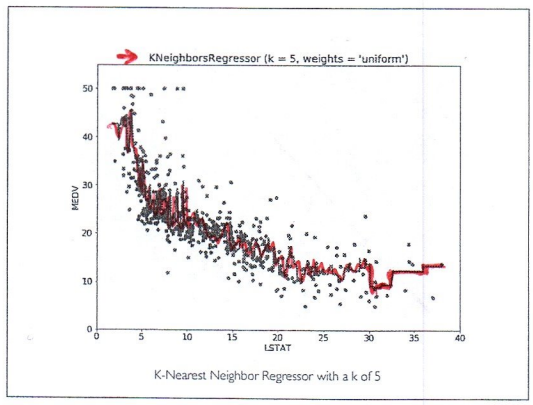
Goal: how much a house was valued in Boston

- Information concerning housing in the area of Boston (MA)
  - collected by U.S Census Service:
  - https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
- 506 cases and 14 variables:
  - CRIM - per capita crime rate by town;
  - CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise);
  - TAX - full-value property-tax rate per $10,000;
  - PTRATIO - pupil-teacher ratio by town;
  - LSTAT - % lower status of the population;
  - ...
  - MEDV - Median value of owner-occupied homes in $1000's
- MEDV is the target variable



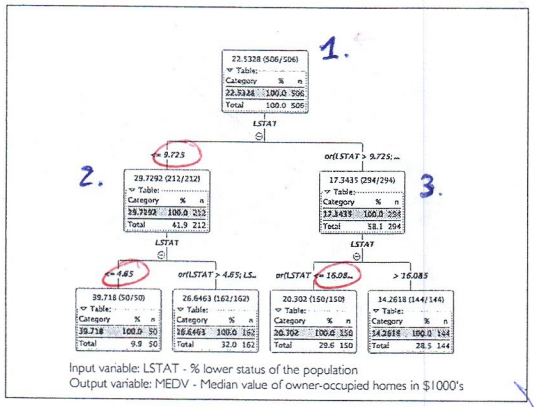
Input variable: LSTAT - % lower status of the population
Output variable: MEDV - Median value of owner-occupied homes in $1000's

What about the narrative of this solution? LSTAT is the percentage (%) of the lower status of the population. We can say that if there are several poor people (1) the value of the house is very low. If the percentage of poor people is very low (2) we're in a rich part of the town and the houses have high values.

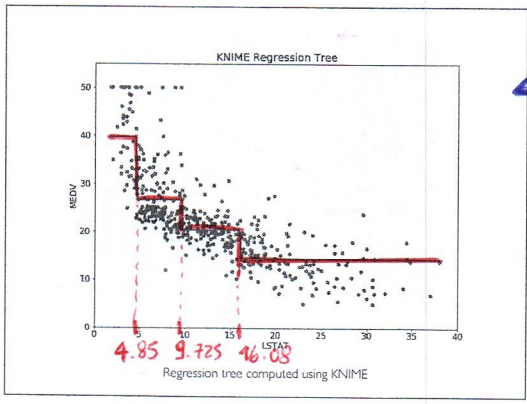


K-Nearest Neighbor Regressor with a k of 5

Another approach is to determine the price of a house basing on the houses in the neighborhood. Here the prediction is build as an average of the 5 nearest points to where we are. If this approach works better we can say that similarities is an important aspect of the problem domain: "if two houses are similar we have more informations (better informations) than just the linear relation".



Another approach: regression tree. It's a tree for which at each node we make a decision. At the beginning there is a root (1). The average value is 22.5328. How can we have a better prediction? We introduce a threshold: if the LSTAT is below 9.725 then the average value is 29.7292, otherwise the average value is 17.3435. By splitting we find subspaces where the results differ a lot. After the first split we can go on splitting. We can split 2. and 3. in two each.



Regression tree computed using KNIME

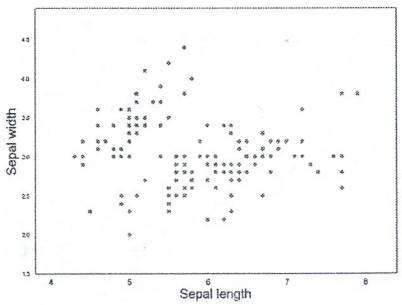
These splits generate a model (piecewise linear model)

What if this works better than the others? We can say that there's a linear relation but it's not as powerful as understanding what are the subspaces in which we have to divide data.

Classification

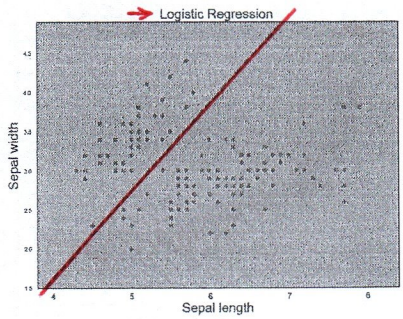
: if we don't have to predict a number but a label

We have to predict a color: some of the points are green and some are blue. We need to build a model that allows to say: "if your datum is here then it'll be blue/green". We want a model that is accurate in predicting what we know and accurate in predicting what we don't know (new points).



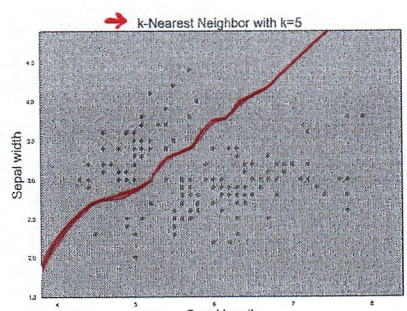
Iris dataset plotted using only two variables and two classes.

Easiest model: linear model.



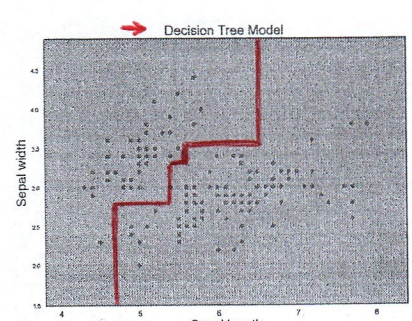
Logistic Regression Model

Another option: **k** nearest neighborhoods. To assign a class to a point we look for the **k** nearest point and the assigned class will be the most frequent among the **k** points.



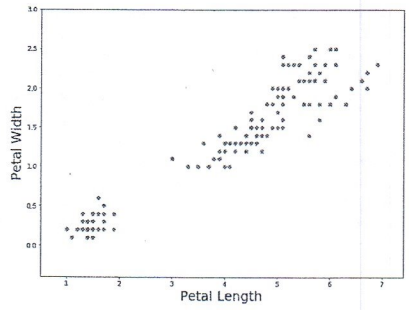
K-Nearest Neighbor model with a k of 5

Another option: decision tree model. The decision tree model considers hyperplanes.



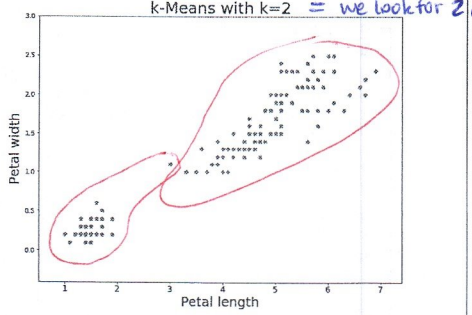
Decision Tree Model

Clustering : we look at the data and we try to find patterns and analogies



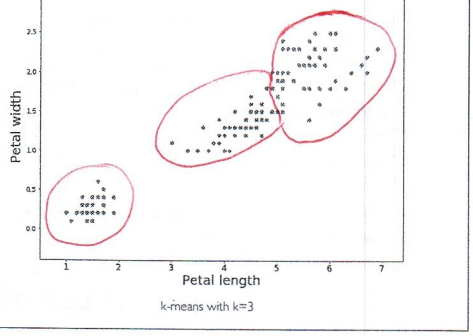
Iris dataset plotted using only two variables without taking into account the known class.

The assumptions we make if the points are nearby then they're very similar. In the PROBLEM SPACE the distance between points is a similarity index.



k-mean with k=2

= we look for 2 groups



k-means with k=3

= we look for 3 groups

Associations

Data Mining Tasks: Associations

Receives after buying stuff:

- {citrus fruit, semi-finished bread, margarine, ready soups}
- {tropical fruit, yogurt, coffee}
- {whole milk}
- {pip fruit, yogurt, cream cheese, meat spreads}
- {other vegetables, whole milk, condensed milk, long life bakery product}
- {whole milk, butter, yogurt, rice, abrasive cleaner}
- {other vegetables, UHT-milk, rolls/buns, bottled beer, liquor (appetizer)}
- {whole milk, cereals}
- ....

Are there interesting products associations?

Frequent Itemsets (Minimum Support 0.1%)


```
     items                      support     count
[1]  {hair spray}               0.001118454 11
[2]  {flower soil/fertilizer}   0.001931876 19
[3]  {rubbing alcohol}          0.001016777 10
[4]  {frozen fruits}            0.001220132 12
[5]  {prosecco}                 0.002033554 20
[6]  {honey}                    0.001525165 15
[7]  {cream}                    0.001321810 13
[8]  {decalcifier}              0.001525165 15
[9]  {organic products}         0.001626843 16
[10] {soap}                     0.002643620 26
...
```


Association Rules (Support, Confidence, Lift)


```
{rice, sugar} => {whole milk}
Support 0.001220132
Confidence 1
Lift 3.913649

{canned fish, hygiene articles} => {whole milk}
Support 0.001118454
Confidence 1
Lift 3.913649

{root vegetables, butter, rice} => {whole milk}
Support 0.001016777
Confidence 1
Lift 3.913649
...
```


Other Tasks

- **Outlier analysis**
  - An outlier is a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- **Trend and evolution analysis**
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- **Text Mining, Topic Modeling, Graph Mining, Data Streams**
- **Sentiment Analysis, Opinion Mining, etc.**
- **Other pattern-directed or statistical analyses**

Relevant Issues

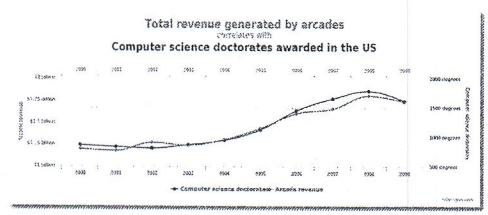
Are all the "Discovered" Patterns Interesting?

- Data Mining may generate thousands of patterns, but typically not all of them are interesting.
- Interestingness measures
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- Objective vs. subjective interestingness measures
  - Objective measures are based on statistics and structures of patterns
  - Subjective measures are based on user's belief in the data, e.g., unexpectedness, novelty, etc.

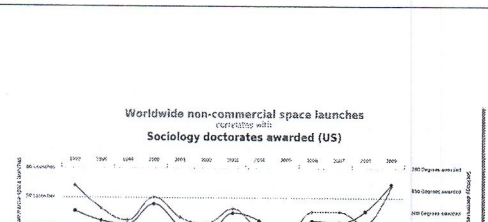
Can We Find All and Only Interesting Patterns?

- Completeness
  - Find all the interesting patterns
  - Can a data mining system find all the interesting patterns?
  - Association vs. classification vs. clustering
- Optimization
  - Search for only interesting patterns:
  - Can a data mining system find only the interesting patterns?
  - Two approaches: (1) first general all the patterns and then filter out the uninteresting ones; (2) generate only the interesting patterns—mining query optimization

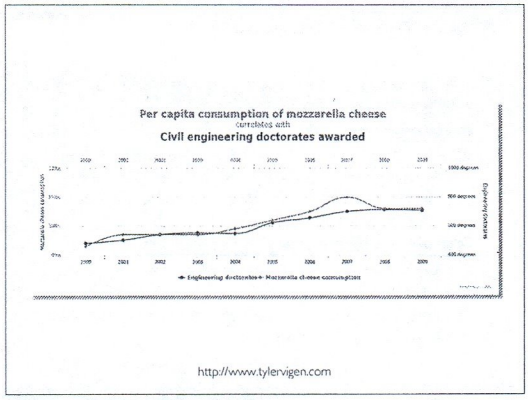
Pitfalls

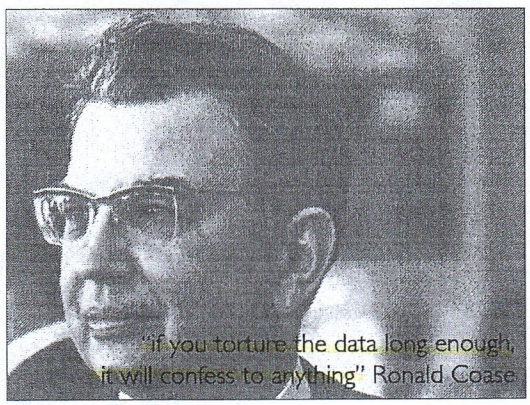


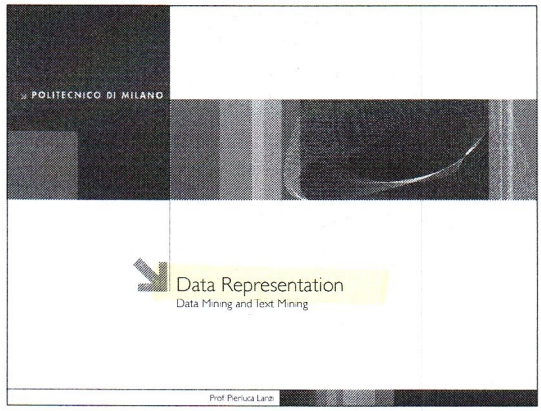
http://www.tylervigen.com



http://www.tylervigen.com


Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded
Engineering doctorates
Mozzarella cheese consumption
http://www.tylervigen.com


"if you torture the data long enough,
it will confess to anything" Ronald Coase



describing data

The Weather Dataset

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	no
Sunny	80	90	True	no
Overcast	83	86	False	yes
Rainy	70	96	False	yes
Rainy	68	80	False	yes
Rainy	65	70	True	no
Overcast	64	65	True	yes
Sunny	72	95	False	no
Sunny	69	70	False	yes
Rainy	75	80	False	yes
Sunny	75	70	True	yes
Overcast	72	90	True	yes
Overcast	81	75	False	yes
Rainy	71	91	True	no

Instances, Attributes, Concepts

- **Instances (aka observations, cases, records, items, examples)**
  - The atomic elements of information from a dataset
  - Each row in previous table corresponds to an instance
- **Attributes (aka variables, features, independent variables)**
  - Measures aspects of an instance
  - Each instance is composed of a certain number of attributes
  - Each column in previous table contains values of an attribute
- **Concept (aka class, target variable, dependent variable)**
  - Special content inside the data
  - Kind of things that can be learned
  - Intelligible and operational concept description
  - Last column of previous table was the class

attribute types

Attributes

- **Numeric Attributes**
  - Real-valued or integer-valued domain
  - Interval-scaled when only differences are meaningful (e.g., temperature)
  - Ratio-scaled when differences and ratios are meaningful (e.g., Age)
- **Categorical Attributes**
  - Set-valued domain composed of a set of symbols
  - Nominal when only equality is meaningful (e.g., domain(Sex) = { M, F})
  - Ordinal when both equality (are two values the same?) and inequality (is one value less than another?) are meaningful (e.g., domain(Education) = { High School, BS, MS, PhD})

Numerical Attributes

- Not only ordered but measured in fixed and equal units
- Examples
  - Attribute "temperature" expressed in degrees
  - Attribute "year"
- Characteristics
  - Difference of two values makes sense
  - Sum or product doesn't make sense
  - Zero point is not defined
- Sometimes they are divided into "discrete" and "continuous"

Nominal Attributes (or Categorical)

- Values are distinct symbols that serve only as labels or names
- Example
  - Attribute "outlook" from weather data
  - Values: "sunny", "overcast", and "rainy"
- Characteristics
  - No relation is implied among nominal values
  - No ordering
  - No distance measure
  - Only equality tests can be performed

Other Types of Attributes

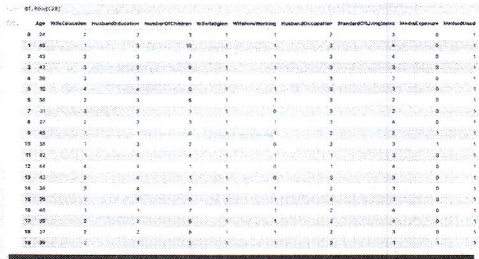
- **Ordinal Attributes** → basically : categorical + ordering
  - Categorical attributes with an imposed order on values
  - No distance between values defined
  - For instance, temperature encoded as "hot","mild", and "cool"
  - Size encoded as "small", "medium", "large", and "jumbo"
- **Ratio Attributes**
  - Numerical attributes for which the measurement scheme defines a zero point (e.g., an attribute representing distance)
- **Binary Attributes**
  - Represented by just two values 0/1

example

Contraceptive Method Choice Data Set

11

https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice



What are the attribute types of these variables?

Contraceptive Method Choice Data Set (Data Description)

12

- Wife's age (numerical)
- Wife's education (categorical) 1=low, 2, 3, 4=high → it's actually a sorting (and to ordinal)
- Husband's education (categorical) 1=low, 2, 3, 4=high → same (↑)
- Number of children ever born (numerical)
- Wife's religion (binary) 0=Non-Islam, 1=Islam → it's actually nominal (categorical)
- Wife's now working? (binary) 0=Yes, 1=No → same (↑)
- Husband's occupation (categorical) 1, 2, 3, 4 → this is categorical (nor numerical, nor ordinal)
- Standard-of-living index (categorical) 1=low, 2, 3, 4=high
- Media exposure (binary) 0=Good, 1=Not good → categorical
- 10. Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term

What are the attribute types of these variables (now)?

In a decision tree is more convenient to use education as number and occupation as categorical. The point is: if we treat education as categorical we lose information on how much a person is educated (it will remain only classes), instead, there's no hidden information in the occupation (4 is not better than 1).

another perspective

Attribute types

14

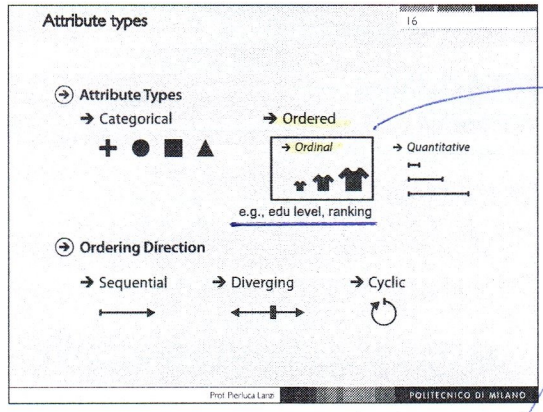
- Attribute Types
  - Categorical
  - Ordered
    - Ordinal
    - Quantitative
- Ordering Direction
  - Sequential
  - Diverging
  - Cyclic

Attribute types

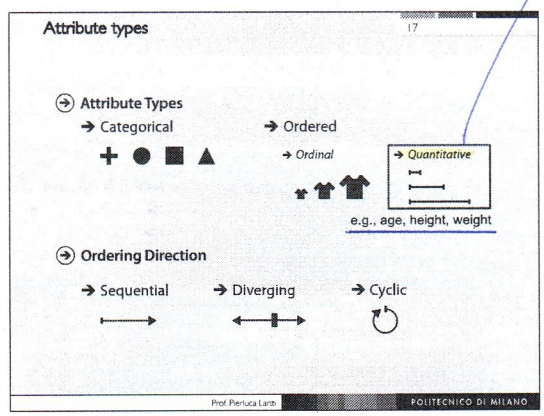
15

- Attribute Types
  - Categorical
    - e.g., gender, race, eye color
  - Ordered
    - Ordinal
    - Quantitative
- Ordering Direction
  - Sequential
  - Diverging
  - Cyclic

having a categorical attribute is like having a shape: we just can show them

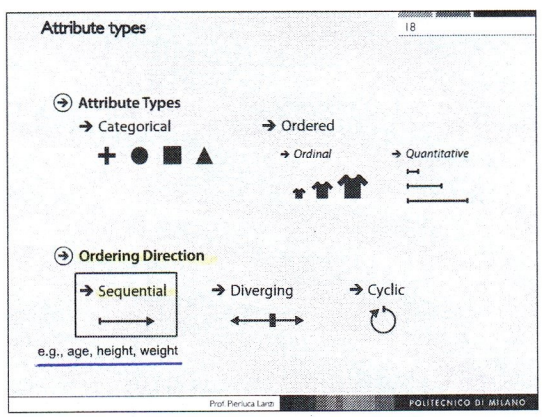


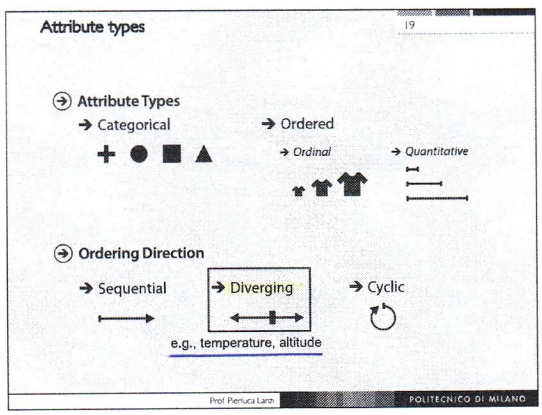
When we have ordinals (ordered) then we can sort them or create a representation that computes them (quantitative).

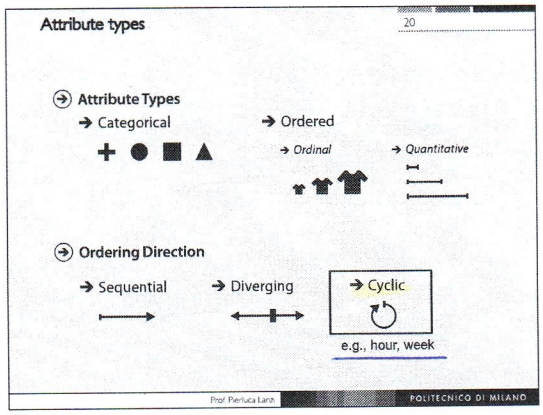


The fact that we can have an order and a quantitative variables, we can have different ordering directions:

(↓)







hierarchies

Some attributes may have an internal hierarchical structure

For example, dates, mail addresses, spatial regions, taxonomies, etc.

day/month/year

nome.cognome@mail.polimi.it

who is writing is a student

more specific from politecnico

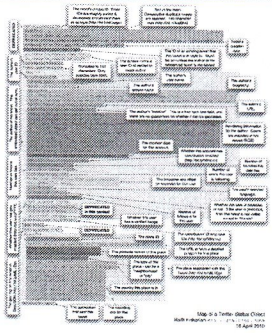
the mail is from Italy

missing values

Why Missing Values Exist?

24

- Faulty equipment, incorrect measurements, missing cells in manual data entry, censored/anonymous data
- Review scores for movies, books, etc.
- Very frequent in questionnaires for medical scenarios
- Censored or anonymous data
- Data from rich representations for which many of the fields might not be used (e.g., Twitter data)



http://socialmedia-class.org/twittertutorial.html

Missing Values

26

- Missing value may have significance in itself
  - E.g. missing test in a medical examination or an empty field in a questionnaire
- They are frequently indicated by out-of-range entries (e.g. max/min float), NaN or special values*(e.g., zero)
- Most schemes assume that is not the case and "missing" may need to be coded as additional value
- Does absence of value have some significance?
  - If it does, "missing" is a separate value
  - If it does not, "missing" must be treated in a special way

What Types of Missing Values?

27

- **Missing not at random (MNAR)**
  - Distribution of missing values depends on missing value
  - E.g., respondents with high income less likely to report it
- **Missing at random (MAR)**
  - Distribution of missing values depends on observed attributes, but not missing value
  - E.g., men less likely than women to respond to question about mental health

Here it depends on the value (if someone earns more then it's more likely that this someone is not going to tell).

Here it depends on the whole column of the attribute: if someone has a certain feature (e.g. being male) then it's more likely to not respond. So, basically it depends on the type of the target, not on some specific values of the target.

What Types of Missing Values?

28

- **Missing completely at random (MCAR)**
  - Distribution of missing values does not depend on observed attributes or missing value
  - E.g., survey questions randomly sampled from larger set of possible questions
- **Identifying MNAR and MAR can be difficult often requires domain knowledge**

Dealing with Missing Values

29

- **Use what you know**
  - Why data is missing
  - Distribution of missing data →
- **Decide on the best strategy to yield the least biased estimates**
  - Deletion Methods (listwise deletion, pairwise deletion)
  - Single Imputation Methods (mean/mode substitution, dummy variable method, single regression)
  - Model-Based Methods (maximum Likelihood, multiple imputation

we consider all the istances with missing values and we check if there are some dependencies. (e.g. in the case of mental health questions we notice that most of missing values are men → it's an information)

Strategies for missing values handling

30

- The handling of missing data depends on the type
- Discarding all the examples with a missing values
  - Simplest approach
  - Allows the use of unmodified data mining methods
  - Only practical if there are few examples with missing values. Otherwise, it can introduce bias
- Fill in the missing value manually ☺
- Convert the missing values into a new value
  - Use a special value for it
  - Add an attribute that indicates if value is missing or not
  - Greatly increases the difficulty of the data mining process
- Imputation methods
  - Assign a value to the missing one, based on the rest of the dataset. Use the unmodified data mining methods.

(Possible issue:)
If the missing values are MNAR we're losing informations. If we eliminate the missing values and it turns out that all people earning a lot did not say it then we're erasing a part of the population from our analysis.

1.

Do Not Impute (DNI)

31

- Simply use the default policy of the data mining method
- Works only if the policy exists
- Some methods can work around missing data

1.1.

List-wise Deletion (Complete Case Analysis)

32

delete all the istances with missing values

- Only analyze cases with available data on each variable
- Simple, but reduces the data
- Comparability across analyses
- Does not use all the information
- Estimates may be biased if data not MCAR

we need massive amount of data

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

1.2.

Pairwise deletion (Available Case Analysis)

33

If we use only 2 attributes out of 3 then we consider missing values only in those 2 that we use. (consider and delete the istances)

- Analysis with all cases in which the variables of interest are present
- Example
  - When using only the first two variables, the missing values of the third variable are not considered
- Advantage
  - Keeps as many cases as possible for each analysis
  - Uses all information possible with each analysis
- Disadvantage
  - Can't compare analyses because sample different each time

If we do 2 different analysis and we consider 2 different set of variables then the eliminated instances will be different in the 2 sets → non comparable analysis

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

2.

Single Imputation Methods

34

- **Mean/mode substitution** (most common value)
  - Replace missing value with sample mean or mode
  - Run analyses as if all complete cases
  - Advantages: Can use complete case analysis methods
  - Disadvantages: Reduces variability
- **Dummy variable control**
  - Create an indicator for missing value (1=value is missing for observation; 0=value is observed for observation)
  - Impute missing values to a constant (such as the mean)
  - Include missing indicator in the algorithm
  - Advantage: uses all available information about missing observation
  - Disadvantage: results in biased estimates, not theoretically driven
- **Regression Imputation**
  - Replaces missing values with predicted score from a regression equation.

if we don't want to lose the information that the missing data has been replaced (so if we pass the dataset to someone else, also this someone else will know that there were missing data)

Imputation methods

35

- Extract a model from the dataset to perform the imputation
  - Suitable for MCAR and, to a lesser extent, for MAR
  - Not suitable for NMAR type of missing data
- For NMAR we need to go back to the source of the data to obtain more information
- Survey of imputation methods available at
  http://sci2s.ugr.es/MVDM/index.php
  http://sci2s.ugr.es/MVDM/biblio.php

inaccurate values

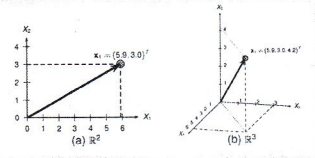
Inaccurate Values

- Data has not been collected for mining it
- Errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes, thus values need to be checked for consistency
- Typographical and measurement errors in numeric attributes, thus outliers need to be identified
- Errors may be deliberate (e.g. wrong zip codes)

the geometric view

The Geometrical View of the Data

- When the data contains only numerical values
  - Every row can be viewed as a point in a d-dimension space
  - Every column as a point in a n-dimensional space



From categorical attributes to numerical ones and Vice-versa

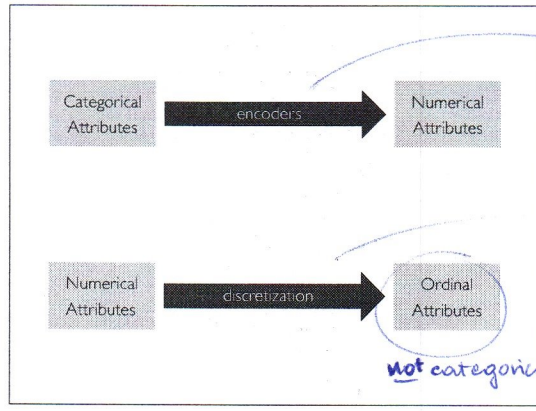
Why do we care about data types?

They influence the type of statistical analyses and visualization we can perform

Some algorithms and functions fit some specific data types best

Check for valid values

Deal with missing values, etc.

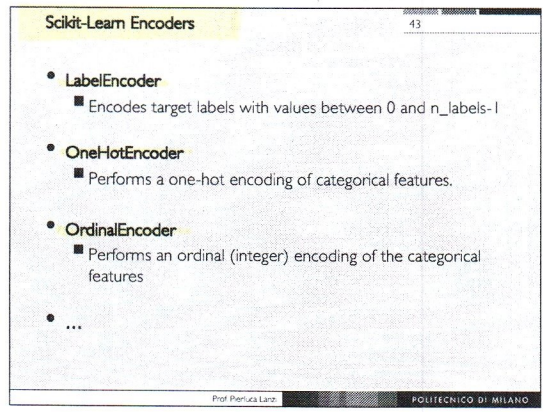


we have to map labels into numbers, there are 3 ways to do it: *

It basically means to create intervals and use them as ordinal attributes

not categorical

*

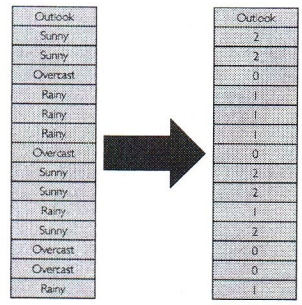


1. Label Encoder

Map a categorical variable described by n values into a numerical variables with values from 0 to n-1

For example, attribute Outlook would be replaced by a numerical variables with values 0, 1, and 2

Remember to keep tracking of the encoding



Label Encoder for the Outlook attribute.

Warning

By replacing a label with a number might influence the process in unexpected ways

In the example, by assigning 0 to overcast and 2 to sunny we give a higher weight to the latter

What happens if we then apply a regression model?

Would the result change with different assigned values?

If we apply label encoding, we should store the mapping used for each attribute to be able to map the encoded data into the original ones

we should always check that different labels encoding do not create too much different results (if so, the labels encoding is influencing and this is not good)

2. One Hot Encoding

Map each categorical attribute with n values into n binary 0/1 variables

Each one describing one specific attribute values

For example, attribute Outlook is replaced by three binary variables Sunny, Overcast, and Rainy

Outlook
Sunny
Sunny
Overcast
Rainy
Rainy
Rainy
Overcast
Sunny
Sunny
Rainy
Sunny
Overcast
Overcast
Rainy

Outlook_Overcast	Outlook_Rainy	Outlook_Sunny
0	0	1
0	0	1
1	0	0
0	1	0
0	1	0
0	1	0
1	0	0
0	0	1
0	0	1
0	1	0
0	0	1
1	0	0
1	0	0
0	1	0

One Hot Encoding for the Outlook attribute.

Warning

One hot encoding assign the same numerical value (1) to all the labels

But it can generate a massive amount of variables when applied to categorical variables with many values

We will discuss discretization later ...

(3.)

data format

Data Format

- Most commercial tools have their own proprietary format
- Most tools import excel files and comma-separated value files

```
Year,Make,Model,Length
1997,Ford,E350,2.34
2000,Mercury,Cougar,2.38
```

```
Year;Make;Model;Length
1997;Ford;E350;2,34
2000;Mercury;Cougar;2,38
```


Attribute-Relation File Format (ARFF)


```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Missing Values in ARFF


```
@relation labor
@attribute 'duration' real
@attribute 'wage-increase-first-year' real
@attribute 'wage-increase-second-year' real
@attribute 'wage-increase-third-year' real
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' real
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' real
@attribute 'shift-differential' real
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' real
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disbility-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
@data
1.5,2,?,[4],?,2,2,?,1,1,'average',?,?,'yes',?,'good'
2,4,5,5.8,?,?,?,35,'ret_allw',?,?,'yes',1,1,'below_average',?,'full',?,'full','good'
?,?,?,?,?,38,'empl_contr',2.5,?,1,1,'generous','yes','half','yes','half','good'
3,3.7,4.5,'tc',?,?,?,?,'yes',3,?,?,?,'yes',?,'good'
```


DSPL: Dataset Publishing Language

- Open format by Google available at
  http://code.google.com/apis/publicdata/
  https://code.google.com/archive/p/dspl/downloads
- Use existing data: add an XML metadata file existing CSV
- Read by the Google Public Data Explorer, which includes animated bar chart, motion chart, and map visualization
- Allow linking to concepts in other datasets
- Geo-enabled: allows adding latitude and longitude data to your concept definitions

model representation

Predictive Model Markup Language

- XML-based markup language developed by the Data Mining Group (DMG) to provide a way for applications to define models related to predictive analytics and data mining
- The goal is to share models between applications
- Vendor-independent method of defining models
- Allow to exchange of models between applications.
- PMML Components: data dictionary, data transformations, model, mining schema, targets, output

data repositories

Publicly Available Datasets

- **UCI repository**
  - http://archive.ics.uci.edu/ml/
  - Probably the most famous collection of datasets
- **Kaggle**
  - http://www.kaggle.com/
  - It is not a static repository of datasets, but a site that manages Data Mining competitions
  - Example of the modern concept of crowdsourcing
- **KDNuggets**
  - http://www.kdnuggets.com/datasets/
- **Google Data Search**
  - https://datasetsearch.research.google.com

Readings

- "Data Mining and Analysis" – Chapter 1
- "Mining of Massive Datasets" – Chapter 1
- Survey of imputation methods
  http://sci2s.ugr.es/MVDM/index.php

POLITECNICO DI MILANO

Data Exploration and Visualization

Data Mining and Text Mining

Prof. Pierluca Lanzi

Before trying anything,
you should know your data!

What is Data Exploration?

- Preliminary exploration of the data aimed at identifying their most relevant characteristics
- What the key motivations?
  - Help to select the right tool for preprocessing and data mining
  - Exploit humans' abilities to recognize patterns not captured by automatic tools
- Related to Exploratory Data Analysis (EDA)
  - Invented/advocated by statistician John Tukey

Exploratory Data Analysis

- "An approach of analyzing data to summarize their main characteristics without using a statistical model or having formulated a prior hypothesis."

- "Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments."
- Resources
  - Seminal book: "Exploratory Data Analysis" by Tukey
    https://books.google.it/books/about/Exploratory_Data_Analysis.html?id=UT9dAAAAIAAJ
  - Online introduction in Chapter 1 of NIST Engineering Statistics Handbook
    http://www.itl.nist.gov/div898/handbook/index.htm
  - Wikipedia article
    http://en.wikipedia.org/wiki/Exploratory_data_analysis

What Data Exploration Techniques?

- Exploratory Data Analysis (as originally defined by Tukey), was mainly focused on
  - Visualization
  - Clustering and anomaly detection (viewed as exploratory techniques)
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory
- In this section, we focus on data exploration using
  - Summary statistics
  - Visualization
- We will return to data exploration when discussing clustering

Summary Statistics

Summary Statistics

- What are they?
  - Numbers that summarize properties of the data
- Summarized properties include
  - Location, mean, spread, skewness, standard deviation, mode, percentiles, etc.
- Most summary statistics can be calculated in a single pass

Frequency and Mode

- The frequency of an attribute value
  - The percentage of time the value occurs in the data set
  - For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points
- However, the mean is very sensitive to outliers
- Thus, the median or a trimmed mean is also commonly used

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Percentiles

- For continuous data, the notion of a percentile is very useful
- p-th percentile
  - Given an ordinal or continuous attribute x and a number p
  - p-th percentile is a value x_p of x such that p% of the observed values of x are less than x_p
- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$

Trimean & Truncated Mean

- **Trimean**
  - It is the weighted mean of the first, second and third quartile

$$TM = \frac{x_{25} + 2x_{50} + x_{75}}{4}$$

- **Truncated** Mean
  - Discards data above and below a certain percentile
  - For example, below the 5th percentile and above the 95th percentile
- **Interquartile Mean**
  - Truncate data at 25th and 75th percentile
  - If the data $(x_1, ..., x_n)$ is sorted by value we have:

$$X_{IQM} = \frac{2}{n} \sum_{i=0.25n+1}^{0.75n} x_i$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \overline{x})^2$$

- However, this is also sensitive to outliers, so that other measures are often used

$$\text{AAD}(x) = \frac{1}{m} \sum_{i=1}^{m} |x_i - \overline{x}|$$

$$\text{MAD}(x) = median\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

useful for variable selection: if a variable has low variance we can consider to get rid of it since it's not informative! Extreme case: constant values, null variance.

Correlation Analysis

- Given two attributes, measure how strongly one attribute implies the other, based on the available data
- Use correlation measures to estimate how predictive one attribute is of another

Karl Pearson

- Linear correlation
  - We often look for linear relationship between variables not all variables are linearly related, of course!
  - Linear correlation measures are symmetric
  - They can be positive (high values of one attribute are likely given high values of the other)
  - Or negative (high values are predictive of low values of the other variable)
  - Latter usually referred to as anti-correlation

Correlation Analysis

- Given two attributes it measure how strongly one attribute implies the other, based on the available data
- **Numerical Variables**
  - For two numerical variables, we can compute the correlation coefficient, Pearson's product moment coefficient
- **Ordinal Variables**
  - We can compute Spearman rank correlation coefficient
- **Categorical Variables**
  - For two categorical variables, A and B, we can compute $\chi 2$ (chi-square) statistic test which tests the hypothesis that A and B are independent
- **Binary Variables**
  - Compute Point-biserial correlation

Correlation vs Causation

- **Correlation does not imply causation**
  - Just because value of one attribute is highly predictive of value of other doesn't mean that forcing the first variable to take on a particular value will cause the second to change
- **Causality has a direction, while correlation typically doesn't**
  - Correlation between high income and owning a Ferrari
  - Giving a person a Ferrari doesn't affect their income
  - But increasing their income may make them more likely to buy a Ferrari
- **Confounding variables can cause attributes to be correlated:**
  - High heart rate and sweating are correlated with each other since they tend to both happen during exercise (confounder)
  - Causing somebody to sweat by putting them in sauna won't necessarily raise their heart rate (it does a little, but not as much as exercise)
  - And giving them beta-blockers to lower their heart rate might not prevent sweating (it might a little, but again not like stopping exercising)

correlations might be misleading we shouldn't assign meanings!

Outliers

- **What are outliers?**
  - Data objects that do not comply with the general behavior or model of the data, that is, values that appear as anomalous
  - Most data mining methods consider outliers noise or exceptions.
- **Outliers may be detected using**
  - Manual inspection and knowledge of reasonable values.
  - Statistical tests that assume a distribution or probability model for the data
  - Distance measures where objects that are a substantial distance from any other cluster are considered outliers
  - Deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group

How Do We Manage Outliers?

- Outliers are typically filtered out by eliminating the data points containing them
- **Trimming**
  - Eliminate the outlier data values
- **Winsorizing**
  - A 10% Winsorizing, consider the 5th and 95th percentiles
  - Set the values below the 5th percentile to the 5th percentile itself
  - Set the values above the 95th percentile to the 95th percentile itself
- Note that, in some applications, outliers are the focus on the analysis. Fraud detection is a typical example of this.

Normalization

Normalization

- When attributes have vastly different scales (e.g., age vs income), it is necessary to normalize them
- Range normalization
  - Converts all values to the range [0,1]

$$x'_i = \frac{x_i - \min_i x_i}{\max_i x_i - \min_i x_i}$$

- Standard Score Normalization
  - Forces variables to have mean of 0 and standard deviation of 1
  - So if data was normally distributed, most of it (68%) will lie in range [-1,+1]

$$x'_i = \frac{x_i - \mu}{\sigma}$$

Visualization

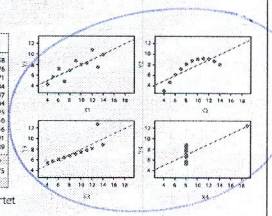
Why visualize data? Anscombe's Quartet

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Why visualize data? Anscombe's Quartet

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

mean, variance and correlation are the same in all the 4 cases

all the summary statistics are the same but the data points are very different

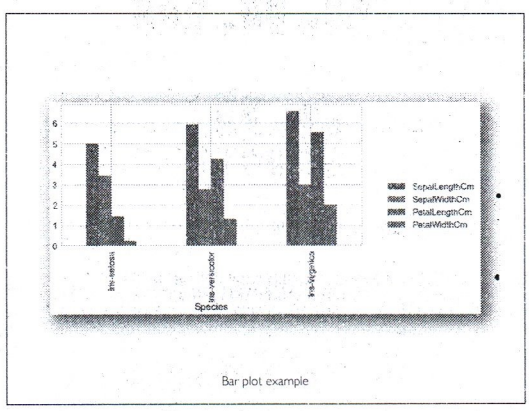
Visualization

- Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported
- Data visualization is one of the most powerful and appealing techniques for data exploration
  - Humans have a well-developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Bar Plots

Bar Plots

- They use horizontal or vertical bars to compare categories.
- One axis shows the compared categories, the other axis represents a discrete value
- Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulate effect (stacked bar graphs)

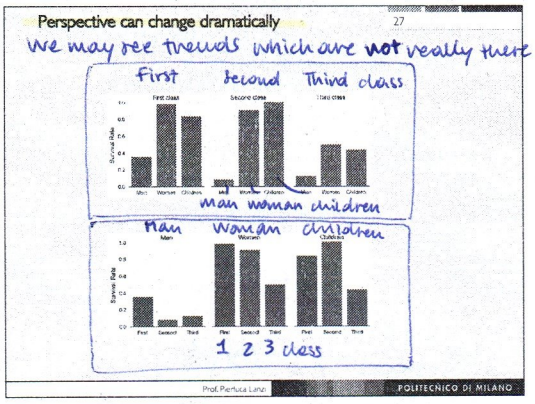


Bar plot example

Perspective can change dramatically

27

We may see trends which are **not** really there!



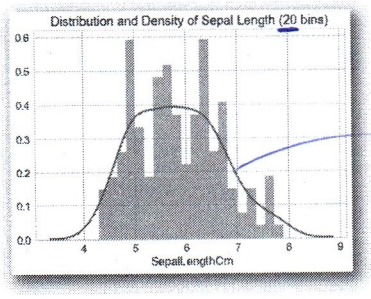
Histograms

Histograms

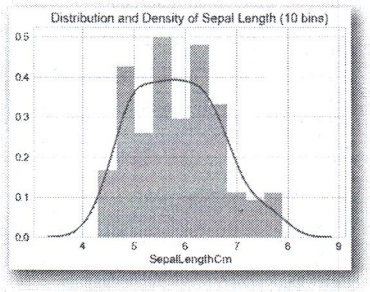
29

We take the range, we split it into intervals and then we count how many data we have in each interval

- They are a graphical representation of the distribution of data introduced by Karl Pearson in 1895
- They estimate the probability distribution of a continuous variables
- They are representations of tabulated frequencies depicted as adjacent rectangles, erected over discrete intervals (bins). Their areas are proportional to the frequency of the observations in the interval.
- The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins

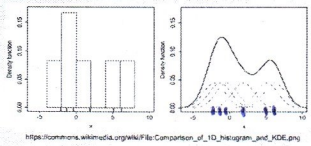


kernel density estimator



Kernel Density Estimates

- What were the lines on the previous plots?
- Smoothed line calculated using kernel-density estimation



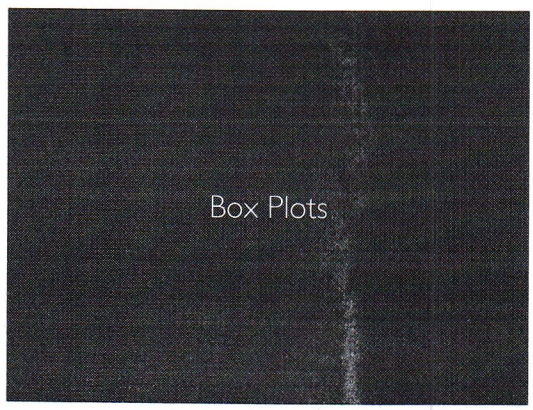
https://commons.wikimedia.org/wiki/File:Comparison_of_1D_histogram_and_KDE.png

- Place small kernel function (e.g. density function of Gaussian distribution) at each observed datapoint
- Curve is simply the summation over the individual

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

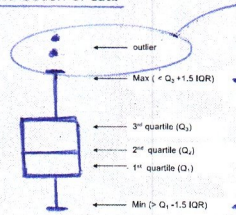
Bandwidth of Kernel Density Estimates

- The bandwidth h of the kernel determines how smooth the estimate is
- Large h generates a smooth function with a possible loss of information
- Small h generates a bumpy function with possible overfitting
- Bandwidth can be determined by cross-validation (hold out some data and estimate the probability)



Box Plots

- Box Plots (invented by Tukey)
- Another way of displaying the distribution of data
- Following figure shows the basic part of a box plot
- IQR is the interquartile range or Q3-Q1

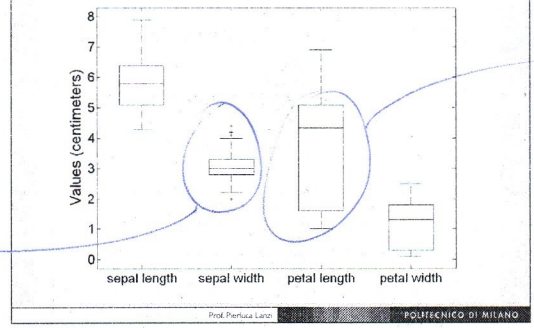


$Q_3 + \frac{3}{2}IQR$

$Q_1 - \frac{3}{2}IQR$

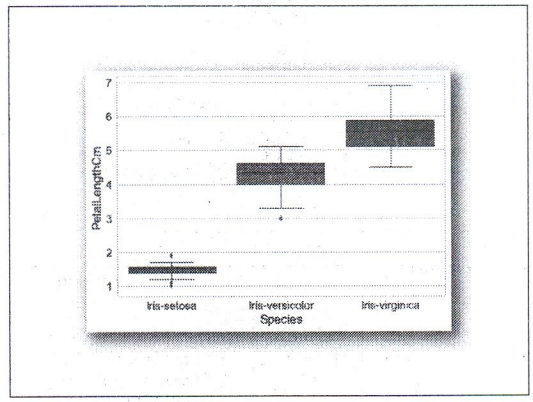
The fact that boxplots name them outliers doesn't mean that they really are. This has nothing to do with outliers detection.

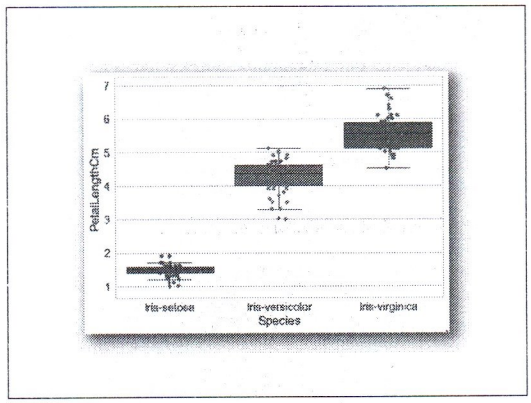
Box Plot Example



Here we can see that the median is almost at the center → symmetric data

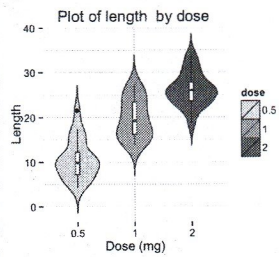
Here instead the median is far from being in the center, so data are not symmetric





Violin Plots

- Uses the kernel density estimate (smoothed histogram) instead of just a box at quartiles

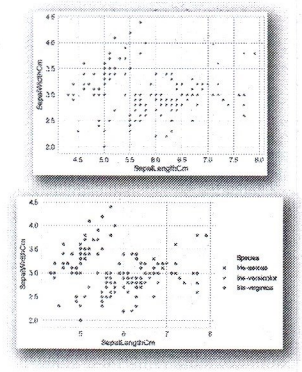


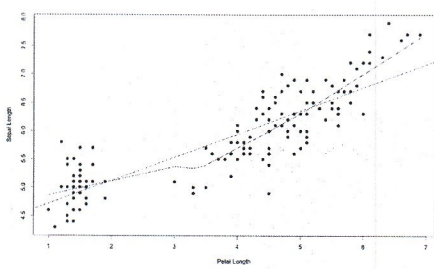
Source: http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization

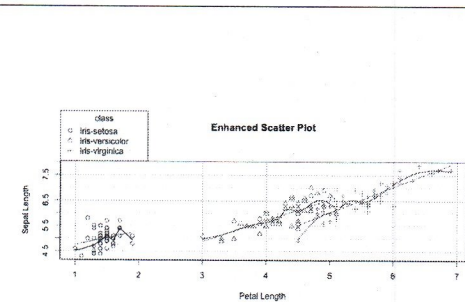
Scatter Plots

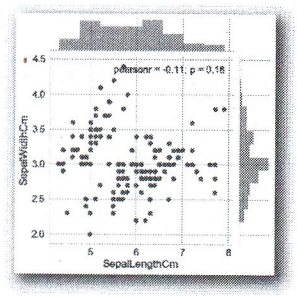
Scatter Plots

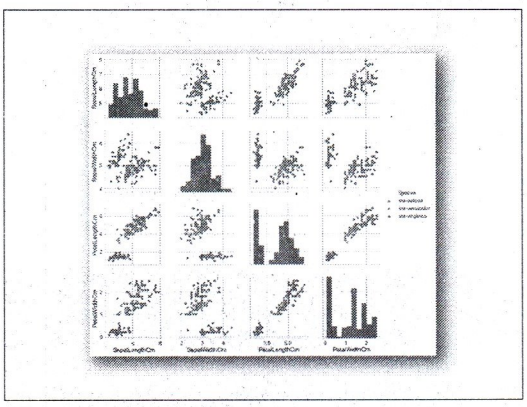
- Used to compare two (or more) attributes
  - Attributes values used to determine the position of the point
  - Two-dimensional scatter plots most common, but three-dimensional plots also used
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
- Examples: http://www.statmethods.net/graphs/scatterplot.html

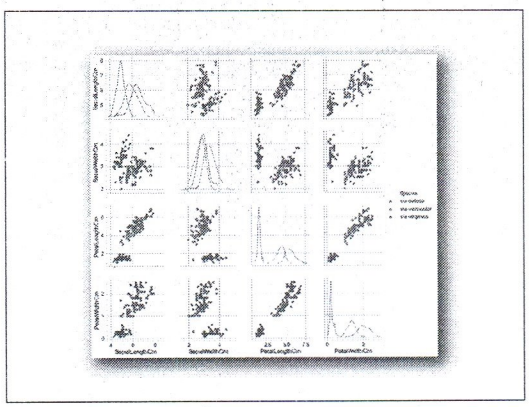












Visualizing Spatial Data

Visualizing Spatial Data

- **Geometry**
  - Geographic
  - Others
- **Scalar fields (one value per cell)**
  - Isocontours
  - Direct volume rendering
- **Vector and tensor fields (many values per cell)**
  - Flow glyphs
  - Geometric (sparse seeds)
  - Textures (dense seeds)
  - Features (globally derived)

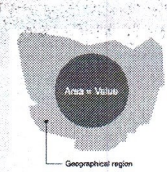
Geometry

Dot Maps

- **What?**
  - Geometry (position)
- **Why?**
  - Locate data in space
- **Remarks**
  - Scale up to hundreds of items
  - Color/shape can encode an additional categorical attribute (reduce scalability)

Bubble Maps

- **What?**
  - Geometry (position)
  - One quantitative attribute
- **Why?**
  - Locate data in space
  - Lookup and compare
- **Remarks**
  - Scale up to hundreds of items
  - Color can encode an additional ca
    (interaction with size).

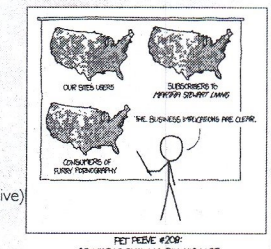


Choropleth Map

- **What?**
  - Geometry (position)
  - One quantitative attribute
- **Why?**
  - Locate data in space
  - Lookup and compare
- **Remarks**
  - Scale up to ~1000 items
  - Hue can encode an additional categorical attribute (better if binary

Bias of Population-related Maps

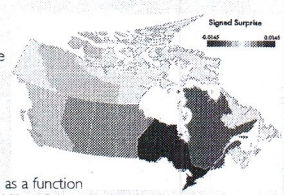
- Using absolute values is dangerous!
- Any map would show population distributionS
- How to deal with this?
  - Visualize per capita (relative)
  - Use statistical models
- See the example at https://xkcd.com/1138



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

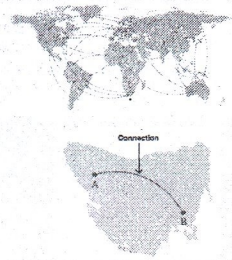
Surprise Maps

- **What?**
  - Geometry (position)
  - One quantitative attribute
  - One derivative attribute
- **Why?**
  - Locate data in space
  - Lookup and compare
- **Remarks**
  - The surprise is computed as a function prior and posterior probability of data distribution.
  - Prior probability is generated with a family of standard models.
- Illustration from https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865



Connection Maps

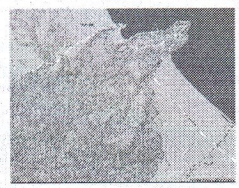
- **What?**
  - Network and positions
- **Why?**
  - Lookup path
  - Identify patterns
- **Remarks**
  - Size of links can encode an additional ordered attribute (3-4 bins at max)



Scalar Fields

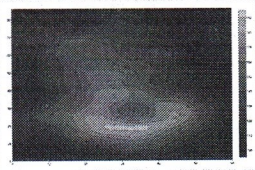
Isocontour Maps

- **What?**
  - Geographic data
  - One quantitative attribute
  - Derived positions
- **Why?**
  - Shape
- **Remarks**
  - The lines are computed from the values of scalar field
  - Area can be filled and color encoded



Isocontour Plots

- **What?**
  - 2D spatial field
  - One quantitative attribute
  - Derived geometry
- **Why?**
  - Shape and patterns
- **Remarks**
  - The lines are computed from the values of scalar field
  - Area can be empty or filled and color encoded



Isosurface Plots

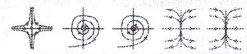
- **What?**
  - 3D spatial scalar field
  - 1 quantitative attribute
  - derived geometry
- **Why?**
  - Shape
- **Remarks**
  - Tree of isosurfaces: positions computed for specific values of the scalar field

Vector Field

position + direction + intensity

Glyph Flow

- **What?**
  - 2D vector fields
- **Why?**
  - Shape and patterns
  - Identify critical points
- **Remarks**
  - Different glyphs can be used to represent vectors
  - Density of grid and jittering



Geometric Flow

- **What?**
  - 2D/3D vector field
  - Derived geometry
- **Why?**
  - Shape and patterns
  - Identify critical points
- **Remarks**
  - Seeding strategy affects the outcome
  - Usage of clustering and color coding improves readability

Texture Flow

- **What?**
  - 2D vector field
- **Why?**
  - Shape and patterns
  - Identify critical points
- **Remarks**
  - Similar to glyphs flow, but computes the flow of a continuous distribution of particles

Feature flow: anatomy

- **What?**
  - 2D vector fields
- **Why?**
  - Shape and patterns
  - Identify critical points
- **Remarks**
  - Similar to glyphs flow
  - But seeding is based on global computing strategy to identify areas with similar behaviors

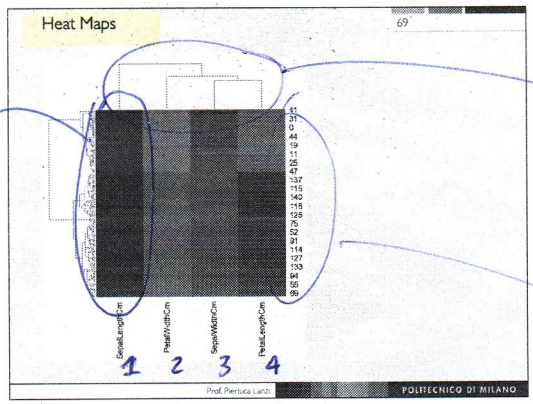
More than Two Dimensions at Once

Two main approaches

Visualize all the dimensions at once
(e.g., heatmaps, spider plots, and Chernoff)

Project the data into a smaller space
and visualize the the projected data

Heat Maps



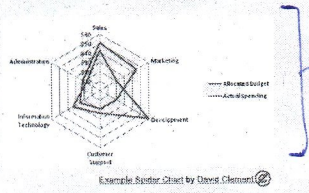
We can see that these values are very different from the others

From here we can obtain a clustering. From this clustering we can see that 3 and 4 are the most similar. Then 2 is the most similar to them. And the least similar is 1.

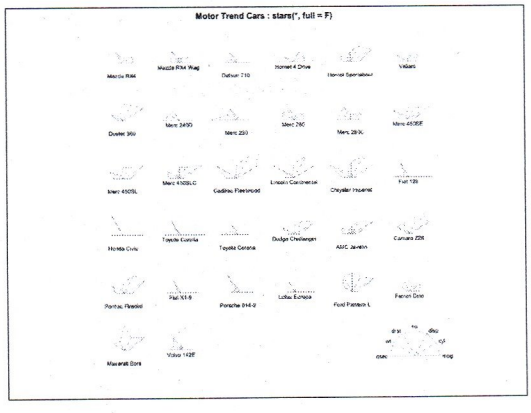
We see that there is a sorting

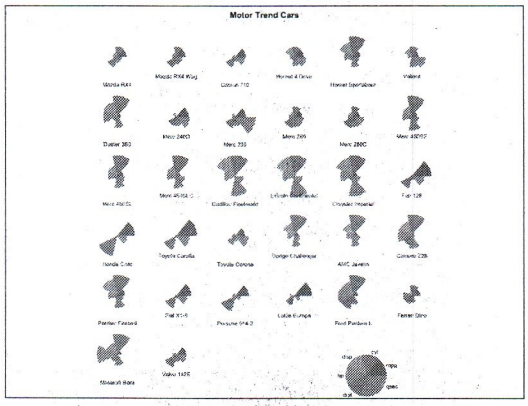
Spider plots, Radar Plots, and Star Plots

- Information radiates outward from central point
- The line connecting the values of an object is a polygon



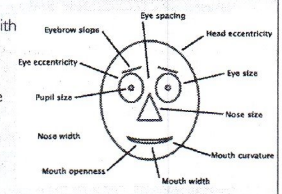
We have a vertix for every variable (feature). If we plot all the datapoints we should be able to detect some patterns.

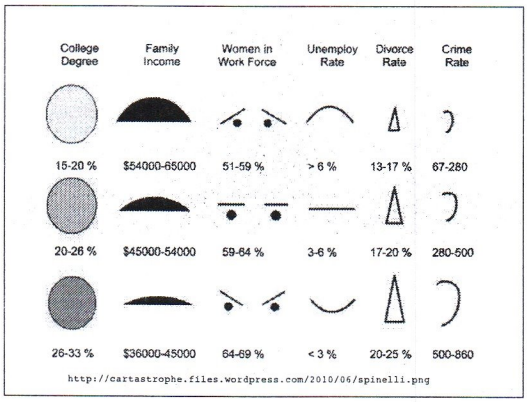


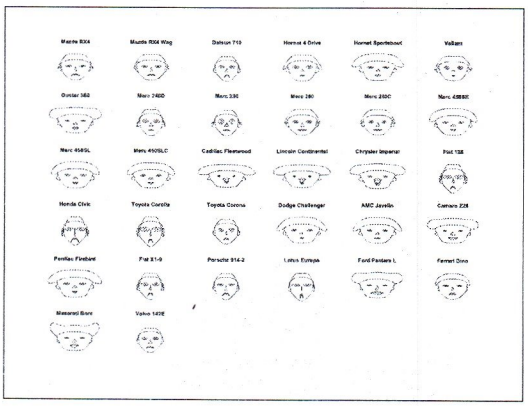


Chernoff Faces

- Approach created by Herman Chernoff
- Associates each attribute with a characteristic of a face
- The values of each attribute determine the appearance of the corresponding facial characteristic
- Each example becomes a separate face
- Relies on human's ability to distinguish faces







Projecting Data into a Lower Dimensional Space

When projecting high-dimensional data into fewer dimensions we can either

1. Find a linear projection
   e.g. use Principle Component Analysis

2. Find a non-linear projection
   e.g. use t-distributed Stochastic Neighbor Embeddings (t-SNE)

1.

Principal Component Analysis

Principal Component Analysis (PCA)

80

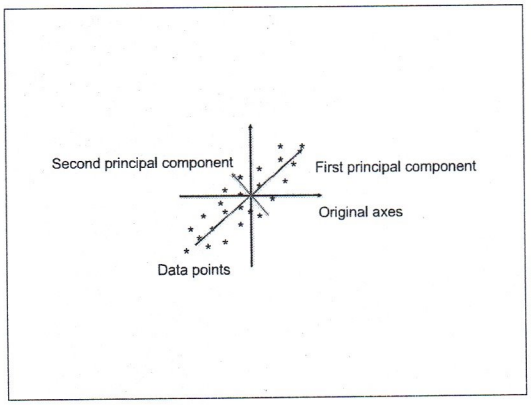
- Typically applied to reduce the number of dimensions of data (feature selection)
- The goal of PCA is to find a projection that captures the largest amount of variation in data
- Given N data vectors from n-dimensions, find k<n orthogonal vectors (the principal components) that can be used to represent data
- Works for numeric data only and it is affected by scale so data usually need to be rescaled before applying PCA

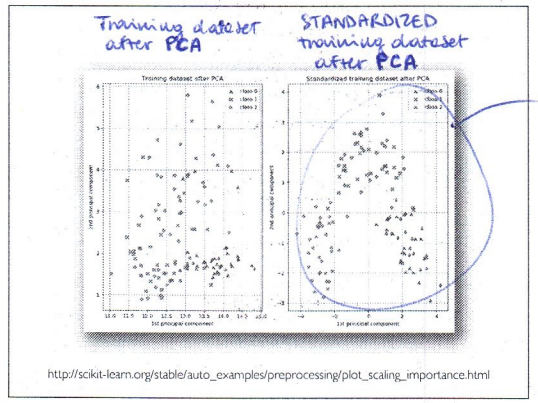
!

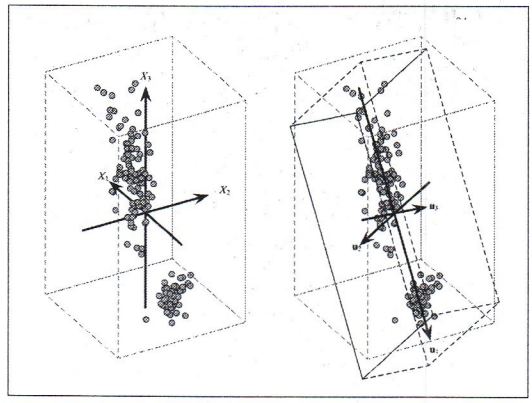
Principal Component Analysis (PCA)

81

- Steps to apply PCA
  - Normalize input data
  - Compute k orthonormal (unit) vectors, i.e., principal components
  - Each input data point can be written as a linear combination of the k principal component vectors
- The principal components are sorted in order of decreasing "significance" or strength
- Data size can be reduced by eliminating the weak components, i.e., those with low variance.
- Using the strongest principal components, it is possible to reconstruct a good approximation of the original data)

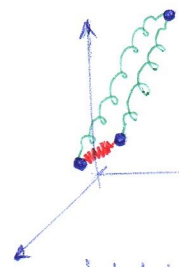






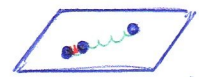
2.

t Distributed Stochastic Neighbor Embedding



Consider a set of points in a multiple-dim space. We put a strong spring ("molla") between two close points and a weak spring between two four points. We do this for all the points combinations.

Once we're done with the springs we let the points fall down on a 2 dimensional plane.



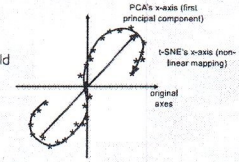
Once the spring has calm down we see that the closest points in heigher dimention are close also in this projection (because of the strong spring)

Note: there is a stochastic component, if we repeat the procedure k times we obtain k different results

Non-linear Dimensionality Reduction

86

- Data in high dimensions never fills the entire space and always lives within some lower-dimensional manifold
- t-SNE is a non-linear dimensionality reduction technique used to map high-dimensional data into 2 or 3 dimensions
- Points from original space mapped onto "map points" in 2D/3D
- Unlike PCA, the mapped points are not linear combination of original attribute values and the axes of mapped space are not linear combination (rotation) of original axes



t Distributed Stochastic Neighbor Embedding

87

- t-SNE tries hard to preserve local distances to nearby points
- Unlike PCA which tries to preserve global (long range) distances between points as much as possible
- t-SNE converts distances between data points to joint probabilities then models orignal points by mapping them to low dimensional map points such that position of map points conserves the structure of the data
- i.e. similar data points are modeled by nearby map points while dissimilar data points are modeled by distant map points

t Distributed Stochastic Neighbor Embedding

- The t-SNE algorithm has two main steps

1. Define a probability distribution over pairs of high-dimensional data points so that:
   - Similar data points have a high probability of being picked
   - Dissimilar points have an extremely small probability of being picked
2. Define a similar distribution over the points in the map space
   - Minimize the Kullback–Leibler divergence between the two distributions with respect to the locations of the map points
   - To minimize the score, it applies gradient descent

Physical Analogy

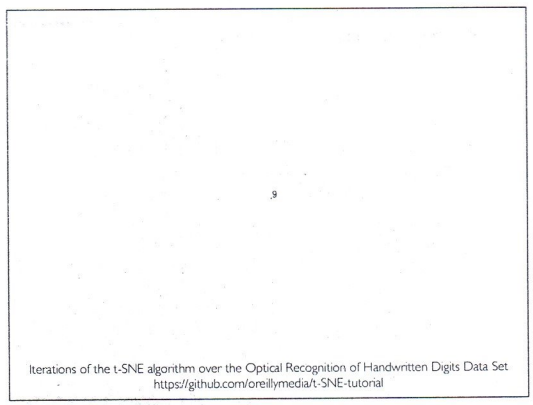
- Assume that map points are all connected with springs.
  - The stiffness of a spring connecting two points depends on the mismatch between the similarity of the two data points and the similarity of the two map points
- Let the system evolve according to the laws of physics
  - If two map points are far apart while the data points are close, they are attracted together
  - If they are nearby while the data points are dissimilar, they are repelled.
- The final mapping is obtained when the equilibrium is reached.

Example: Handwritten Digit Dataset

- Optical Recognition of Handwritten Digits Data Set from the UCI machine learning repository

  http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

- Contains 1797 images with 8x8 pixels each



Iterations of the t-SNE algorithm over the Optical Recognition of Handwritten Digits Data Set
https://github.com/oreillymedia/t-SNE-tutorial

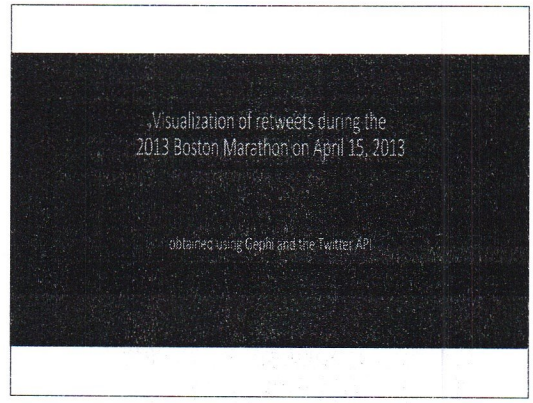
t-SNE Caveats

- The parameter perplexity says (loosely) how to balance attention between local and global aspects of the data
- Different initializations will lead to different results
- Should be applied to data with a "reasonable" number of dimensions (e.g. 30-50)
- If the data have more dimensions, another dimensionality reduction algorithm should be applied

Force-directed Layout

- Idea of mapping complicated data into 2D is not limited to high dimensional data ….
- We can map any graph of data points into 2D provided we have some (dis)similarity value between pairs of nodes
  - Such as the Euclidean distance between them in higher dimensional space
  - Or their joint probability under a Gaussian kernel (in case of t-SNE)
  - Or Pearson's correlation, Spearman's Rank correlation, chi-squared, etc.
- It works by moving points around in the mapped 2D space until convergence
- Technique is called force-directed layout. There are many algorithms implemented in Gephi (https://gephi.org), on next slide we have an example of a graph generated using it

Other Visualizations



Some Interesting Pages

- https://python-graph-gallery.com
- https://python-graph-gallery.com/category/seaborn/
- https://github.com/matplotlib/AnatomyOfMatplotlib
- https://github.com/rasbt/matplotlib-gallery
- https://seaborn.pydata.org/examples/index.html
- http://bokeh.pydata.org/en/latest/

Study Material

- **"Data Mining and Analysis" – Chapter 2 & 3**
- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**
  - http://lvdmaaten.github.io/tsne/
  - https://www.youtube.com/watch?v=RJVL80Gg3lA
  - http://alexanderfabisch.github.io/t-sne-in-scikit-learn.html
  - http://jmlr.csail.mit.edu/papers/volume9/vandermaaten08a/vandermaaten08a.pdf
  - https://distill.pub/2016/misread-tsne/
- **Principal Component Analysis**
  - "Data Mining and Analysis" – Chapter 7
  - http://sebastianraschka.com/Articles/2015_pca_in_3_steps.html
- **Data Visualization**
  - PhD Course by Daniele Loiacono