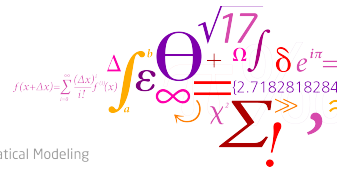
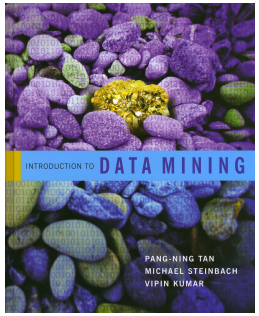
02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 1.1-1.4



Lecture schedule

1. **Introduction** **(Tan 1.1-1.4)**

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.2 +(A)+ B.1)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

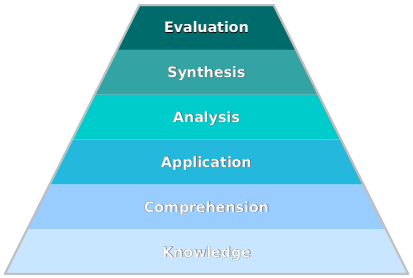
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3 + 8.5.7)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

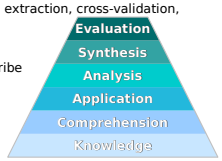
12. Putting it all together: Summary and overview
13. Mini project presentation

Blooms taxonomy



Learning objectives

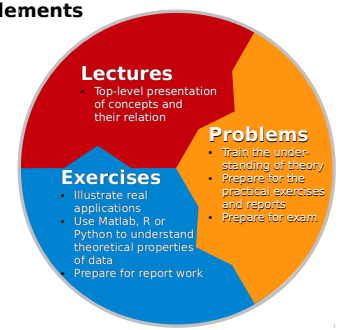
1. Describe the major steps involved in data modeling from preparing the data, modeling the data to evaluating and disseminating the results. (Knowledge)
2. Discuss key machine learning concepts such as feature extraction, cross-validation, generalization and over-fitting, prediction and curse of dimensionality. (Comprehension)
3. Sketch how the data modeling methods work and describe their assumptions and limitations. (Knowledge and Comprehension)
4. Match practical problems to standard data modeling problems such as regression, classification, density estimation, clustering and association mining. (Comprehension and Application)
5. Apply the data modeling framework to a broad range of application domains in medical engineering, bio-informatics, chemistry, electrical engineering and computer science. (Application)
6. Compute the results of the data modeling framework by use of Matlab, R or Python. (Application)
7. Use visualization techniques and statistics to evaluate model performance, identify patterns and data issues. (Analysis)
8. Combine and modify data modeling tools in order to analyze a data set of their own and disseminate the results of the analysis. (Application, Analysis, Synthesis and Evaluation)



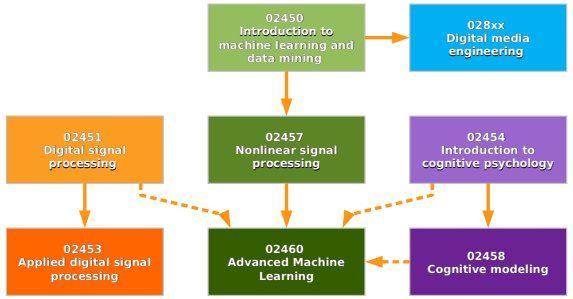
Examination

- 4 hours written examination (multiple choice)
  Concentrated on the same topics as the lectures and exercises and linked to the learning objectives.
- Group reports
  Report Deadlines:
  - Report 1: 30th September, **Feature extraction and visualization**
  - Report 2: 4th November, **Supervised learning: Classification and regression**
  - Report 3: 2nd December, **Unsupervised learning: Clustering and density est.**
- Final grade based on an overall assessment of the reports and written exam

Course elements



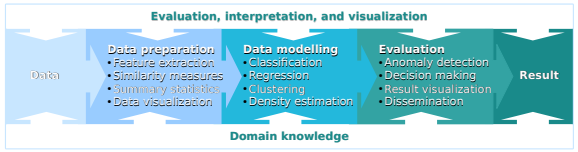
Relevant courses at Section for Cognitive Systems



Pretest

The purpose of the pretest is to assess the students background and academic level in order to allow the teachers to adjust the presentation of the course material as well as measure student learning. This pretest will not be graded and will not influence exam results in any way. We do not expect you to be able to answer all questions.

Data modeling framework



We are entering the era of big data

- Google index 30 trillion webpages
(http://venturebeat.com/2013/03/01/how-google-searches-30-trillion-web-pages-100-billion-times-a-r
- 100 hours of video are uploaded to youtube every minute resulting in 16 years of content every day
(source: youtube)
- We have sequenced more than 1000 peoples genome of $3.8 \cdot 10^9$ base pairsz
(source: K. P. Murphy "Machine Learning")
- Walmart handles more than 1 mio. transactions per hour and has databases containing more than $2.5 \cdot 10^{15}$ bytes of information
(source: K. P. Murphy "Machine Learning")
- Each night the worlds astronomy laboratories store high-resolution of the night sky of around a terabyte (10^{12})
(source: Stephen Marsland "Machine Learning An Algorithmic Perspective")
- In total, the four main detectors at the Large Hadron Collider (LHC) produced 13 petabytes (10^{15}) of data in 2010
(source: wikipedia "Big Data")
- Facebook handles 50 billion photos from its user base.
(source: wikipedia "Big Data")
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide
(source: wikipedia "Big Data")

Every day, we create 2.5 quintillion (10^{18}) bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.

Source: http://www-01.ibm.com/software/data/bigdata/

"We are drowning in information and starving for knowledge"

John Naisbitt, author of "Megatrends"

What is machine learning?

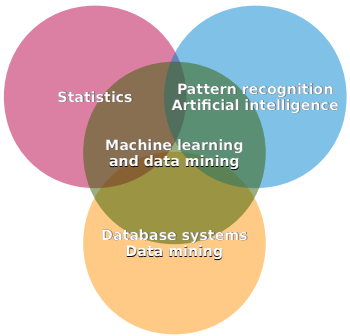
- Arthur Samuel (1959)
  - ***Machine learning:* "Field of study that gives computers the ability to learn without being explicitly programmed"**
    - Samuels wrote a checkers playing program
      - Had the program play 10000 games against itself
      - Work out which board positions were good and bad depending on wins/losses
- Tom Michell (1999)
  - ***Well posed learning problem:* "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."**
    - The checkers example,
      - E = 10000 games
      - T = playing checkers
      - P = if you win or not


Arthur Samuel (1901-1990)


Tom Michell

Source: http://www.holehouse.org/mlclass/08_Neural_Networks_Representation.html

Machine learning and data mining



But machines cannot really think

Human abilities examined

It is true machines are not very smart. However for many tasks humans are even worse.

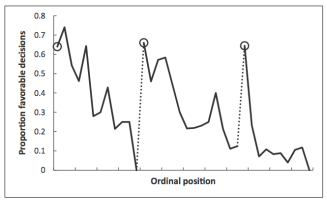
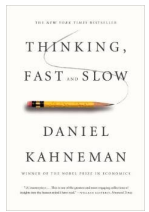


Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on *x* axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

(Source: Danziger, S., J. Levav, and L. Avnaim-Pesso, 2011:Extraneous factors in judicial decisions. Proc. Nat. Acad. Sci.)

Human abilities examined

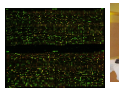


The number of studies reporting comparisons of clinical and statistical predictions has increased to roughly two hundred, but the score in the contest between algorithms and humans has not changed. **About 60% of the studies have shown significantly better accuracy for the algorithms. The other comparisons scored a draw in accuracy**, but a tie is tantamount to a win for the statistical rules, which are normally much less expensive to use than expert judgement. No exception has been convincingly documented.

The range of predicted outcomes has expanded to cover medical variables such as longevity of cancer patients, the length of hospital stayes, the diagnosis of cardiac disease, and the susceptibility of babies to sudden infant death syndrome; economic measures such as the prospect of success for new businesses, the evaluation of credit risks by banks, and the future career satisfaction of workers; questions of interest to government agencies, including assessment of the suitability of foster parents, the odds of recidivism among juvenile offenders, and the likelihood of other forms of violent behaviour; and the miscallaneous outcomes such as the evaluation of scientific presentations, the winners of football games, and the future prices of Bordeaux wine. Each of these domains entails a significant degree of uncertainty and unpredictability. We describe them as "low-validity environments.". **In every case, the accuracy of experts was matched or exceeded by a simple algorithm.** (Kahneman 2011)

Applications

- **Chemistry**
  - Spectrometry
  - Chemical sensors
- **Audio processing**
  - Spoken digit classification
  - Music genre classification
- **Image processing**
  - Hand-written digit recognition
  - Image tagging and classification
  - Number plate recognition
- **Informatics**
  - Collaborative filtering
  - Text corpus analysis
  - Spam filters
  - Computer games
- **Biomedical**
  - Micro-array gene analysis
  - Medical Imaging
- **Financial data mining**
  - Market predictions
- **Climate data**
  - Weather forecast



Current trend in machine learning

Simple Machine-Learning methods

+ Lots of data

+ Fast computers

= Solution to many "hard" problems

Advanced applications of machine learning

Armies of Expensive Lawyers, Replaced by Cheaper Software (*NYTimes, 2011*)

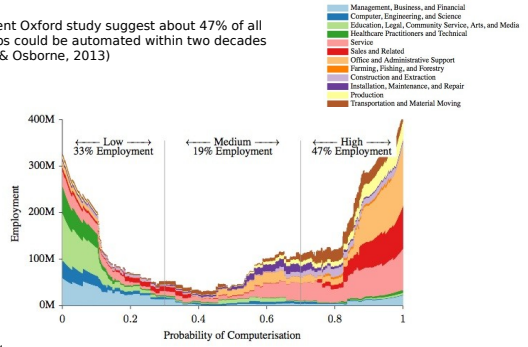
"a lot of people who used to be allocated to conduct document review are no longer able to be billed out," said Bill Herr, who as a lawyer at a major chemical company used to muster auditoriums of lawyers to read documents for weeks on end.

In August 2012, the [Google autonomous car] team announced that they have completed over 300,000 autonomous-driving miles (500,000 km) accident-free ... Four U.S. states have passed laws permitting autonomous cars as of December 2013 (Wikipedia)

In 2011, Watson competed on Jeopardy! against former winners Brad Rutter and Ken Jennings. Watson received the first place prize of $1 million (Wikipedia)

Machine learning as a disruptive technology

- A recent Oxford study suggest about 47% of all US jobs could be automated within two decades (Frey & Osborne, 2013)

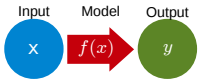


Data Mining and Machine Learning Tasks

Predictive tasks (Supervised learning)

– Use some variables to predict unknown or future values of other variables

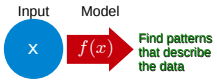
- **Classification**
  - Discrete output
  (Determine which class a new data object belongs to)
- **Regression**
  - Continuous output
  (Determine the output value from the input variables)



Descriptive tasks (Unsupervised learning)

– Find human-interpretable patterns that describe the data

- **Clustering**
  - Discover group structure in data
- **Association rule discovery**
  - Discover how data objects relate to each other
- **Anomaly detection**
  - Find data objects that are abnormal

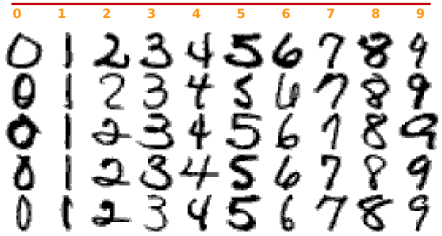


Classification: Definition

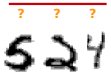
- Given a collection of data objects (**training set**)
  - Each object has associated a number of features
  - Each object belongs to a certain class
- Define a **model** for the class given the other features
- Goal: Assign a class label to a **previously unseen object**

Classification: Example

Training set

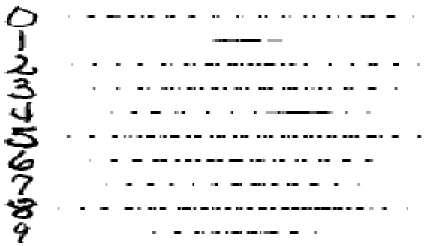


Classify



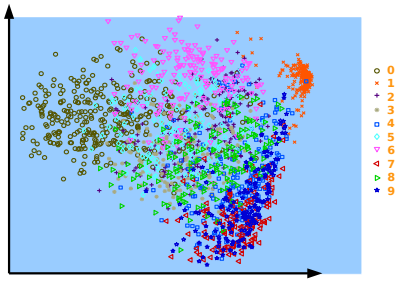
Classification: Example

Data representation

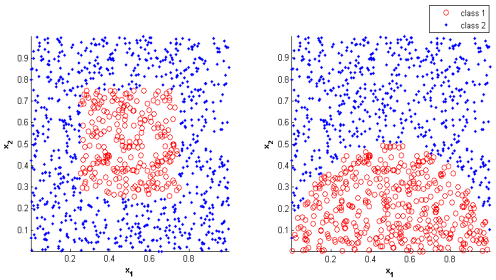


Classification: Example

Visualization

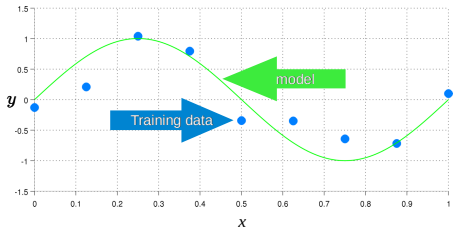


How would you characterize the two classes and how would you determine what class a new observation belongs to?



Regression: Definition

- Given a collection of data objects
  - Each object has associated a number of features
  - Each object has associated a **continuous valued variable**
- Define a **model** for the variable given the features
- Goal: Predict the value of the variable for a **previously unseen object**



Regression: Example

- Predict **sales amounts** of new product based on
  - advertising expenditure
- Predict **wind velocity** as a function of
  - temperature, humidity, and air pressure
- Predict the value of a **stock market index** based on
  - previous index time series and market indicators

Clustering: Definition

- Given a collection of data objects
  - Each object has associated a number of features
  - A measure of **similarity** between objects is defined
- Goal: **Group the objects** into clusters such that
  - Objects within each cluster are similar
  - Objects in separate clusters are less similar

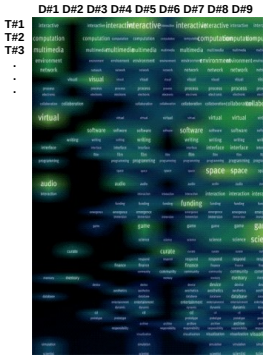
Clustering: Definition

- Given a collection of data objects
  - Each object has associated a number of features
  - A measure of **similarity** between objects is defined
- Goal: **Group the objects** into clusters such that
  - Objects within each cluster are similar
  - Objects in separate clusters are less similar

Clustering: Example

Document clustering

- **Goal**
  - Find groups of similar documents based on the words appearing in them
- **Approach**
  - Identify frequently occurring words in each document
  - Define a similarity measure based on the word frequencies
  - Perform clustering to find groups of documents
- **Motivation**
  - Use the clusters to relate a new document to existing documents
  - Better search algorithms: Return documents that are similar but do not have the exact search keywords



Association rule discovery: Definition

- Given a set of **records**
  - Each containing a number of **items from a set**
- Goal: Produce dependency rules
  - Predict the occurence of an item based on occurences of other items

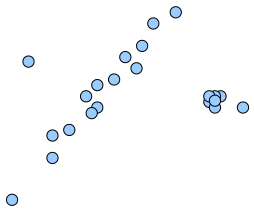
Association rule discovery: Example

Market basket analysis

Training set	Rules discovered
1. {Bread, Coke, Milk}	{Milk} ► {Coke}
2. {Beer, Bread}	{Diaper, Milk} ► {Beer}
3. {Beer, Coke, Diaper, Milk}	
4. {Beer, Bread, Diaper, Milk}	
5. {Coke, Milk}	

Anomaly detection: Definition

- Given a collection of data objects
  - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour



Anomaly detection: Example

- Credit card **fraud detection**
  - Recognize dubious credit card transactions based on the transaction history of the card holder
- Detection of **outliers** in data measurements
  - Remove erroneous measurements due to misreading from an instrument
- **Fault detection** in system health monitoring
  - Detect when a wind turbine performs poorly due to ice coating on blades

Discussion

Which of these activities are machine learning / data modeling tasks?
(Tan Ex. 1.1)

a) Dividing the customers of a company according to their gender
b) Dividing the customers of a company according to their profitability
c) Computing the total sales of a company
d) Sorting a student database based on student identification number
e) Predicting the outcomes of tossing a (fair) pair of dice
f) Predicting the future stock price of a company using historical data
g) Monitoring the heart rate of a patient for abnormalities
h) Monitoring seismic waves for earthquake activities
i) Extracting the frequencies of a sound wave

Group discussion

You all have different backgrounds and experiences

- Give an example of a data modeling problem from
  - your previous work or
  - your interests or
  - your imagination
- Which of the five modeling tasks could be applied to solve the problem?

Data modeling tasks

- **Classification**
  - Determine which class a new data object belongs to
- **Regression**
  - Learn a functional relationship from input to output
- **Clustering**
  - Discover group structure in data
- **Association rule discovery**
  - Discover how data objects relate to each other
- **Anomaly detection**
  - Find data objects that are abnormal

Exercises

All the exercises will be in Matlab, Python and R.

–Two options for running Matlab on your computer:
1)Install Matlab on your computer and run it using internet connection to a license server: http://www.gbar.dtu.dk/downloads/#
2) Run Matlab on the GBAR from your computer using thinlinc: http://gbar.dtu.dk/ThinLinc

–Python and R is freely available
(exercise 1 today will guide you through how to install the programs)

You should form groups of 2-3 people for the exercises and for the 3 (group) reports.

Each group will give feedback to the teachers on the lectures and exercises of one of the course weeks.

Lecture schedule

Report 1
1. **Introduction (Tan 1.1-1.4)**

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.2 +(A)+ B.1)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

Report 2
5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

Report 3
9. K-means and hierarchical clustering (Tan 8.1-8.3 + 8.5.7)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project presentation

See also the course homepage for the lecture outline and reading material for each week: www.imm.dtu.dk/courses/02450


Section for Cognitive Systems
DTU Compute

02450 Introduction to Machine Learning and Data Mining

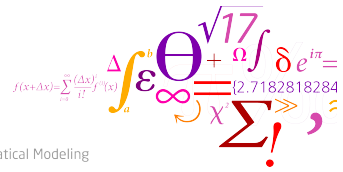
Machine learning and data mining

Lecture schedule

Today's exercise:

- **Form groups for project work**
  For the projects 1, 2 and 3 you must work in groups of 2-3 students. The sooner you form the groups the better. Once you have formed groups please send me en email at tuhe@dtu.dk with your information in the format:
  `Alice Hansen (s113589) aliceh@gmail.com`
  `Bob Carlsen (s095458) bobc@yahoo.com`
  `Charlie Jensen (s125199) charliej@mail.com`
  - **Please send the mail before 9th September**
- **Find a dataset to analyze throughout the course:**
  Each group will find a dataset of their own. Either your own dataset, a dataset you find yourself or for instance taken from one of the resources given in the guideline *"FindingADatasetForTheReport.pdf"* on Campusnet. The 3 group reports will be based on the dataset that in turn will be analyzed by the various approaches taught during the course. Once you have found a dataset you need to have the dataset approved by me for the course.
  - **Deadline for finding and having the dataset approved: 16th September**
  - **First report must be uploaded to campusnet 30th September. Notice file upload is closed after 13:00.**
- **Familiarize yourself with Matlab, Python or R:**
  Todays exercise is a brush up course on either Matlab, Python or R targeted for those not familiar with these programming language. We recommend you use Matlab unless you are much more familiar with Python or R.
- **Exercises are in building 358 rooms 006, 029, 031 and 032.**
  Instructor with expertise in Python in rooms 006 & 029 and instructor with expertice in R in 029. All instructors have expertice in Matlab.

02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

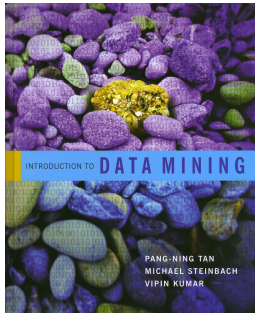
Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 2.1-2.3 + B1 (+ A)

Feedback Groups of the day:
Rasmus Olsen
Oliver Naaby

Jesper Baltzersen
Steen Barkholt
Mads Slotsbo

Signe Bendsen Brusgaard
Amalie Sofie Ekstrand

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. **Data and feature extraction (Tan 2.1-2.3 + (A) + B.1)**
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

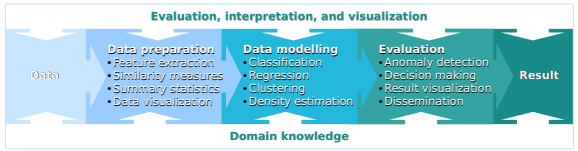
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



Todays learning objectives:
Understand types of data, their attributes and data issues.
Be able to apply principal component analysis for data visualization and feature extraction.

What is data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Also known as variable, field, characteristic, or feature
- Collection of attributes describe an object
  - Also known as record, point, case, sample, entity, or instance

Attributes

	ID	Age	Gender	Name
Data objects	1	31	F	Alex
	2	24	M	Ben
	3	52	F	Cindy
	4	35	M	Dan
	5	58	M	Eric
	6	46	F	Fay
	7	42	M	George

Discrete / continuous attributes

- **Discrete**
  - Finite (or countably infinite) set of values
  - Examples:
    - Zip codes
    - Counts
    - Set of words in a collection of documents
  - Often represented as integer variables
- **Continuous**
  - Has real numbers as attribute values
  - Examples:
    - Temperature
    - Height
    - Weight.
  - Often represented as floating point variables

Types of attributes

- **Nominal:** Objects belong to a category (Equal / Not equal)
  - ID numbers
  - Eye color
  - Zip codes
- **Ordinal:** Objects can be ranked (Greater than / Less than)
  - Taste of potato chips on a scale from 1-10
  - Grades
  - Height in {short, medium, tall}
- **Interval:** Distance between objects can be measured (Addition / Subtraction)
  - Calendar dates
  - Temperature in Fahrenheit and Celcius
- **Ratio:** Zero means absence of what is measured (Multiplication / Division)
  - Length
  - Time
  - Counts
  - Temperature in Kelvin

Qualitative
Quantitative

Discussion

- **Classify the following attributes**
  a) Military rank
  b) Angles measured in degrees
  c) A persons year of birth
  d) A persons age in years
  e) Coat check number
  f) Distance from center of campus
  g) Number of patients in a hospital
  h) Sea level

- **Discrete**
  - Finite (or countably infinite) set of values
- **Continuous**
  - Real number

- **Nominal** (Equal / Not equal)
  - Objects belong to a category
- **Ordinal** (Greater than / Less than)
  - Objects can be ranked
- **Interval** (Addition / Subtraction)
  - Distance between objects can be measured
- **Ratio** (Multiplication / Division)
  - Zero means absence of what is measured

Types of data sets

- **Record data**
  - Collection of data objects and their attributes
  - Representation: Table
- **Relational data**
  - Collection of data objects and their relation
  - Representation: Graph
- **Ordered data**
  - Ordered collection of data objects
  - Representation: Sequence

Record data example: Market basket data

• **Transaction data table**

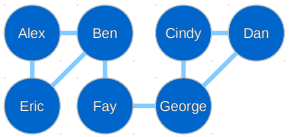
ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

• **Matrix**

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Relational data example: Who knows who?

• **Graph**

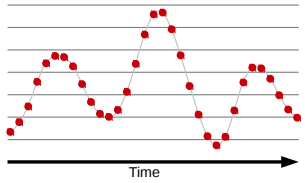


• **Matrix**

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	0	0	1	1	0
C	0	0	0	1	0	0	1
D	0	0	1	0	0	0	1
E	1	1	0	0	0	0	0
F	0	1	0	0	0	0	1
G	0	0	1	1	0	1	0

Ordered data example: Time series

• **Sequence**



• **Matrix**

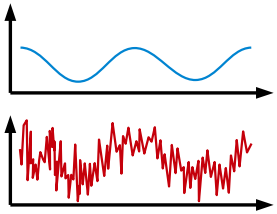
Time	Value
0	1.3
1	1.8
2	2.5
3	3.6
4	4.4
5	4.7
6	4.6
7	4.3
8	2.4
9	2.1
10	2.0
11	2.3
12	3.1

Data quality

- **Data is of high quality if they**
  - Are fit for their intended use
  - Correctly represent the phenomena they correspond to
- **Examples of quality problems**
  - Noise
  - Outliers
  - Missing values

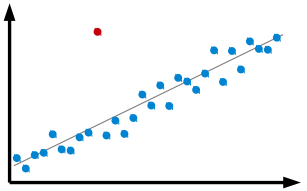
Noise

- **Definition**
  - Unwanted pertubation to a signal
  - Unwanted data
- **Reasons for noise**
  - Fundamental limits in measurement accuracy
  - Interference from other signals
  - Measurement of attributes not related to the data modeling task
- **Handling noise**
  - Exclude noisy attributes
  - Remove noise by filtering
  - Include a model of the noise



Outliers

- **Definition**
  - Data objects which are significantly different from most others
- **Reasons for outliers**
  - Measurement error
  - Natural property of data
- **Handling outliers**
  - Identify & exclude outliers
  - Model the outliers



Missing values

- **Definition**
  - No value is stored for an attribute in a data object
- **Reasons for missing values**
  - Information is not collected
    - People decline to give their age
  - Attribute is not applicable
    - Annual income is not applicable to children
- **Handling missing values**
  - Eliminate data objects
  - Estimate missing values (e.g. an average)
  - Ignore the missing value in analysis

ID	Age	Gender	Name
1	31	F	Alex
2	(?)	M	Ben
3	52	F	Cindy
4	35	(?)	Dan
5	(?)	M	Eric
6	(?)	F	Fay
7	42	M	(?)

Discussion

- A group of people were asked to write how many children they have
  - Their response was this

- A research assistant typed the results into a table
  - His table looked like this

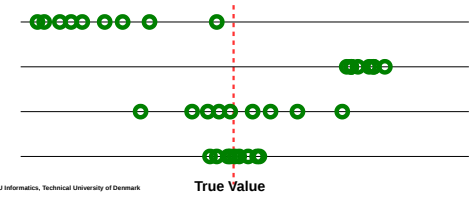
Children	3	1	0	2	7	5	15	0	1	3	-2	0	0	0	1

- Are there any data quality issues?
  - Noise?
  - Outliers?
  - Missing values?
- Why have these issues occured, and how should they be handled?

Precision, Bias and Accuracy

Assume we make repeated measurements of the same underlying quantity and use this set of values to calculate a mean value (average) that serves as our estimate of the true value.

- **Definition 2.3 (Precision):** The closeness of repeated measurements (of the same quantity) to one another (often measured by standard deviation)
- **Definition 2.4 (Bias):** A systematic variation of measurements from the quantity being measured.
- **Definition 2.5 (Accuracy):** The closeness of measurements to the true value of the quantity being measured.



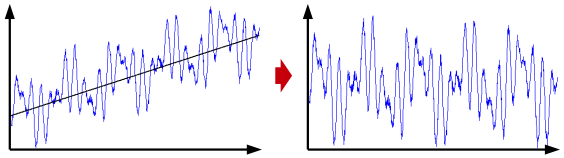
Data preprocessing and dimensionality reduction

- **Aggregation**
  - Combining several attributes into a single attribute
- **Sampling**
  - Selecting a representative subset of data points
- **Dimensionality reduction**
  - Project data to a low-dimensional subspace **(explained today)**
- **Feature subset selection**
  - Choose a subset of attributes
- **Feature extraction**
  - Create new features from existing attributes
- **Discretization and binarization**
  - Reduce continuous attributes to discrete
- **Attribute transformation**
  - Apply a fixed transformation to an attribute

Filtering

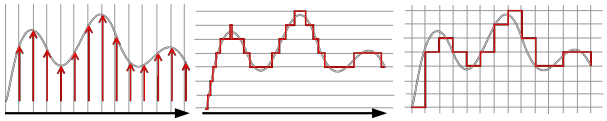
- Eliminating, suppressing, or attenuating certain aspects of the data
  - Noise removal in audio signals
  - Elimination of common words in text documents
  - Removal of background in images
  - Removal of examples which are corrupted
  - De-trending data (if it is not stationary)

Example of de-trending data



From Analog to Digital

- Data are often **analog** and needs to be converted to a **digital** format



Time
Discrete time, Continuous values.

Time
Discrete values, Continuous time

Time
Discrete values, Discrete time

- Some data are "born" digital
  - On/off sensors
  - Questionnaires
  - User ratings

Chemical sensor data

- Example: The nano-nose

	A	B	C	D	E	F	G	H	I	J	K
1	Nanonose		A		B		C		D		
2	Sample type	Concentration									
3	Water	9200	95,5	21	6	11,94231	64,13462	21,49856	5,56784	1,174135	
4	Water	9200	94,5	17	5	5,484615	63,20577	19,65856	4,968	1,083444	
5	Water	9200	92	16	3	11,01769	62,38654	19,81312	5,19346	0,944655	
6	Water	4600	59	7,5	2,5	3,598462	35,16346	6,876207	1,441724	0,146654	
7	Water	4600	51	7,5	2,5	4,865385	34,05769	6,757241	1,612968	0,055663	
8	Water	4600	50	8	2,5	3,980769	33,61538	6,773103	1,778552	0,075009	
9	Water	2300	27,5	4	1,5	2,2	18,35577	2,798333	0,3635	0,030383	
10	Water	2300	27,5	4,5	1,5	2,2	17,25	2,629667	0,854833	0	
11	Water	2300	27	4	1,5	3,1	17,47115	2,645	0,624833	0	
12	Water	1150	13,5	1,5	0	1,705231	9,730769	1,114065	0,495938	0	
13	Water	1150	13	2	0,5	2,653846	10,61538	1,078125	0,452813	0	
14	Water	1150	13,25	2,5	0,25	2,653846	8,846154	0,626563	0,115	0	
15	Water	575	4,5	0,8	0	0,442308	2,432692	0,651667	0,529	0	
16	Water	575	4	0,8	0	0,884615	5,871154	0,406533	0,348333	0	
17	Water	575	5,7	0,5	0	1,320923	6,759615	0,329667	0,2593	0	
18	Water	290	1,5	0	0	0	1,167308	0,133256	0,144	0	

Bag of words

- First three sentences on **wikipedia.org**
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification

Bag of words

- First three sentences on **wikipedia.org**
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification
- We will treat **this text** as a data set and create a bag-of-words model of it

Bag of words

- Elimination of common words (so-called stop words)
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification

bag-of-words model methods document classification

Bag of words

- Stemming

Word	Sentence		
	1	2	3
bag-of-word*	1		1
model*	1	1	1
simplif*	1		
assum*	1		
natural*	1		
languag*	1		
process*	1		
information*	1		
retriev*	1		
text*		1	
sentence*		1	
document*		1	1
represent*		1	
unorder*		1	
collect*		1	
word*		2	
disregard*		1	
grammar*		1	
order*		1	
method*			1
classif*			1

Image representation

- Example: Handwritten digits
- Preprocessing
  - Digitalization
  - Centering
  - Rotation
  - Scaling

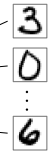
$$M_0 = \begin{bmatrix} 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0.3 & 1 & 0.2 & 0 & \dots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

- Vectorization

$$x_0 = \begin{bmatrix} 0 & \cdots & 0 & 0.3 & 1 & 0.2 & 0 & \cdots & 0 \end{bmatrix}^\top$$

- Matrix representation of data set

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix}$$

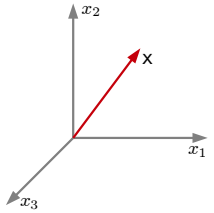


If each image is 16×16 pixels then $\boldsymbol{X}$ is a $N \times 256$ matrix.

Vector space representation

- All these data objects have a vector space representation

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}$$



Vectors and matrices

- Data represented as vectors and matrices
  - Linear algebra useful for manipulating and analyzing data
- We will derive the Principal Component Analysis (PCA) and discuss the Singular Value Decomposition (SVD)
  - First a (brief) highlight of linear algebra
  - PCA is very important for data visualization

Vectors and matrices

- Common matrix notation

$$\boldsymbol{A}, A, \overline{\overline{A}} \qquad \boldsymbol{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,M} \\ \vdots & & \vdots \\ a_{N,1} & \cdots & a_{N,M} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

- Common vector notation

$$\boldsymbol{x}, x, \overline{x}, \vec{x} \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} \in \mathbb{R}^N$$

Matrix Multiplication

- Two matrices can be multiplied $\boldsymbol{AB} = \boldsymbol{C}$
  - if the number of columns in the first equals the number of rows in the second

$$\underset{L \times M}{\boldsymbol{A}} \times \underset{M \times N}{\boldsymbol{B}} = \underset{L \times N}{\boldsymbol{C}}$$

$$\underset{3\times4 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 2 & 3 & 4 \end{bmatrix}} \underset{4\times5 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & a & \cdot \\ \cdot & \cdot & \cdot & b & \cdot \\ \cdot & \cdot & \cdot & c & \cdot \\ \cdot & \cdot & \cdot & d & \cdot \end{bmatrix}} = \underset{3\times5 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & x_{3,4} & \cdot \end{bmatrix}}$$

$$x_{3,4} = 1 \cdot a + 2 \cdot b + 3 \cdot c + 4 \cdot d$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = ?$$

Matrix transpose

- The transpose of a matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad \boldsymbol{A}^\top = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 4 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

- Transpose of a sum

$$(\boldsymbol{A} + \boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$$

- Transpose of a product

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$$

$$(\boldsymbol{Ax})^\top \boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{y} = \boldsymbol{x}^\top (\boldsymbol{A}^\top \boldsymbol{y})$$

The identity matrix

- Ones on the diagonal and zeros everywhere else

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad \boldsymbol{I}^\top = \boldsymbol{I}$$

- Multiplying by the identity does not change anything

$$\boldsymbol{IA} = \boldsymbol{A}$$

$$\boldsymbol{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\boldsymbol{I}_2 \boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Matrix inverses

- For a square matrix, the inverse satisfies

$$\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

- Inverse of a product of square matrices

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$$

- Transpose of inverse

$$(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$$

Norms

- The norm of a vector is usually written as

$$\|\boldsymbol{x}\|_p = \left(\sum_{n=1}^{N} |x_n|^p \right)^{1/p}$$

  - Of particular interest is the 1-norm and the 2-norm

$$\|\boldsymbol{x}\|_1 = \sum_{n=1}^{N} |x_n| \qquad \|\boldsymbol{x}\|_2 = \sqrt{\sum_{n=1}^{N} |x_n|^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$$

- The Frobenius norm of a matrix

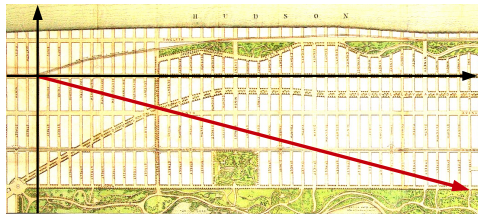
$$\|\boldsymbol{X}\|_F^2 = \sum_{i,j} x_{i,j}^2 = \text{trace}(\boldsymbol{X}\boldsymbol{X}^T) = \text{trace}(\boldsymbol{X}^T\boldsymbol{X})$$

Where trace takes the sum of the diagonal elements, i.e.

$$\text{trace}(\boldsymbol{A}) = \sum_{i=1}^{N} a_{i,i}$$

Discussion

- I want to go from **61st and West End Ave.** to **110th and Central Park West**
- I have created a coordinate system and a vector to my destination



- Which vector norm should I use to measure the length of my trip?

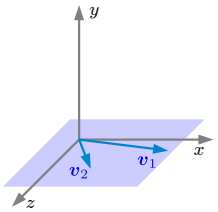
$$\|\boldsymbol{x}\|_p = \left(\sum_{n=1}^{N} |x_n|^p \right)^{1/p} \qquad \|\boldsymbol{x}\|_1 = \sum_{n=1}^{N} |x_n| \qquad \|\boldsymbol{x}\|_2 = \sqrt{\sum_{n=1}^{N} |x_n|^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$$

Vector spaces

- Vector spaces can be of arbitrary size
- Typically defined using a matrix of basis vectors
- The basis vectors must be **linearly independent**

$$\boldsymbol{V}\boldsymbol{x} = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{x} = \boldsymbol{0}$$

$$\boldsymbol{V} = \begin{bmatrix} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_N \\ | & | & & | \end{bmatrix}$$



Vector spaces

- Often the vectors are taken to be **mutually orthogonal** and of **unit lengt**

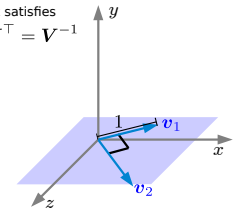
$$\boldsymbol{v}_i^\top \boldsymbol{v}_j = 0 \qquad \|\boldsymbol{v}_i\|_2 = 1$$

$$\boldsymbol{V} = \begin{bmatrix} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_N \\ | & | & & | \end{bmatrix}$$

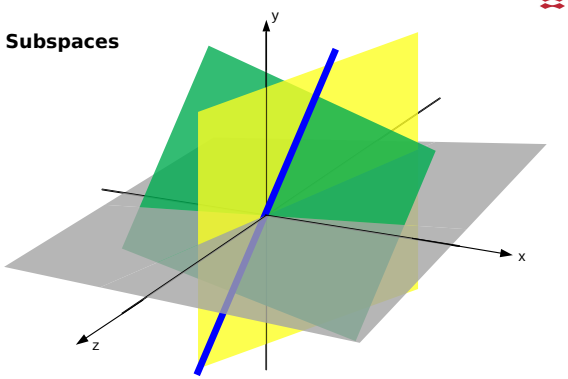
  - This defines an orthonormal basis for the vector space

- An orthonormal matrix satisfies

$$\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}, \quad \boldsymbol{V}^\top = \boldsymbol{V}^{-1}$$

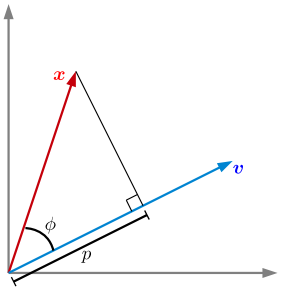


Subspaces



Projection

- Projection onto a vector



- Angle between vectors

$$\cos(\phi) = \frac{\boldsymbol{v}^\top \boldsymbol{x}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2}$$

- Length of projection

$$p = \|\boldsymbol{x}\|_2 \cos(\phi) = \frac{\boldsymbol{v}^\top \boldsymbol{x}}{\|\boldsymbol{v}\|_2}$$

- Projection onto unit vector

$$p = \boldsymbol{v}^\top \boldsymbol{x}$$

Projection onto subspace

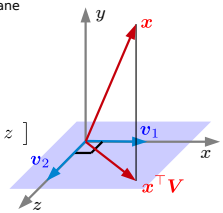
- **Projection onto a subspace**
  - Subspace defined by a orthonormal basis matrix
  - Projection given by

$$\boldsymbol{x}^\top \boldsymbol{V}$$

- **Example:** Projection of 3-D vector onto the (x,z) plane

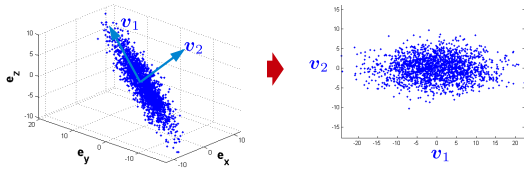
$$\boldsymbol{V} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\boldsymbol{x}^\top V = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} x & z \end{bmatrix}$$



Visualizing high dimensional data

- Humans are good at percieving data in 2-D
  - However we cannot visualize high dimensional data
- We can project high dimensional data to a lower dimensional subspace
- But what is a good projection?
  - Account for as much of the variation as possible

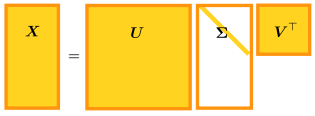


Singular Value Decomposition (SVD)

(Eugino Beltrami & Camille Jordan, independently, 1873-1874)

Any N x M matrix can be decomposed as follows:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$



$\sigma_1, \ldots, \sigma_M$ is known as the singular values

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_M \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M \geq 0$$

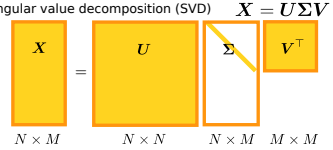
$$\boldsymbol{U} = \begin{bmatrix} | & & | \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_N \\ | & & | \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} | & & | \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_M \\ | & & | \end{bmatrix}$$

$$\text{if } i \neq j\text{: } \boldsymbol{\Sigma}_{i,j} = 0, \quad \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_{N \times N}, \quad \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_{M \times M}$$

Principal component analysis (PCA)

(Karl Pearson, 1901)

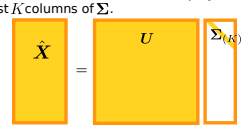
1) Subtract the mean from each attribute
2) Apply singular value decomposition (SVD) $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$



3) Select first K columns of $\boldsymbol{V}$ (the PCA projection operation) and first K columns of $\boldsymbol{\Sigma}$.

$$\boldsymbol{V}_{(K)} = \begin{bmatrix} | & & | \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_K \\ | & & | \end{bmatrix}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{v}_k = \sigma_k^2 \boldsymbol{v}_k$$



(PCA components or PCA projection of the data)

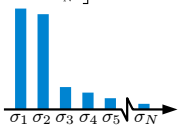
(PCA loadings)

Principal component analysis $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

- Entries in the diagonal matrix $\boldsymbol{\Sigma}$ are called **singular values**
  - They are sorted (largest first)
  - Indicate how much variability is explained by the corresponding component
    - 1st component explains most of the variability
    - 2nd component explains most of the remaining variability
    - Etc.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{bmatrix} \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$$

- Singular value spectrum



- Show that

$$\|\boldsymbol{X}\|_F^2 = \sum_i \sigma_i^2$$

where

$$\sigma_i = \boldsymbol{\Sigma}_{i,i}$$

Fraction of the variation in the data explained by the i[th] principal component is given by:

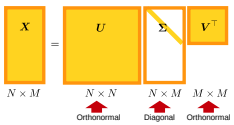
$$\frac{\sigma_i^2}{\sum_j \sigma_j^2}$$

And by the first K principal components

$$\frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_j \sigma_j^2}$$

Hints:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$



$$\|\boldsymbol{X}\|_F^2 = \mathrm{trace}(\boldsymbol{X}\boldsymbol{X}^\top)$$

$$\mathrm{trace}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{trace}(\boldsymbol{B}\boldsymbol{A})$$

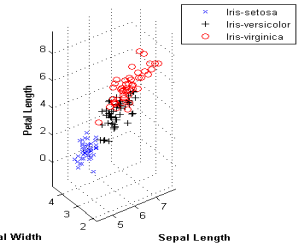
Fishers Iris Data

Three types of flowers:
Iris Setosa, Iris Versicolor, Iris Virginica

Flower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

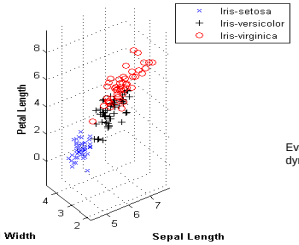
We will presently consider the first 3 attributes, i.e. Sepal length, Sepal Width and Petal Length.

3D scatter plot of the data



What fraction of the total variation in the data will the first principal component account for?

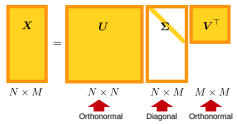
3D scatter plot of the data



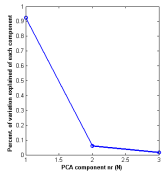
What fraction of the total variation in the data will the first principal component account for?

1) Subtract the mean
2) Apply singular value decomposition (SVD

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

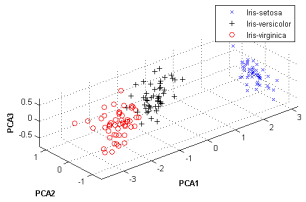


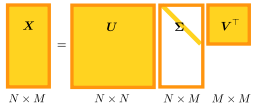
Evaluate the singular values to determine how much of the dynamics is lost when reducing the dimensionality



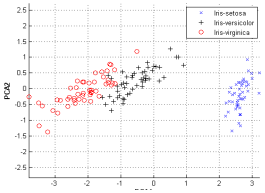
Visualizing the data projected onto the space of the principal components

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$





$$PCA1: \boldsymbol{p}_1 = \boldsymbol{X}\boldsymbol{v}_1 = \boldsymbol{u}_1\sigma_1$$
$$PCA2: \boldsymbol{p}_2 = \boldsymbol{X}\boldsymbol{v}_2 = \boldsymbol{u}_2\sigma_2$$
$$PCA3: \boldsymbol{p}_3 = \boldsymbol{X}\boldsymbol{v}_3 = \boldsymbol{u}_3\sigma_3$$



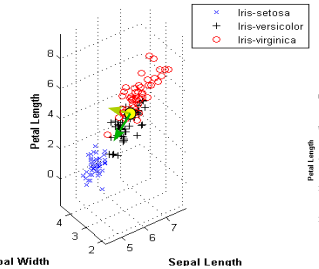
The principal directions V

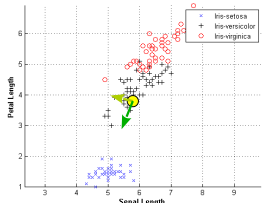
Sepal Length
Sepal Width
Petal Length

$$\boldsymbol{V} = \begin{bmatrix} -0.39 & -0.64 & -0.66 \\ 0.09 & -0.74 & 0.66 \\ -0.92 & 0.20 & 0.35 \end{bmatrix}$$

$$\boldsymbol{\mu} = \begin{bmatrix} 5.8 \\ 3.1 \\ 3.8 \end{bmatrix}, \quad \boldsymbol{v}_1 = \begin{bmatrix} -0.39 \\ 0.09 \\ -0.92 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} -0.64 \\ -0.74 \\ 0.20 \end{bmatrix}$$

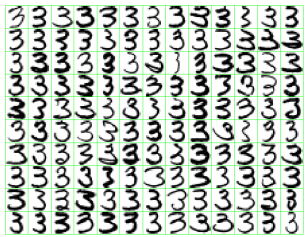
Sepal Length
Sepal Width
Petal Length





Visualization of hand written digits

- Data: Hand written 3's



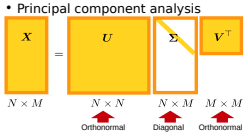
(Figure taken from "Elements of Statistical learning" by Trevor Hastie, Robert Tibshirani, Jerome Friedman)

- Data matrix

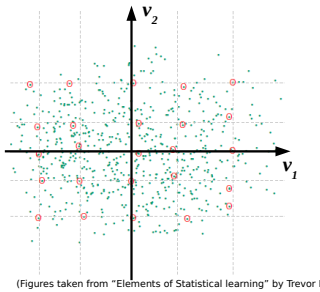
$$\boldsymbol{X} = \begin{bmatrix} \text{—} & \boldsymbol{x}_1 & \text{—} \\ \text{—} & \boldsymbol{x}_2 & \text{—} \\ & \vdots & \\ \text{—} & \boldsymbol{x}_N & \text{—} \end{bmatrix}$$

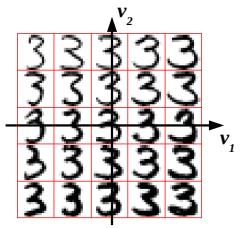
If each image is 16×16 pixels then $\boldsymbol{X}$ is a $N \times 256$ matrix.

- Principal component analysis



Visualization of hand written digits



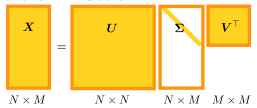


(Figures taken from "Elements of Statistical learning" by Trevor Hastie, Robert Tibshirani, Jerome Friedman)

What is the dynamics captured by the first two prinicipal components?

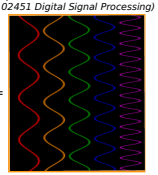
Data and Domain driven feature extraction

PCA is an example of a data driven approach for feature extraction
i.e., we define from data the features extracted in terms of the projections $\boldsymbol{V}^{(PCA)}$ that preserve most of the variation in the data

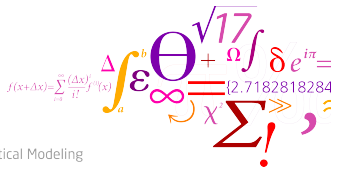


The fourier transform is an example of a domain driven approach for feature extraction
i.e., *in the analysis of sound good features are often to use spectral representations. These can be found by projecting the data using the so-called fourier transform matrix $\boldsymbol{V}^{(FFT)}$ where the components are defined as specific frequencies such that the projection of the data onto these frequencies defines the extend to which these frequencies are present in the data. (you can learn much more about this in 02451 Digital Signal Processing)*

$\boldsymbol{V}^{(FFT)}=$



02450 Introduction to machine learning and data mining (Preliminary Version)

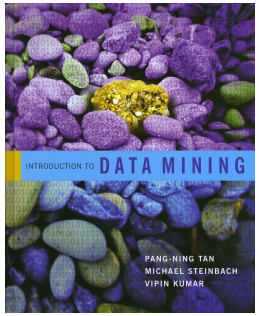


DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 2.4 + 3.1-3.2 + C1-C2

Groups of the day:
TBA



Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.3 + B1 (+ A))
3. **Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)**
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

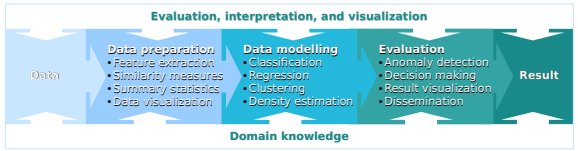
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3+8.5.7)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework

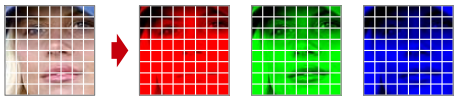


Todays learning objectives:
Be able to calculate various measures of similarity and dissimilarity.
Understand how various summmary statistics are calculated and can be interpreted
Explain and apply Bayes theorem
Understand the normal and multi-variate normal distribution and the role of the covariance matrix

Example: Principal component analysis of images

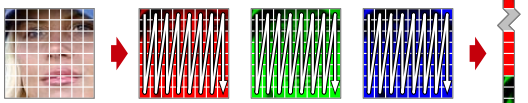
• 1000 images, 86 x 86 pixels, 3 RGB intensities
Tamara Berg "Faces in the wild"

Preprocessing



• Each image
  – 86 x 86 pixels
  – 3 RGB intensities
• Split image into red, green, and blue color channels

Preprocessing



• Concatenate all pixel color values in one long vector
  – 86 x 86 x 3 = 22.188
  – Image is now represented as a 22.188 dimensional vector

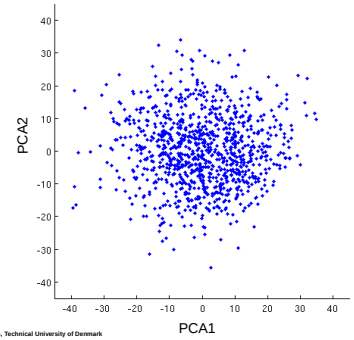
• Put all 1000 images into a big matrix
  – 1000 x 22.188

Principal component analysis (PCA)

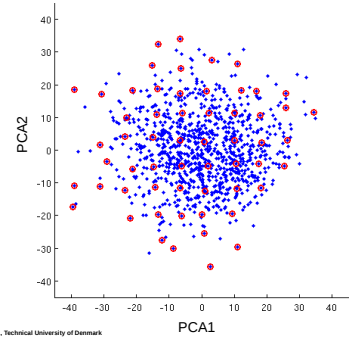
1. **Subtract the mean**
2. **Compute the singular value decomposition (SVD)**
   – Orthogonal linear transformation
   – Transforms data to a new coordinate system
     • Greatest variance along the first axis
     • Second greatest variance along the second axis
     • Etc.

• **Plot data in the transformed coordinate system**
   – Corresponds to looking at data from an angle where it is most spread out

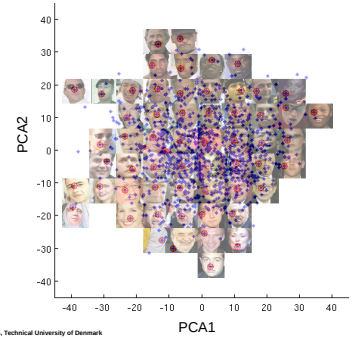
PCA of face images



PCA of face images

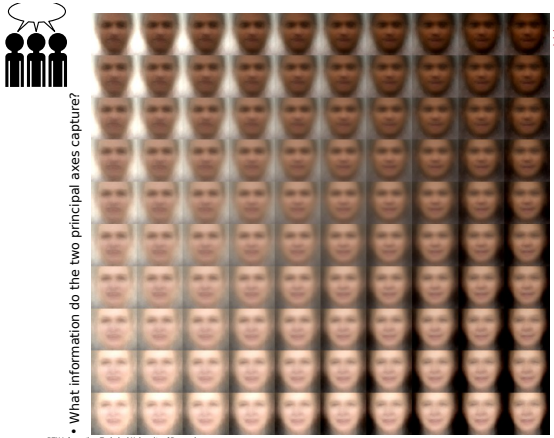


PCA of face images

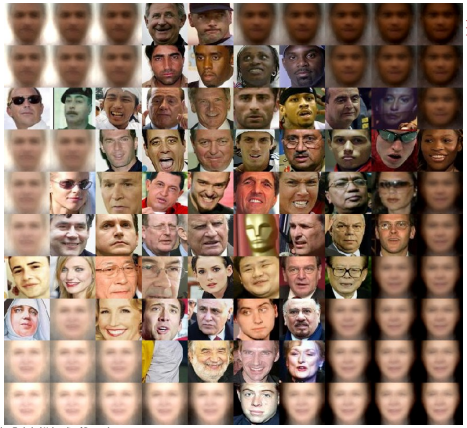


Discussion

- What information do the two principal axes capture?



- What information do the two principal axes capture?

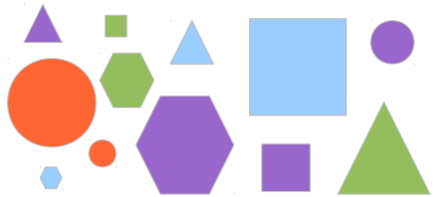


- What information do the two principal axes capture?

Similarity / Dissimilarity

- **Definition**
  - A numerical measure of *how alike/different* two data objects are
  - Often defined on the interval [0,1]
- **Similarity / dissimilarity between two data objects**

$$s(x, y), \quad d(x, y)$$



Dissimilarity measures

- **Euclidean distance (2-norm)**

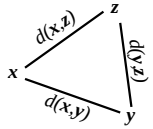
$$d_2(x, y) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

- **Minkowski distance (p-norm)**

$$d_p(x, y) = \left(\sum_{n=1}^{N} |x_n - y_n|^p \right)^{1/p}$$

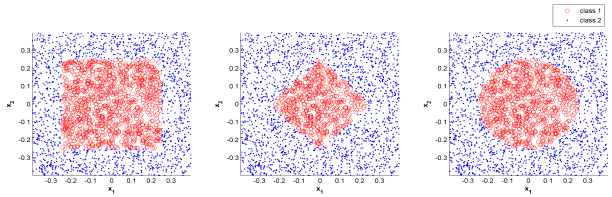
Properties of dissimilarity measures

- **Positivity**
  $d(x, y) \geq 0$
  $d(x, y) = 0 \Leftrightarrow x = y$
- **Symmetry**
  $d(x, y) = d(y, x)$
- **Triangle inequality**
  $d(x, z) \leq d(x, y) + d(y, z)$
- Measures that satisfy all these properties: **Metrics**



Minkowski distance

Which Minkowski distance (p-norm) can be used to separate the two classes, by measuring the distance to the origo (0,0)?



$$d_p(x, y) = \left(\sum_{n=1}^{N} |x_n - y_n|^p \right)^{1/p}$$

Binary similarity measures

- **Simple matching coefficient (SMC)**
  - Symmetric: Counts present and absent attributes equally

$$\mathrm{SMC}(x,y) = \frac{f_{00}+f_{11}}{K}$$

- **Jaccard coefficient**
  - Asymmetric: Counts only present attributes

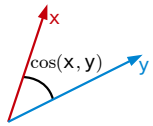
$$\mathrm{J}(x,y) = \frac{f_{11}}{K-f_{00}}$$

K : Total number of attributes
f_{00} : Number of attributes where $x_k = y_k = 0$
f_{11} : Number of attributes where $x_k = y_k = 1$

Continuous similarity measures

- **Cosine similarity**

$$\cos(x,y) = \frac{x^\top y}{\|x\|_2\|y\|_2}$$



- **Extended Jaccard coefficient**

$$\mathrm{EJ}(x,y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

Calculate the SMC, Jaccard, Cosine and Extended Jaccard similarity between customer 1 and customer 2 in the market basket data below.

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	0	1	1	0	1

$$\mathrm{SMC}(x,y) = \frac{f_{00}+f_{11}}{K}$$
$$\mathrm{J}(x,y) = \frac{f_{11}}{K-f_{00}}$$
$$\cos(x,y) = \frac{x^\top y}{\|x\|_2\|y\|_2}$$
$$\mathrm{EJ}(x,y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

K : Total number of attributes
f_{00} : Number of attributes where $x_k = y_k = 0$
f_{11} : Number of attributes where $x_k = y_k = 1$

Empirical statistics

- **Empirical mean**

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- **Empirical covariance**

$$\mathrm{cov}(x,y) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})$$

- **Empirical variance**

$$\mathrm{var}(x) = \mathrm{cov}(x,x) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})^2$$

- **Empirical standard deviation**

$$\mathrm{std}(x) = \sqrt{\mathrm{var}(x)}$$

Correlation

- **Measure of linear relation**

$$\mathrm{corr}(x,y) = \frac{\mathrm{cov}(x,y)}{\mathrm{std}(x)\mathrm{std}(y)}$$

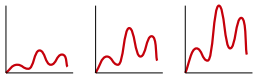
  - A correlation of **1** or **-1** means there is a perfect linear relation

$$x_k = a y_k + b$$

Invariance

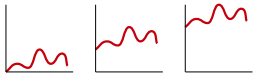
- **Scale**

$$d(x,y) = d(\alpha x, y)$$



- **Translation**

$$d(x,y) = d(\beta + x, y)$$

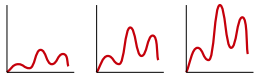


Discussion

- When would a **scale invariant** similarity measure be useful
  - Give an example
- When would a **translation invariant** similarity measure be useful
  - Give an example

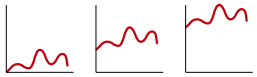
Scale invariance

$$d(x,y) = d(\alpha x, y)$$

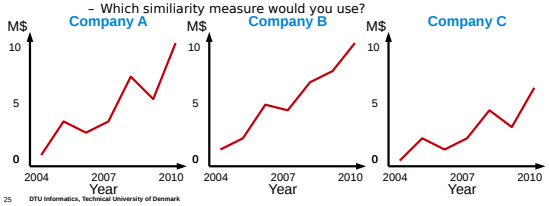


Translation invariance

$$d(x,y) = d(\beta + x, y)$$



- Which of these three companies are similar and in what way?
  - Which similiarity measure would you use?



Issues in proximity calculation

- **What to do when attributes have different**
  - **Scale?**
    - Standardize attributes
    - Use scale invariant similarity measure
  - **Type?**
    - Compute similarities for each attribute and combine
  - **Importance?**
    - Compute a weighted similarity measure

Standardization

- **Attributes have different scales.**
  - Example:
    - **Number of children** ~ 0-5
    - **Age** ~ 0-100 years
    - **Annual income** ~ 0-50.000 €

- Unless we do something, **Annual income** will dominate
  - **Standardization**: Subtract mean and divide by standard deviation

$$x_k^* = \frac{x_k - \bar{x}_k}{\mathrm{std}(x_k)}$$

Combining heterogeneous attributes

- **Attributes have different type**
  - Example:
    - **Age:** Continuous
    - **Education:** Binary
      - Primary (yes/no)
      - Secondary (yes/no)
      - Tertiary (yes/no)

- Similarity measure must handle **continuous** and **binary** features
  - **Compute similarities for each attribute and combine**

$$s_{\text{Age}} = d_1(x_{\text{Age}}, y_{\text{Age}})^{-1} \qquad s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

$$s(x, y) = \tfrac{1}{2}(s_{\text{Age}} + s_{\text{Edu.}})$$

Weighting attributes by importance

- **Attributes have different importance**
  - Example:
    - **Age:** Very important
    - **Education:** Less important

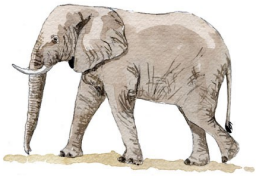
- Similarity measure must take **importance** into account
  - Introduce **importance weights** for each attribute

$$s_{\text{Age}} = d_1(x_{\text{Age}}, y_{\text{Age}})^{-1} \qquad s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

$$s(x, y) = 0.99 \cdot s_{\text{Age}} + 0.01 \cdot s_{\text{Edu.}}$$

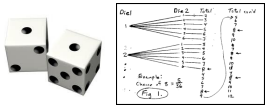
Discussion

- The following **attributes** are measured for a herd of elephants
  - Weight
  - Height
  - Tusk length
  - Trunk length
  - Ear area
  - Gender
- **Based on these measurements**
  - How would you evaluate how similar elephants are?
  - Justify your answer

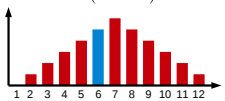


Probabilities

- **Discrete: Probability mass**
  - Example: The sum of two dice

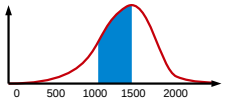


$P(X = v)$



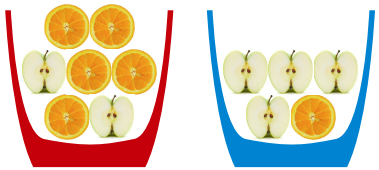
- **Continuous: Probability density**
  - Example: Lifetime of a light bulb

$$P(a \leq X \leq b) = \int_a^b p(X)\mathrm{d}X$$



Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Probabilities

- Basic rules of probability
  - Sum rule

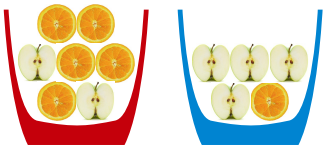
$$p(x) = \sum_y p(x, y)$$

  - Product rule

$$p(x, y) = p(x|y)p(y)$$

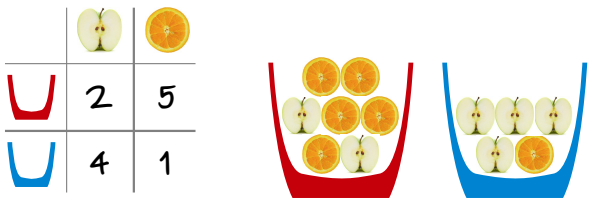
  - Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$



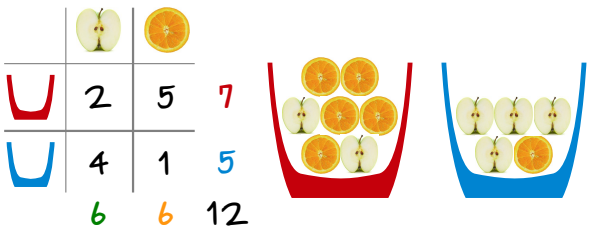
Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



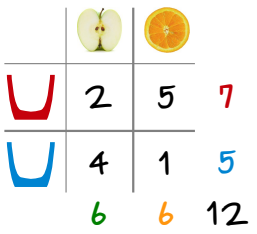
Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r, o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

$$= \frac{5/7 \cdot 7/12}{6/12} = 5/6$$

Medical test

A medical test for a given disease

- Correctly identifies the disease 99% of the time (true positives), and
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.

You go to the doctor to get tested, and the test turns out to be positive.
What is the probability you have the disease?

Hints:

- Identify from the text:
  p(Positive|Disease)
  p(Positive|No Disease)
  p(Disease)
  p(No Disease)
- Use the basic rules of probability given to the right to find:
  p(Disease|Positive)

$$p(x) = \sum_y p(x,y)$$

$$p(x,y) = p(x|y)p(y)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Frequency and mode

- **Frequency:** Percentage of time a value occurs
  - Example: Given the attribute **Gender** and a representative population of people, the value **Female** occurs about 50% of the time
- **Mode:** The most frequent attribute value
  - Example: Given the attribute **Operating System** and a representative population of computers, the value **Microsoft Windows** is the mode
- The notions of frequency and mode are typically used with categorical data

Percentiles

- **Percentiles:** Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the **p**th percentile is a value $\mathbf{x}_p$ of **x** such that **p** percent of the observed values of **x** are less than $\mathbf{x}_p$.
  - Example: The 10th percentile of **x** is the value $\mathbf{x}_{10\%}$ such that 10% of all values are less than $\mathbf{x}_{10\%}$.
- **Median:** The 50th percentile
  - Sort the numbers and take the middle value
    (if there are an even number of values, average the two middle values)

Measures of location

- **Mean:** Average

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- **Median:** Sort the numbers

$$\text{median}(x) = \begin{cases} x_{R+1} & \text{if } N \text{ is odd } (N = 2R+1) \\ \frac{1}{2}(x_R + x_{R+1}) & \text{if } N \text{ is even } (N = 2R) \end{cases}$$

Discussion

- What are the frequencies and the mode of these numbers and what is the mean and median?

  0, 1, 1, 3, 5, 590

  - Explain also what to be careful about when using the mean and median

- **Frequency:** Percentage of time a value occurs
- **Mode:** The most frequent attribute value
- **Mean:** Average

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- **Median:** Sort the numbers

$$\text{median}(x) = \begin{cases} x_{R+1} & \text{if } N \text{ is odd } (N = 2R+1) \\ \frac{1}{2}(x_R + x_{R+1}) & \text{if } N \text{ is even } (N = 2R) \end{cases}$$

Measures of spread

- **Range**

$$\text{range}(x) = \max(x) - \min(x)$$

- **Variance**

$$variance(x) = s_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

- **Absolute average deviation (AAD)**

$$\text{AAD}(x) = \frac{1}{N}\sum_{n=1}^{N}|x_n - \bar{x}|$$

- **Median absolute deviation (MAD)**

$$\text{MAD}(x) = \text{median}\{|x_1 - \bar{x}|, \ldots, |x_N - \bar{x}|\}$$

- **Interquartile range (IQR)**

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

Expected values

- Discrete random variable

$$\mathbb{E}\left[g(X)\right] = \sum_i g(x_i) P(X = x_i)$$

- Continuous random variable

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(X) p(X) \mathrm{d}X$$

Statistics

- **Mean**

$$\bar{x} = \mathbb{E}\left[x\right]$$

- **Covariance**

$$\mathrm{cov}(x, y) = \mathbb{E}\left[(x - \bar{x})(y - \bar{y})\right]$$

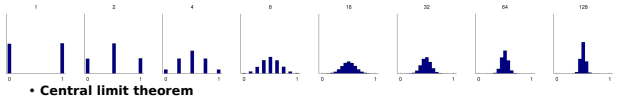
- **Variance**

$$\mathrm{var}(x) = \mathrm{cov}(x, x) = \mathbb{E}\left[(x - \bar{x})^2\right]$$

- **Standard deviation**

$$\mathrm{std}(x) = \sqrt{\mathrm{var}(x)}$$

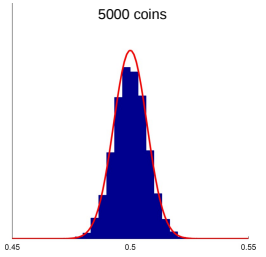
Normal distribution



- **Central limit theorem**
  - The mean of a large number of random variables will tend to a Normal distribution irrespective of the distribution of the random variables
    (Under certain conditions)

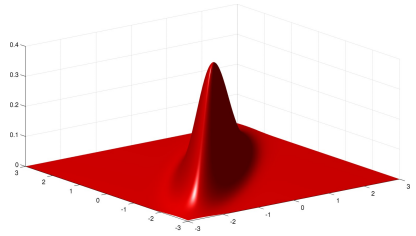
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Example:** Proportion of heads when flipping
  - 1 coin, 2 coins, 4 coins etc.



Multivariate Normal distribution

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$



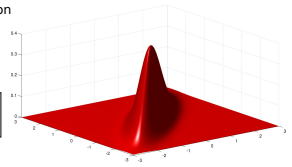
Multivariate Normal distribution

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

- Example: 2-dimensional Normal distribution

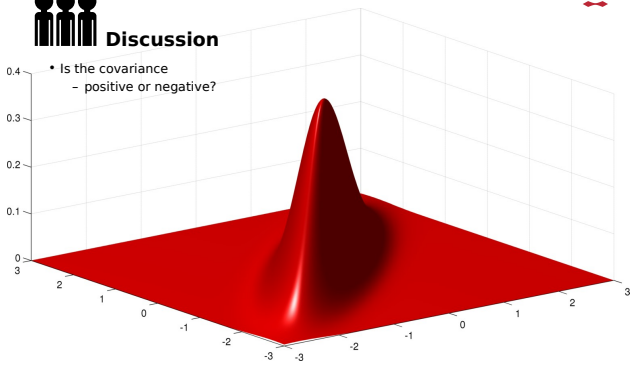
$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{bmatrix}$$



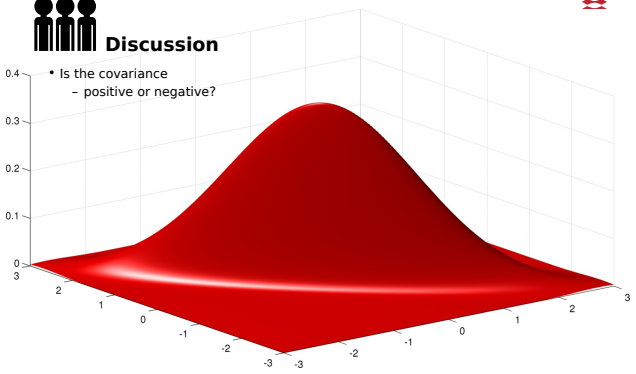
Discussion

- Is the covariance
  - positive or negative?



Discussion

- Is the covariance
  - positive or negative?



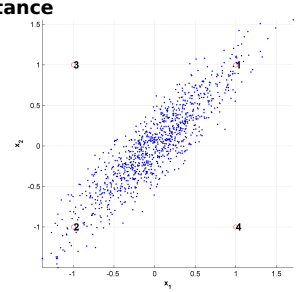
The Mahalanobis distance

How far are x_1 and x_2 apart?

- mahalanobis(x_1,x_2) = 4.2
- $d_{\text{Euclidean}}(x_1,x_2)^2$ = 8.0

How far are x_3 and x_4 apart?

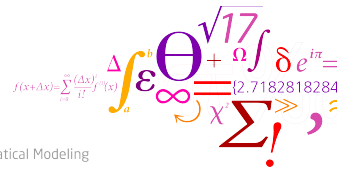
- mahalanobis(x_3,x_4) = 80
- $d_{\text{Euclidean}}(x_3,x_4)^2$ = 8.0



$$\mathrm{mahalanobis}(x, y) = (x\ y)^\top \Sigma^{-1} (x\ y)$$

$$d_{Euclidean}(x, y)^2 = (x\ y)^\top I^{-1} (x\ y)$$

02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 3.3

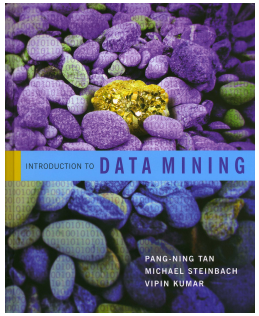
Feedback Groups of the day:
Søren Trads Steen
Janus Nørtoft Jensen

Veli Kerim Celik
Fadi Abdul Halim
Ziad Bahlawan

Andreas Jacobsen
Asger Anker Sørensen
Phong Le Trung

Anders Ulrik Eliasen
Rasmus Sten Andersen
Rasmus Lund Bendtsen

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. **Data visualization (Tan 3.3)**

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3, 8.5.7)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Probabilities (revisited from last week)

- Basic rules of probability
  - Sum rule

$$p(x) = \sum_y p(x, y)$$

  - Product rule

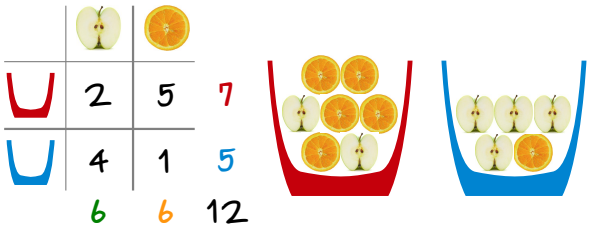
$$p(x, y) = p(x|y)p(y)$$

  - Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

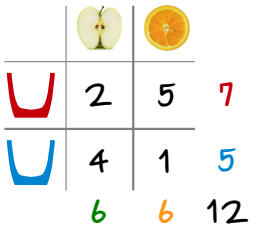
Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

$$= \frac{5/7 \cdot 7/12}{6/12} = 5/6$$

The news paper "Slam the Glam"

News media agency Reuters Bureau sends stories to the Economist:

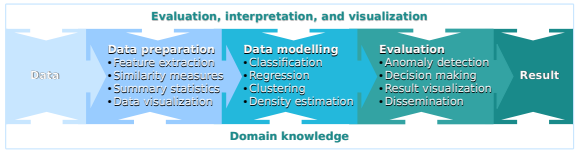
- 80% of the news stories from Reuters are positive and 20% of the news stories are negative.
- 90% of the negative news stories are published in the Economist while only 5% of the positive stories are published.

What is the probability that a a story from Reuters is published in the Economist given it is positive?

Hints:

- *Sum Rule* $p(x) = \sum_y p(x, y)$
- *Product rule* $p(x, y) = p(x|y)p(y)$
- *Bayes theorem* $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$

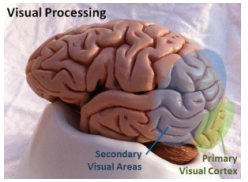
Data modeling framework

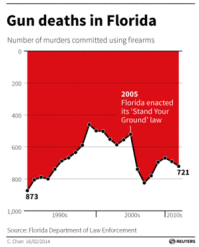


Todays learning objectives:
Be able to understand and apply a wide range of data visualization approaches
Understand the ACCENT principle and Tufte's guidelines

Visualization

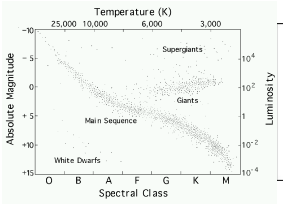
- A main function of the brain is to process visual information
- We can exploit this capacity using visualization of the data in order to:
  - Detect new patterns, i.e. exploratory data analysis)
  - **Dissiminate results, i.e. visualizations/plots in written work (todays lecture)**
- Since we are making use of the brains visual system we should take into account how the visual system work



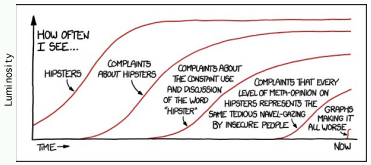


Visualization

- A visualization is an attempt to communicate/convince a reader about a particular point supported by the data.
- Keep the audience in mind (compare a popular talk to a technical article)



(technical article)



(popular talk)

(Sources: http://cass.ucsd.edu/, http://xkcd.com)

Three questions to ask of any visualization

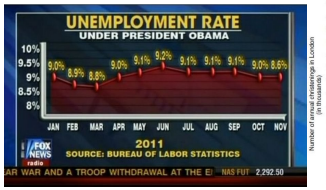
(From http://junkcharts.typepad.com/ (JC)

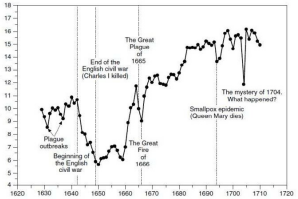
- What is the **Question?**
  - Any visualization should attempt to answer a particular question or argue a particular point.
  - Without a question there is no need for the visualization
- What does the **Data** say?
  - The data should be relevant for the question.
  - Applying appropriate transformations (not to few, not to many)
  - discarding irrelevant attributes
  - remove outliers
- What does the **Visual** say?
  - The visual element being choosen should address the question being asked using the data
  - Type of plot, use of colors, use of other graphical elements to help/confuse the reader

- This talk will mainly consider the **visual** question

Visualization as technical writing

- Just as there are no recipe for writing, there are no definate recipes for plotting
- However there are guidelines, often more apparent for writing than for illustrations
  - The purpose of the text is to communicate an idea *(vs. plots has a purpose)*
  - Be grammatically correct *(vs. elementary "rules" of good plotting)*
  - Ensure the text is readable *(vs. labels, legends or lines nobody can read)*
  - Avoid long/complicated paragraph *(vs. plots that are overly complicated)*
  - Dont lie or gloss over inconvenient facts. *(vs. distort data in a plot)*





Common difficulties for visualizations

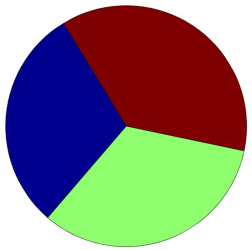
- **Representation:** How will you map objects, attributes, and relations to visual elements?
  - Positions, lengths, colors, areas, orientation

- **Arrangement:** How will you display the visual elements?
  - Viewpoint, transparency, separation, grouping

- **Selection:** How will you handle a large number of attributes and data objects?
  - Display a subset, focus on a region of interest, show summaries

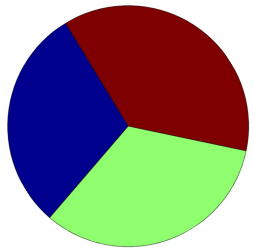
Representation

- **Area represents proportion**
  - Which is smallest, middle, and largest?
  - What are the proportions approximately?

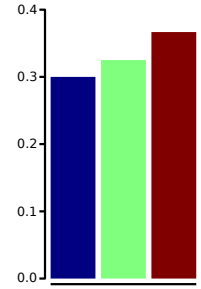


Representation

- **Area represents proportion**
  - Which is smallest, middle, and largest?
  - What are the proportions approximately?



- **Height represents proportion**



Arrangement

- **Placement of visual elements**
  - Can make a great difference in how easy it is to interpret data

- **Example**

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

Selection

- Elimination or de-emphasis of certain objects or attributes

- A subset of **attributes**
  - **Why?** A graph can only show so many attributes – focus on the relevant
  - **How?**
    - Dimensionality reduction
    - Plot pairs of attributes

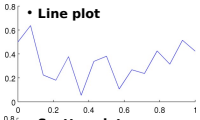
- A subset of **objects**
  - **Why?** A graph can only show so many objects – focus on the relevant
  - **How?**
    - Random sampling
    - Display of region of interest
    - Use density estimation

Types of plots

- **Distribution of a single attribute**
  - Histogram
  - Empirical cumulative distribution
  - Percentile plots
  - Box plot
- **Relation between attributes**
  - 2D histogram
  - Heat maps and contour plots
  - Scatter plots
- **Visualization of high-dimensional objects**
  - Matrix plots
  - Parallel coordinates
  - Star plots

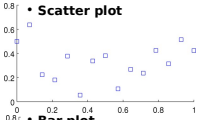
Basic plots

- **Line plot**



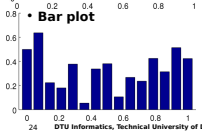
```
plot(x,y);
```

- **Scatter plot**



```
plot(x,y,'s');
scatter(x,y,'s')
```

- **Bar plot**



```
bar(x,y);
```


The iris data set

- **Three flowers**
  - 50 instances of each class, 150 in total
- **Attributes**
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

Fower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	**3.2**	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	**5.9**	**3.0**	**5.1**	**1.8**

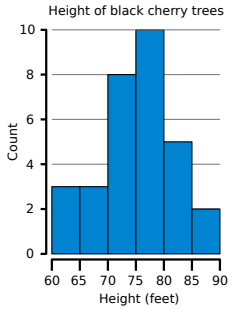
$X^{\text{Observation x Attribute}}$

Matlab var.	Type	Size	Description
`X`	Numeric	$N \times M$	Data matrix: The rows correspond to N data objects, each of which contains M attributes.
`y`	Numeric	$N \times 1$	Class index: For each data object, y contains a class index, $y \in \{0, 1, \ldots, C-1\}$, where C is the total number of classes.
`classNames`	Cell array	$C \times 1$	Class names: Name (string) for each of the C classes.
`attributeNames`	Cell array	$M \times 1$	Attribute names: Name (string) for each of the M attributes.
`N`	Numeric	Scalar	Number of data objects.
`M`	Numeric	Scalar	Number of attributes.
`C`	Numeric	Scalar	Number of classes.

Distribution of a single attribute

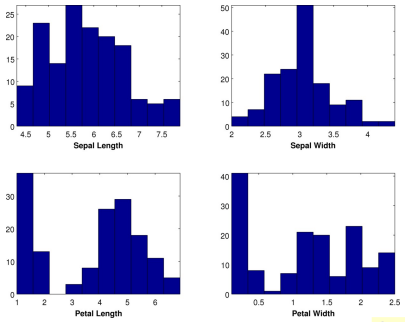
Histograms

- **Shows distribution of a single variable**
  - Divide the values into bins
  - Bar plot of the number of values in bin
  - Height indicates count of values
  - Shape determined by
    - Distribution of data
    - Number of bins / bin width



H = {60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89}

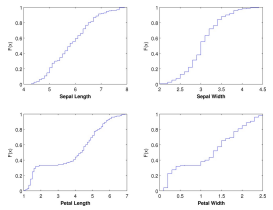
Histograms of the Iris data attributes




```
for m=1:M
    subplot(2,2,m)
    hist(X(:,m),10);
    axis tight;
    xlabel(attributeNames{m})
end
```


Empirical cumulative distributions

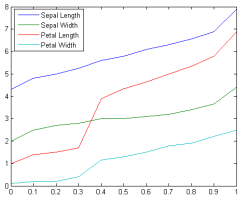
Cumulative Distribution: Given an ordinal or continuous attribute **x** and a value **z**, the commulative distribution of **F(z)** is the fraction of observations less than **z**.




```
for m=1:M
    subplot(2,2,m);
    ecdf(X(:,m));
    xlabel(attributeNames{m});
end
```

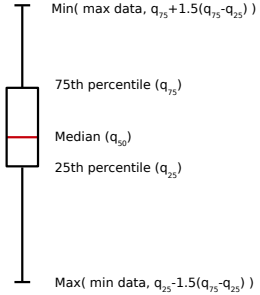

Percentile plots

Percentiles: Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the **p**th percentile is a value $\mathbf{x}_p$ of **x** such that **p** percent of the observed values of **x** are less than $\mathbf{x}_p$.

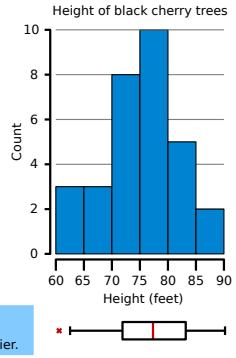



```
prctile = 0:0.1:1;
Y = quantile(X,prctile);
plot(prctile,Y);
legend(attributeNames);
```

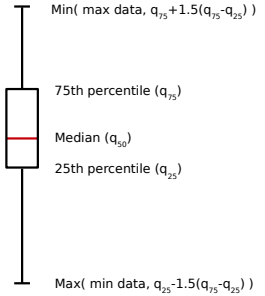
Box plots

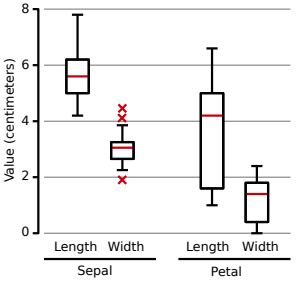


The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.



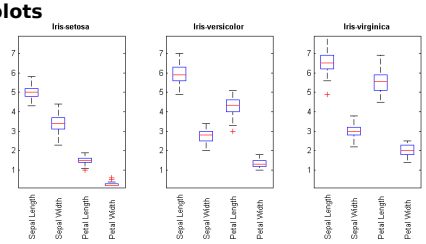
Box plots





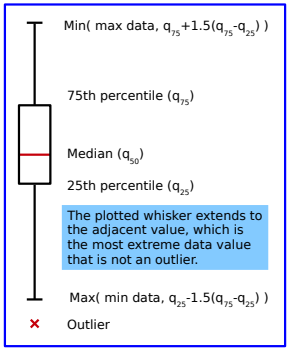
The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.

Box plots




```
y_lim=[min(min(X)) max(max(X))];
for c = 1:C
    subplot(1,C,c)
    boxplot(X(y==c-1,:),'labels',attributeNames,'labelorientation','inline');
    axis([0.5 M+0.5,y_lim])
    title(classNames{c},'FontWeight','bold');
end
```

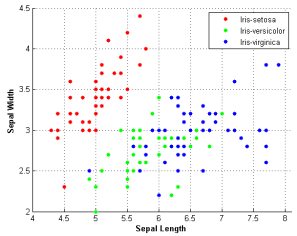

Homework: Draw boxplot for dataset D={0, 1, 1, 3, 3, 4, 4, 10} (boxplot(D); **in matlab**)



Relation between attributes

Scatter plots

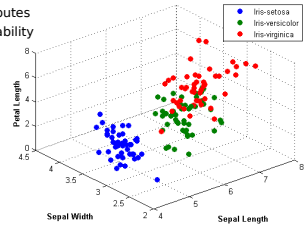
- Shows **relation** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability




```
gscatter(X(:,1),X(:,2),classNames(y+1));
grid on;
xlabel(measurementType(1),'fontweight','bold');
ylabel(measurementType(2),'fontweight','bold');
```


Scatter plots

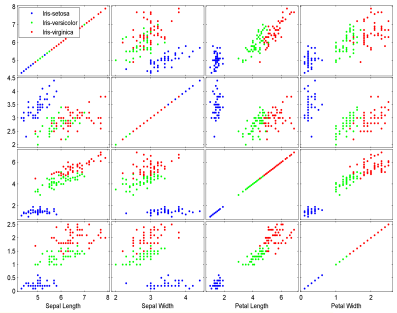
- Shows **relation** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability
  - 3D plots are often confusing; avoid if possible




```
hold all;
for c = 1:C
    scatter3(X(y==c-1,1),X(y==c-1,2),X(y==c-1,3),'filled');
end
grid on;
xlabel(attributeNames{1},'FontWeight','bold');
ylabel(attributeNames{2},'FontWeight','bold');
zlabel(attributeNames{3},'FontWeight','bold');
legend(classNames)
```


Scatter plots

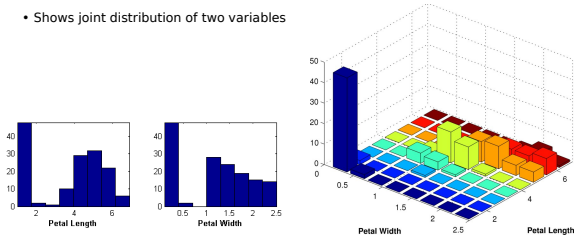
- Scatter plot matrix
  - All pairs of attributes




```
gplotmatrix(X,X,classNames(y+1),[],[],[],'on','',attributeNames);
```


Two-dimensional histograms

- Shows joint distribution of two variables



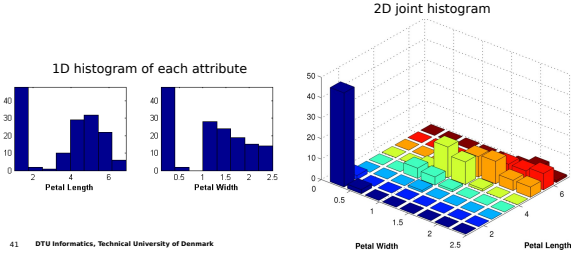
1D histogram of each attribute

2D joint histogram

```
nBins = 8;
ind = [3 4];  % Indices of the two variables to create the 2D histogram from
[n,x,data] = hist2d(X(:,ind)',nBins);
bar3xy(x(1,:),x(2,:),n);
axis tight;
xlabel(attributeNames{ind(1)},'FontWeight','bold')
ylabel(attributeNames{ind(2)},'FontWeight','bold')
```


Let B be the number of bins for each mode in the 1-D and 2-D histogram (B=8 in the figures below). How many values are there to be estimated in the 1-D and 2-D histogram as a function of B? What happen to the number of estimated values if we estimated a 3-D histogram of the joint distribution between three attributes? What happens if we estimated a n-D histogram of the joint distribution of n attributes?

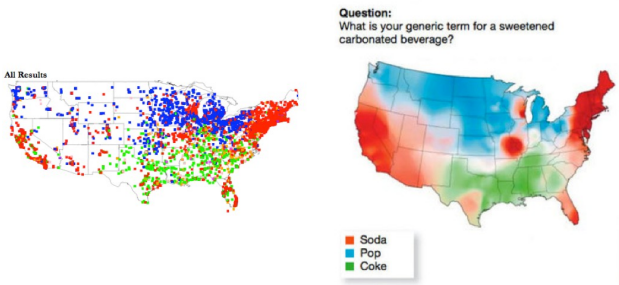
2D joint histogram



1D histogram of each attribute

Other approaches to 2D density visualization

- 3-D plots should be avoided when possible. Consider instead **heatmaps**, **contour** plots or **images** instead of 2D histograms



- When using colors, choose a color scheme which use one color per class and let the intensity of the color indicate strength.

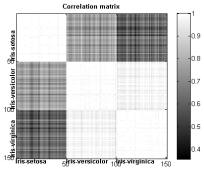
Visualization of high-dimensional objects

Matrix plots

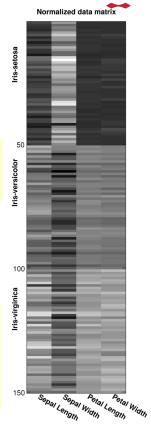
- **Plot of raw data matrix**
  - Useful when objects are sorted according to class
  - Typically, attributes are normalized

```
Z = zscore(X); % Standardize data
imagesc(Z);
colormap(gray)
title('Normalized data matrix');
```

- **Plots of similarity matrices**
  - Useful for visualizing the relation between objects

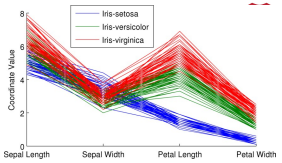



```
C = corrcoef(X');
h2 = imagesc(C);
colormap(gray);
axis equal;
axis tight;
colorbar;
title('Correlation matrix');
```



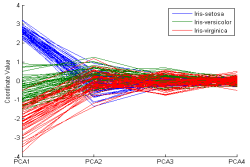
Parallel coordinates

- **Plot high-dimensional data**
- Instead of perpendicular axes
  - Use parallel axes
- Attribute values are plotted as a point
  - and the points are connected by a line
- Each object is represented as a line
- Lines representing a group of objects
  - Are similar in some sense
  - Ordering of attributes is important in seeing such groupings



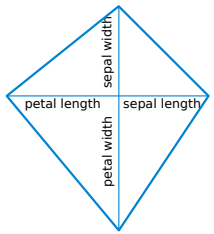

```
parallelcoords(X,'group',classNames(y+1),...
'labels',attributeNames);
```

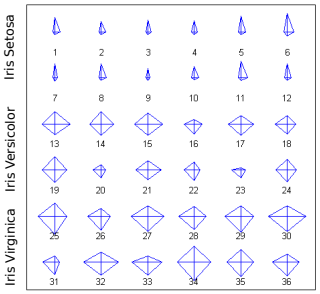
```
% Project data by PCA
Y = bsxfun(@minus, X, mean(X));
[U, S, V] = svd(Y);
Z = U*S;
% Plot data in space spanned by PCA
parallelcoords(Z,'group',classNames(y+1),...
'labels',{'PCA1','PCA2','PCA3','PCA4'});
```



Star plots

- **Plot high-dimensional data**
- Similar to parallel coordinates
  - Axes radiate from center
  - Connecting line is a polygon






```
indGlyph=[1:12 50+(1:12) 100+(1:12)];
glyphplot(X(indGlyph,:), 'glyph','star');
```

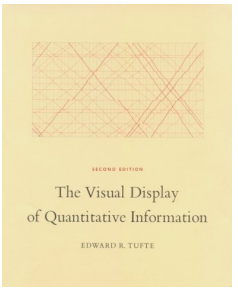

ACCENT

- **Apprehension**
  - Is it easy to see what is important in the graph?
- **Clarity**
  - Are the most important elements visually most prominent?
- **Consistency**
  - Have you used the same colors, shapes, etc. as in other graphs?
- **Efficiency**
  - Does it convey its information in the most simple and efficient way?
- **Necessity**
  - Are all elements of the graph necessary to represent data?
- **Truthfulness**
  - Does the graph represent the data correctly?

Tufte's guidelines

- **Graphical excellence**
  - Well-designed presentation of interesting data – a matter of
    - substance, statistics, and design
  - Complex ideas communicated with
    - clarity, precision, and efficiency
  - Gives the viewer
    - the greatest number of ideas
    - in the shortest time
    - with the least ink
    - in the smallest place.
  - Nearly always multivariate
  - Requires telling the truth about the data
  - Maximise Data-ink ratio:

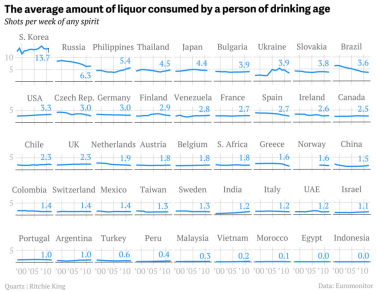
$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used}}$$



Examples

Example of the three questions in practice:

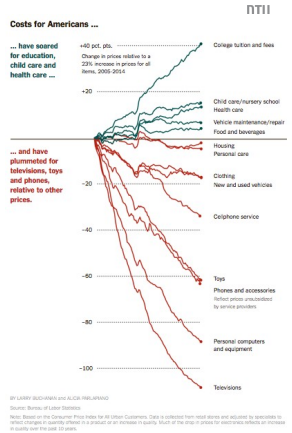
- What is the **objective**?
- What does the **data** say?
- What does the **visual** say?



Source: Richie King, JC.

Example of the three questions in practice:

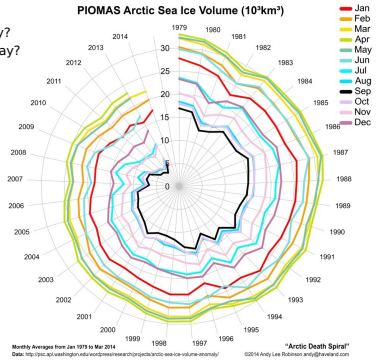
- What is the **objective**?
- What does the **data** say?
- What does the **visual** say?



Source: NYT's CPI Component Price Drop chart (JC)

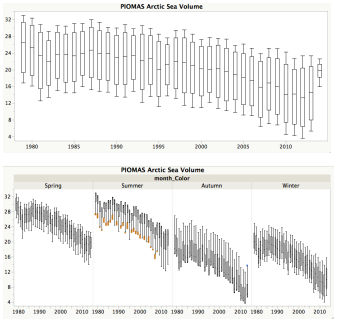
Example of the three questions in practice:

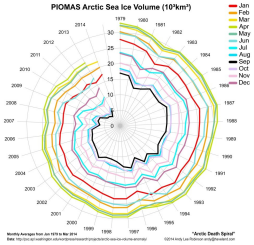
- What is the **objective**?
- What does the **data** say?
- What does the **visual** say?



Example of the three questions in practice:

- What is the **objective**?
- What does the **data** say?
- What does the **visual** say?

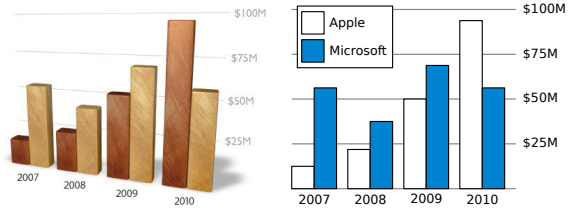




Source: JC.

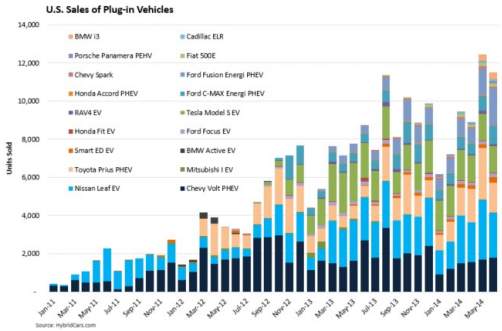
Apprehension: Is it easy to see what is important in the graph?
Clarity: Are the most important elements visually most prominent?
Consistency: Have you used the same colors, shapes, etc. as in other graphs?
Efficiency: Does it convey its information in the most simple and efficient way?
Necessity: Are all elements of the graph necessary to represent data?
Truthfulness: Does the graph represent the data correctly?

- Compare the graphs using ACCENT

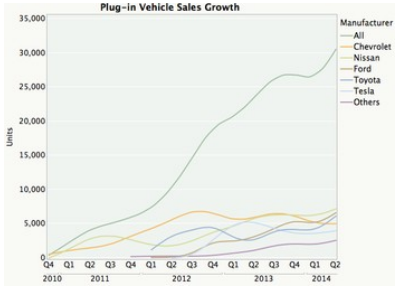


Problems with bar charts

(Source: JC)



The problem with bar charts (cont.)

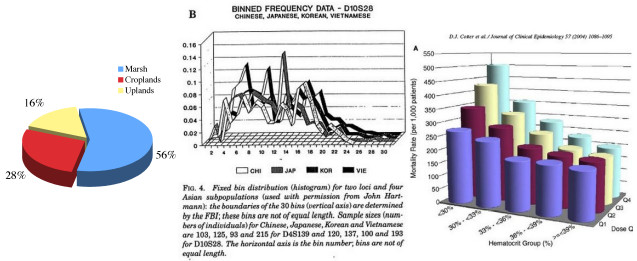


- Bar charts use a lot of space to convey little information. If one wish to convey a lot of information bar charts should be avoided

Source: JC

The horrorcabinet

- No matter who you are writing to (even the Pointy-haired boss or 3rd grade students) you should never, ever use the **trivial persuit**, **spaghetti** or **stairs** plot

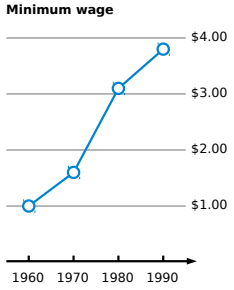


Apprehension: Is it easy to see what is important in the graph?
Clarity: Are the most important elements visually most prominent?
Consistency: Have you used the same colors, shapes, etc. as in other graphs?
Efficiency: Does it convey its information in the most simple and efficient way?
Necessity: Are all elements of the graph necessary to represent data?
Truthfulness: Does the graph represent the data correctly?

- Compare the graphs using ACCENT

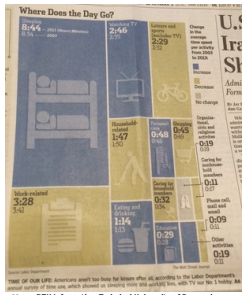
Minimum wage

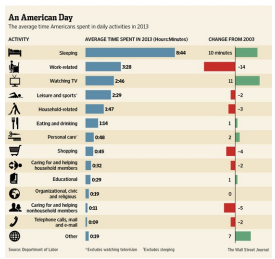
1960		$1.00
1970		$1.60
1980		$3.10
1990		$3.80



Apprehension: Is it easy to see what is important in the graph?
Clarity: Are the most important elements visually most prominent?
Consistency: Have you used the same colors, shapes, etc. as in other graphs?
Efficiency: Does it convey its information in the most simple and efficient way?
Necessity: Are all elements of the graph necessary to represent data?
Truthfulness: Does the graph represent the data correctly?

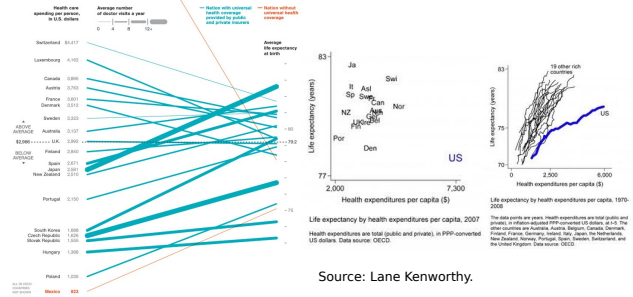
- Compare the graphs using ACCENT *(Source: WSJ & JC)*





Apprehension: Is it easy to see what is important in the graph?
Clarity: Are the most important elements visually most prominent?
Consistency: Have you used the same colors, shapes, etc. as in other graphs?
Efficiency: Does it convey its information in the most simple and efficient way?
Necessity: Are all elements of the graph necessary to represent data?
Truthfulness: Does the graph represent the data correctly?

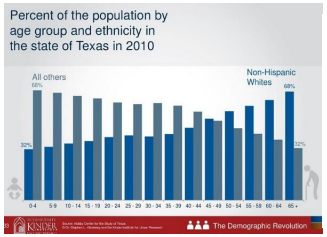
- Compare these graphs using ACCENT

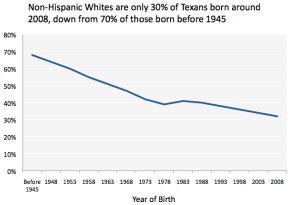


Source: Lane Kenworthy.

Apprehension: Is it easy to see what is important in the graph?
Clarity: Are the most important elements visually most prominent?
Consistency: Have you used the same colors, shapes, etc. as in other graphs?
Efficiency: Does it convey its information in the most simple and efficient way?
Necessity: Are all elements of the graph necessary to represent data?
Truthfulness: Does the graph represent the data correctly?

- Compare these graphs using ACCENT (Source: JC)





Imagine you have a dataset where some of the attributes are numeric given in the matrix X but you also have a categorical attribute given by TXT (see below). You would like to carry out a PCA on the data taking both the numeric and categoric attributes into account. How would you proceed?

One-out-of-K coding

Age (Standardized)	Height	Weight	Nationality	Denmark	Norway	Sweden
-0.2248	-0.4762	-0.2097	'Sweden'	0	0	1
-0.5890	0.8620	0.6252	'Sweden'	0	0	1
-0.2938	-1.3617	0.1832	'Sweden'	0	0	1
-0.8479	0.4550	-1.0298	'Sweden'	0	0	1
-1.1201	-0.8487	0.9492	'Norway'	0	1	0
2.5260	-0.3349	0.3071	'Norway'	0	1	0
X= 1.6555	0.5528	0.1352	TXT= 'Norway'	X_tmp= 0	1	0
0.3075	1.0391	0.5152	'Norway'	0	1	0
-1.2571	-1.1176	0.2614	'Norway'	0	1	0
-0.8655	1.2607	-0.9415	'Sweden'	0	0	1
-0.1765	0.6601	-0.1623	'Norway'	0	1	0
0.7914	-0.0679	-0.1461	'Denmark'	1	0	0
-1.3320	-0.1952	-0.5320	'Denmark'	1	0	0
-2.3299	-0.2176	1.6821	'Sweden'	0	0	1
-1.4491	-0.3031	-0.8757	'Sweden'	0	0	1
0.3335	0.0230	-0.4838	'Sweden'	0	0	1
0.3914	0.0513	-0.7120	'Denmark'	1	0	0
0.4517	0.8261	-1.1742	'Sweden'	0	0	1
-0.1303	1.5270	-0.1922	'Norway'	0	1	0
0.1837	0.4669	-0.2741	'Denmark'	1	0	0

```
[X_tmp, attributeNames_tmp]=categoric2numeric(TXT);
X=[X X_tmp];
attributeNames=[attributeNames; attributeNames_tmp];
```

X=

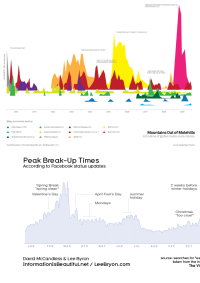
Age	Height	Weight	Denmark	Norway	Sweden
-0.2248	-0.4762	-0.2097	0	0	1.0000
-0.5890	0.8620	0.6252	0	0	1.0000
-0.2938	-1.3617	0.1832	0	0	1.0000
-0.8479	0.4550	-1.0298	0	0	1.0000
-1.1201	-0.8487	0.9492	0	1.0000	0
2.5260	-0.3349	0.3071	0	1.0000	0
1.6555	0.5528	0.1352	0	1.0000	0
0.3075	1.0391	0.5152	0	1.0000	0
-1.2571	-1.1176	0.2614	0	1.0000	0
-0.8655	1.2607	-0.9415	0	0	1.0000
-0.1765	0.6601	-0.1623	0	1.0000	0
0.7914	-0.0679	-0.1461	1.0000	0	0
-1.3320	-0.1952	-0.5320	1.0000	0	0
-2.3299	-0.2176	1.6821	0	0	1.0000
-1.4491	-0.3031	-0.8757	0	0	1.0000
0.3335	0.0230	-0.4838	0	0	1.0000
0.3914	0.0513	-0.7120	1.0000	0	0
0.4517	0.8261	-1.1742	0	0	1.0000
-0.1303	1.5270	-0.1922	0	1.0000	0
0.1837	0.4669	-0.2741	1.0000	0	0

Some guidelines I follow for technical writing

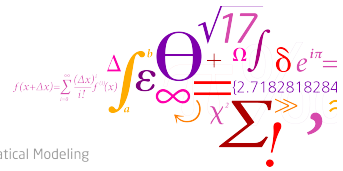
- Everything should be vector graphics at all times (PDF or EPS; the zoom-in test)
- Print out the graphics and ensure it is readable
- Never use a piechart. Never use something *worse* than a pie chart
- Avoid 3D plots if at all possible (heat-maps or contour plots are preferable)
- Under no circumstances must you add a 3D effect to a plot
- Excel is not for serious work
- Always use a white background. Don't add colors unless it improves the plot
- The best plots are often **"plot(x,y,'.-');"**. Can the point be made like this?
- Turn of the box around the plot ("**box off;**" in matlab)
- Consider the scale/location of axis: "**axis equal;**" if axis are of the same type
- Automate as much as possible (labels, legends, axis). Avoid the Matlab plot editor
- Try to avoid graphics programs. I prefer **Matlab+matlabfrag+Latex+Tikz**
- Use image captions. Add a *"mini-conclusion"* to the caption to tell a reader what he should/can take away from a graphics if it is not apparent
- Ask others about your plot without explaining them at first
- Be aware a typical reader won't read your main text
- **Use common sense**

Making good data visualizations is an art

For some interesting data visualizations see also
http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html
http://www.informationisbeautiful.net/
http://www.junkcharts.typepad.com/



02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 4.1-4.3 + Appendix D

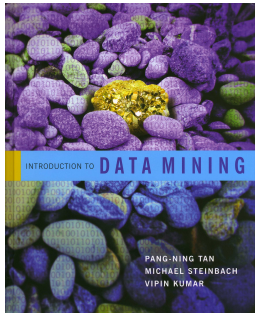
Feedback Groups of the day:
Anne Brandt Nielsen
Julie Sofie Bjørn Olsen
Emilie Lundbye Dalsgaard

Harri Laine
Istvan Szonyi

Astrid Ottosen
Nanna Elkjær Møller
Dzenita Omanovic

Andreas Olsen
Shafi Nazar
Qian Junging

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

Report 1

1. Introduction **(Tan 1.1-1.4)**

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.2 +(A)+ B.1)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

Report 2

5. **Decision trees and linear regression (Tan 4.1-4.3 + D)**
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)

Report 3

8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

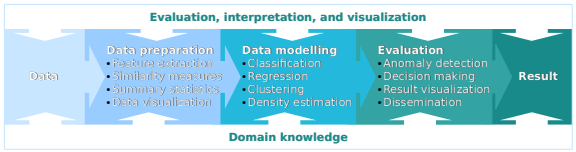
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3 + 8.5.7)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

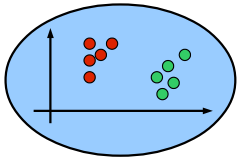
12. Putting it all together: Summary and overview
13. Mini project presentation

Data modeling framework

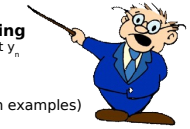


After today you should be able to:
Explain what supervised learning is
Explain the difference between classification and regression
Be able to evaluate classifiers in terms of the confusion matrix, error rate and accuracy
Understand the principles behind decision trees and Hunt's algorithm
Apply and interpret decision trees, linear regression and logistic regression

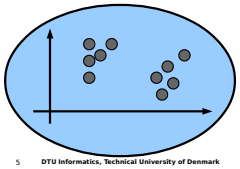
Supervised and Unsupervised learning



Supervised Learning
Input data $\mathbf{x}_n$ and output y_n



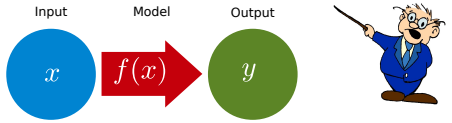
(Generalize from known examples)



Unsupervised Learning
Input data $\mathbf{x}_n$ alone

(Exploratory analysis)

Supervised learning



- **Data**
  - Inputs and outputs *(this is what we are given)*

$$\{x_n, y_n\}_{n=1}^{N}$$

- **Model**
  - Function that maps inputs to outputs *(what we are trying to determine)*

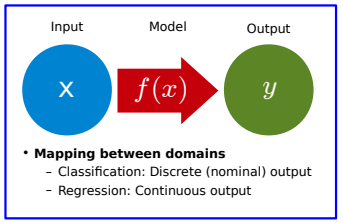
$$f(x)$$

- **Cost function**
  - Dissimilarity measure between observation and prediction *(how we tell if a model is good or bad)*

$$d(y, f(x))$$

- **Types of supervised learning**
  - Regression: Continuous output **y**
  - Classification: Discrete output **y**

Give an example of a classification and a regression problem and explain what the model *f(x)* can be used for.



- **Mapping between domains**
  - Classification: Discrete (nominal) output
  - Regression: Continuous output

Classification

- **Definition:** Learning a function that maps a data object to a discrete class

- **Why classify?**
  - Descriptive modeling
    - Explain / understand the relation between attributes and class
  - Predictive modeling
    - Predict the class of a new data object

Confusion matrix

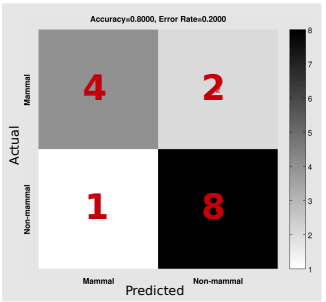
- Visualization of actual versus predicted class labels

- **Accuracy**
  (Number of correctly predicted observations divided by the total number of observations)

$$\frac{4+8}{4+2+1+8} = 80\%$$

- **Error rate**
  (Number of in-correctly predicted observations divided by the total number of observations)

$$\frac{2+1}{4+2+1+8} = 20\%$$



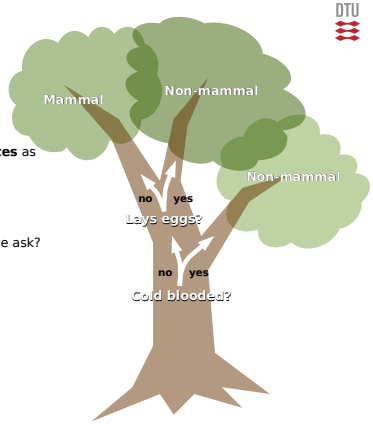
Decision trees

- Remember the game “20 questions to the professor”? (see also www.20q.new)

Q1. Is it an Animal? Yes.
Q2. Can you hold it? No.
Q3. Does it live in groups (gregarious)? Yes.
Q4. Are there many different sorts of it? No.
Q5. Can it jump? Yes.
Q6. Does it eat seeds? No.
Q7. Is it white? Sometimes.
Q8. Is it black and white? No.
Q9. Does it have paws? Yes.
Q10. Can you see it in a zoo? Yes.
Q11. Does it roar? Yes.
Q12. Is it worth a lot of money? Yes.
Q13. Does it have spots? Yes.
Q14. Is it multicoloured? Yes.
Q15. Can you make money by selling it? Yes.
Q16. Does it live in the jungle? Yes.
Q17. I guessed that it was a leopard? Wrong.
Q18. Does it like to play? Yes.
Q19. I guessed that it was a cheetah? Wrong.
Q20. I am guessing that it is a siberian tiger? Correct.

Decision trees

- Ask a series of questions until a conclusion is reached
- **Example**: Classify **vertebrates** as
  - **Mammal** or
  - **Non-mammal**
- **Learning task**
  - Which questions should we ask?



Hunts algorithm

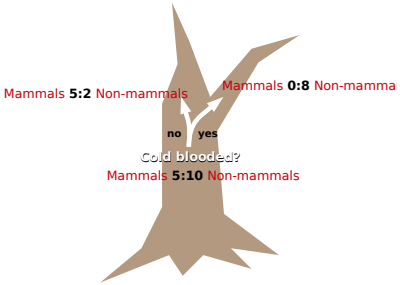
- Assign all data objects to the root

Hunts algorithm

- Select an attribute test condition
  - Find a good question to ask

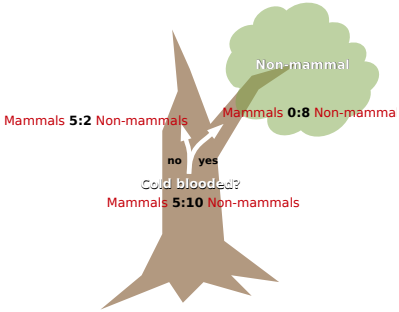
Hunt's Algorithm

- Partition the data objects into subsets according to the test condition



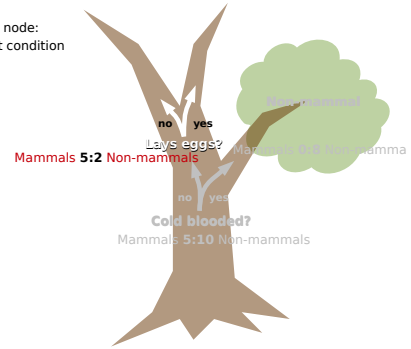
Hunts algorithm

- If all data objects belong to the same class
  - Create a leaf node



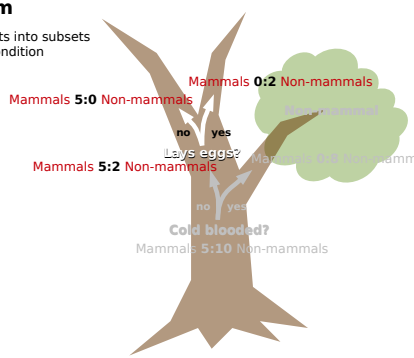
Hunts algorithm

- Repeat for each non-leave node:
  - Select an attribute test condition



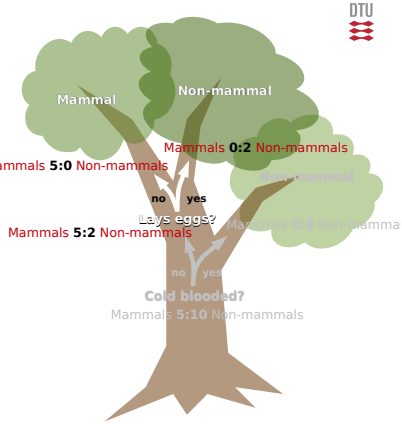
Hunts algorithm

- Partition the data objects into subsets according to the test condition



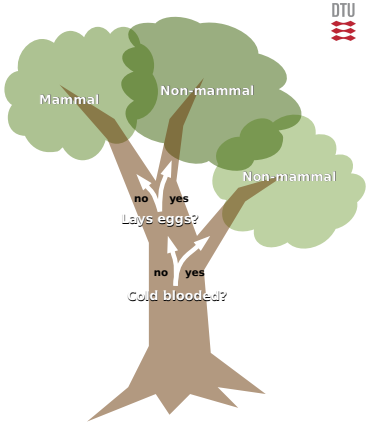
Hunts algorithm

- If all data objects belong to the same class
  - Create a leaf node



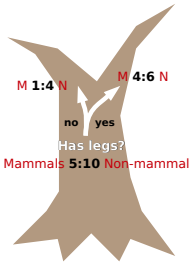
Hunts algorithm

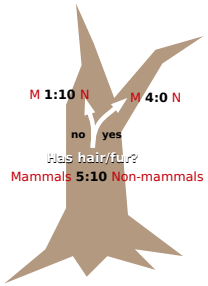
- But how do we find the **best question** at each step?



Selecting the best split

- Which of these two questions is best and why?





Selecting the best split

- Consider a large number of possible splits
- Compute a measure of impurity after the proposed split
  - For each new branch of the tree
  - Compute weighted average impurity
- Choose split that reduces impurity most

Selecting the best split: Impurity measures

- Compute the purity gain, Δ

$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i\|t) log_2 p(i\|t)$	$\Delta = ?$	$\Delta = ?$
$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} \left(p(i\|t)\right)^2$	$\Delta = ?$	$\Delta = ?$
$\text{Class. error}(t) = 1 - \max_i p(i\|t)$	$\Delta = ?$	$\Delta = ?$

$p(i|t)$ Fraction of objects that belong to class i
$N(v_j)/N$ Fraction of animals in branch v_j

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

Hint: Look at page 158-160 in Tan et al.

Selecting the best split: Impurity measures

- Compute the purity gain, Δ

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i|t) log_2 p(i|t)$$

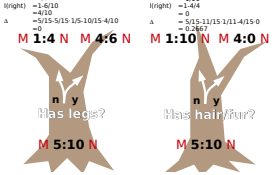
$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} \left(p(i|t)\right)^2$$

$$\text{Class. error}(t) = 1 - \max_i p(i|t)$$

$p(i|t)$ Fraction of objects that belong to class i

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

Hint: Look at page 158-160 in Tan et al.

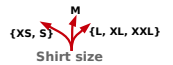


Which splits to consider

- **Binary**

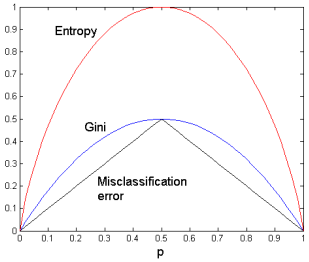
- **Nominal**

- **Ordinal**

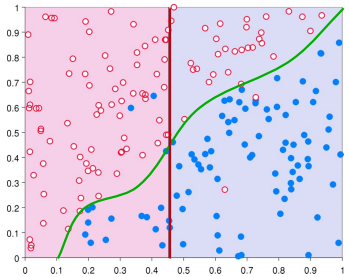


- **Continuous**

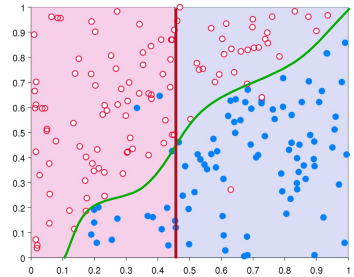
For a two class problem



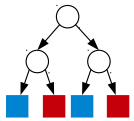
Classification Trees

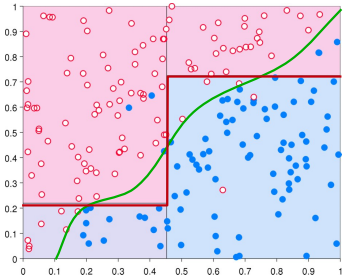


Classification Trees

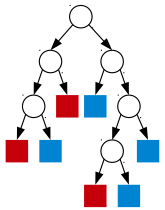


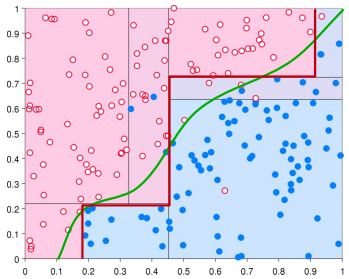
Classification trees





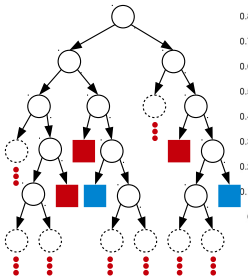
Classification trees

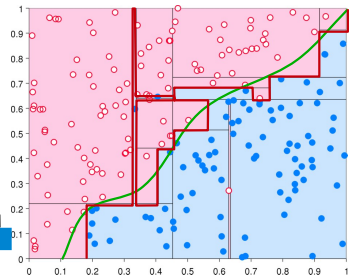




Classification trees

Common stopping criteria:
All records have the same class label
The number of observations have fallen below some minimum treshold



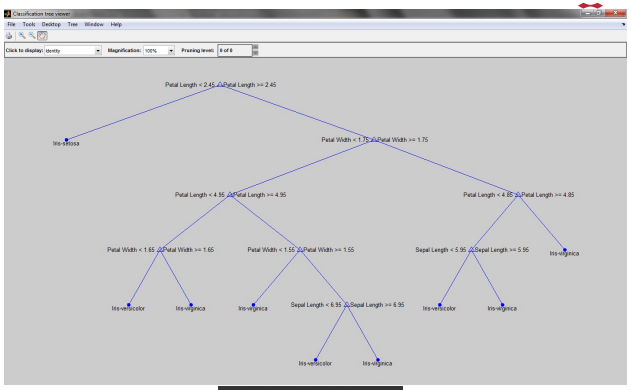


The iris data set

- **Three flowers**
  - 50 instances of each class, 150 in total
- **Attributes**
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

Fower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	**3.2**	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	**5.9**	**3.0**	**5.1**	**1.8**

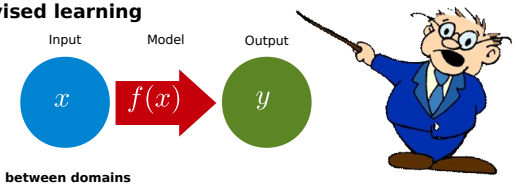
$\boldsymbol{X}^{\text{Observation x Attribute}}$



What would the following iris flower be classified as?

Sepal Length	Sepal Width	Petal Length	Petal Width
4.9	3.5	3.0	2.0

Supervised learning



- **Mapping between domains**
  - Classification: Discrete (nominal) output
  - Regression: Continuous output

Supervised learning

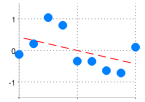
- **Data**
  - Inputs and outputs $\{x_n, y_n\}_{n=1}^{N}$
- **Model**
  - Function that maps inputs to outputs $f(x)$
- **Cost function**
  - Dissimilarity measure between data and model $d\big(y, f(x)\big)$

Regression

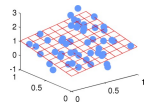
- **Definition:** Learning a function that maps a data object to a continuous-valued output
- **Why Regression?**
  - Descriptive modeling
    - Explain / understand the relation between attributes and continous-valued output
  - Predictive modeling
    - Predict the output value of a new data object

Linear regression

- 1-dimensional inputs $f(x) = w_0 + w_1 x$



- 2-dimensional inputs $f(x) = w_0 + w_1 x_1 + w_2 x_2$



- K-dimensional inputs $f(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$

Linear regression

- K-dimensional inputs

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

- Non-linearly transformed inputs

$$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_K x^K$$

$$f(x) = w_0 + w_1 sin(x) + w_2 cos(x)$$

Linear regression

- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

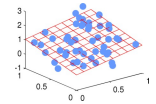
- Non-linearly transformed inputs

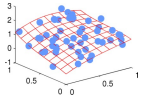
$$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_K x^K$$

$$f(x) = w_0 + w_1 sin(x) + w_2 cos(x)$$

- **Example**

$$f(x) = w_0 + w_1 x_1 + w_2 x_2$$





$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + w_6 x_1^3 + w_7 x_1^2 x_2 + w_8 x_1 x_2^2 + w_9 x_2^3$$

Vector notation

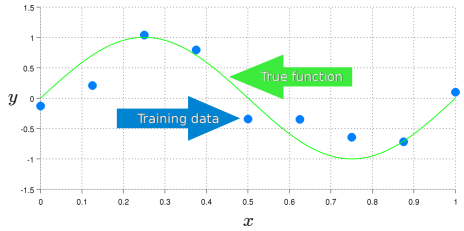
- The linear model can be written compactly using vector notation

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

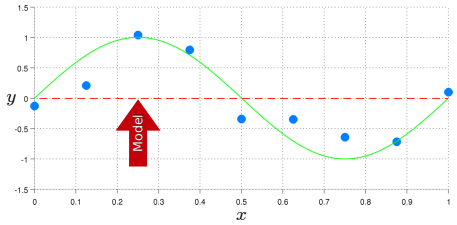
$$= \sum_{k=0}^{K} w_k x_k = \boxed{x^\top w}$$

- where $x_0 = 1$

Linear regresssion



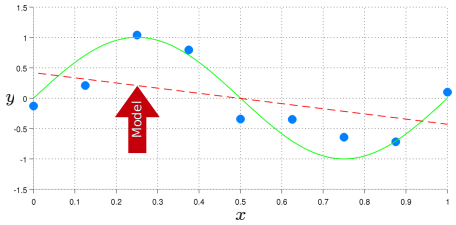
Linear regresssion



Model

$$f(x) = w_0$$

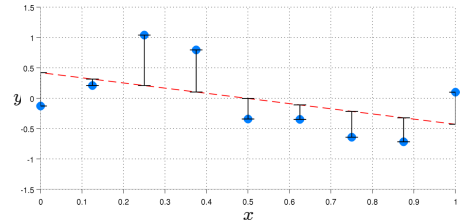
Linear regresssion



Model

$$f(x) = w_0 + w_1 x$$

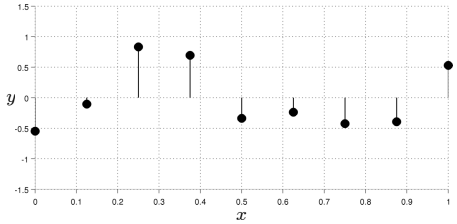
Residual error



Model

$$f(x) = w_0 + w_1 x$$

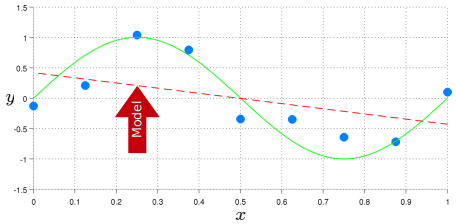
Residual error



Model

$$f(x) = w_0 + w_1 x$$

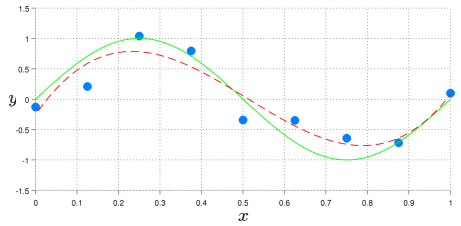
Linear regression



Model

$$f(x) = w_0 + w_1 x$$

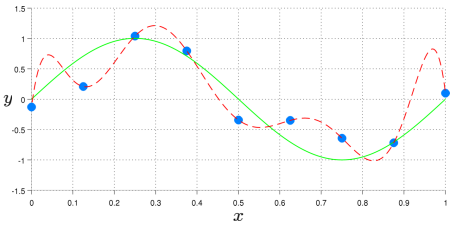
Linear regression



Model

$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

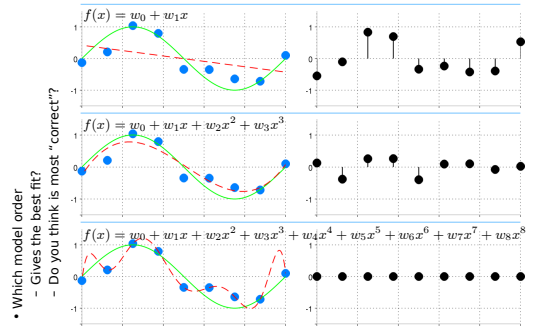
Linear regression



Model

$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6 + w_7 x^7 + w_8 x^8$$

Model order



- Which model order
  - Gives the best fit?
  - Do you think is most "correct"?

Estimating parameters

- **How do we compute the parameters?**
  - Most simple approach: **Minimize cost function over data set**
  - **Data** $\{x_n, y_n\}_{n=1}^N$
  - **Model** $f(x) = x^\top w$
  - **Cost function** $d\big(y, f(x)\big)$
  - **Parameters** $w = \arg\min_w \sum_{n=1}^N d\big(y_n, f(x_n)\big)$

Least Squares Regression

- **Cost function:** Squared error
  $d\big(y, f(x)\big) = \big(y - f(x)\big)^2$

  Model: Linear regression
  $f(x) = x^\top w$

- **Parameters**

$$w = \arg\min_w \sum_{n=1}^N d\big(y_n, f(x_n)\big) = \arg\min_w \underbrace{\sum_{n=1}^N (y_n - x_n^\top w)^2}_{E = \|y - Xw\|^2}$$

$$\frac{\partial E}{\partial w} = 2(y - Xw)^\top X = 0$$

$$\Rightarrow 2y^\top X = 2w^\top X^\top X$$

$$\Rightarrow w = (X^\top X)^{-1} X^\top y$$

Logistic Regression

(for binary classification, $y \in \{0,1\}$)

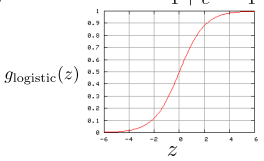
- **Cost function:**
  negative log of the Bernoulli distribution
  $d(y, f(x)) = -y \log f(x) - (1-y)\log(1 - f(x))$

  Model: Logistic link function
  $f(x) = g_{\text{logistic}}(x^\top w)$

- **Parameters**

$$w = \arg\min_w \sum_{n=1}^N d\big(y_n, f(x_n)\big)$$

$$g_{\text{logistic}}(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$



Interpretation of f(x): The probability that the observation belongs to class 1

Generalized linear model

- **Cost function:** Choose one
  $d\big(y, f(x)\big)$

  Model: Linear + non-linear link
  $f(x) = g_{\text{link}}(x^\top w)$

- **Parameters:** Optimize using numerical optimization methods

$$w = \arg\min_w \sum_{n=1}^N d\big(y_n, f(x_n)\big)$$

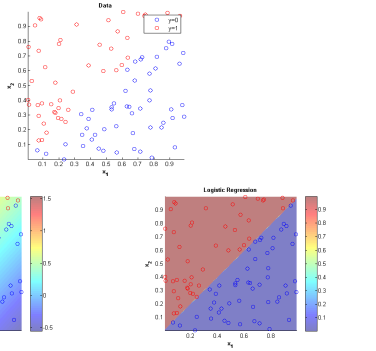
Optimize

Matlab: glmfit
Python: sklearn.linear_model
R: glm

$$g_{\text{identity}}(z) = z$$

$$g_{\text{logistic}}(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

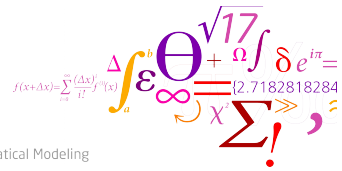
Linear vs. Logistic Regression for a classification problem



$f(x)=0.5463-1.1081x_1+1.0009x_2$

$f(x)=1/[1+\exp(-(14.5-3316x_1+3352x_2))]$

02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 4.4-4.6

Feedback Groups of the day:
Pietro Davide Doldo, Josephine Robert, Late Vitolina

Jakob Frederiksen, Rasmus Hansen

Lise Oppermann Aagesen, Sophie Nielsen, Isabel Amalia Jepsen

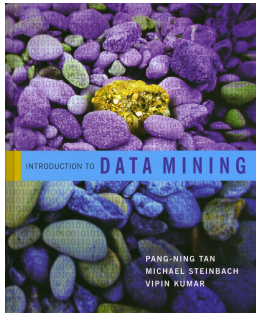
Andrew Marc, Bjarki Sigurjonsson

Rasmus Lau Petersen, Daniel Tolboe Handler

Wen Hsin Li, Robert Lyck, Viivi Halla-aho

Christian Fiksdal Brandes, Anna Emilie Henius

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. **Overfitting and performance evaluation (Tan 4.4-4.6)**
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

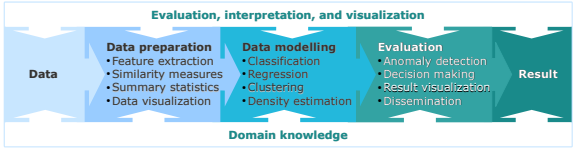
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3+8.5.7)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

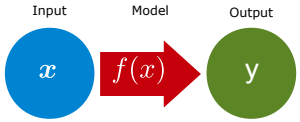
12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:
Explain the difference between training and test (generalization) error
Explain how cross-validation can be used for model selection
Apply forward and backward selection
Prune decision trees
Understand the Bias-Variance tradeoff as illustrated for regularized least squares estimation
Test the significance of classifiers

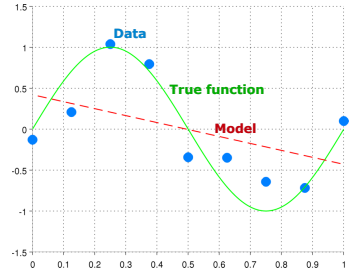
Supervised learning



- **Mapping between domains**
  - Classification: Discrete output
  - Regression: Continuous output

Linear regression

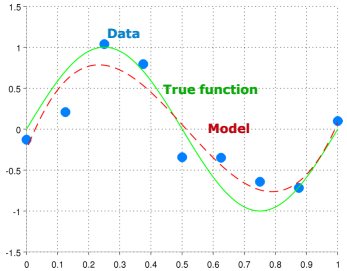
- Bad fit
- **Too simple model**



$$f(x) = w_0 + w_1 x$$

Linear regression

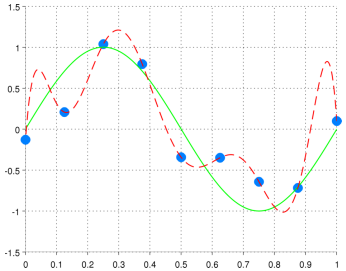
- Reasonable fit
- **Reasonable model**



$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Linear regression

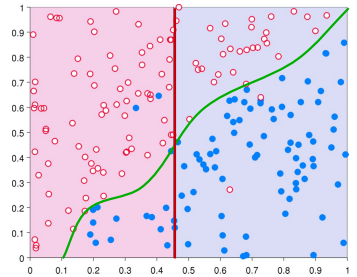
- Perfect fit
- **Too complex model**



$$f(x) = w_0 + w_1 x + \cdots + w_8 x^8$$

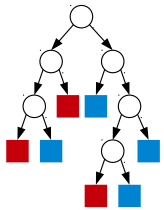
Regression trees

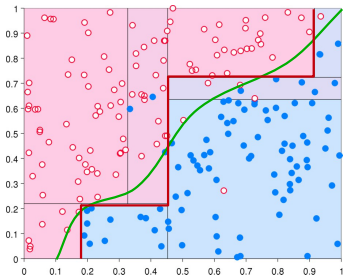
- Bad fit
- **Too simple model**



Regression trees

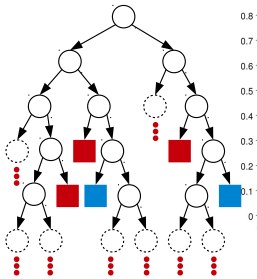
- Reasonable fit
- **Reasonable model**

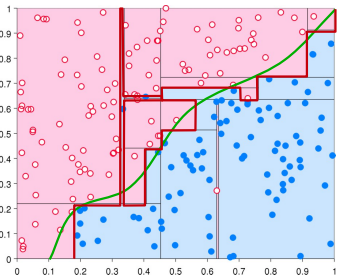




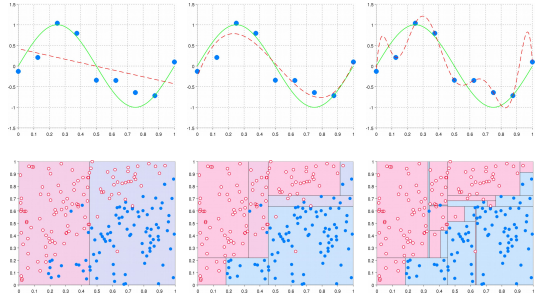
Regression trees

- Perfect fit
- **Too complex model**



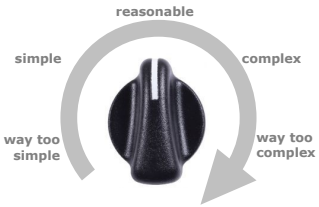


Model overfitting



Control the model complexity

- Find **parameter** or **mechanism** in model that controls complexity



Lex Parsimoniae, Law of parsimony

Given two models with same predictive performance, the simpler model is preferred over the more complex model - William of Occam

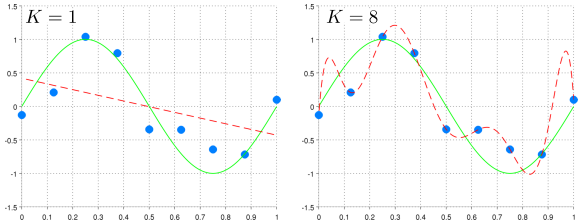
"Everything should be made as simple as possible, but not simpler" - Einstein

Linear regression

- Linear regression on non-linearly transformed inputs (polynomials)

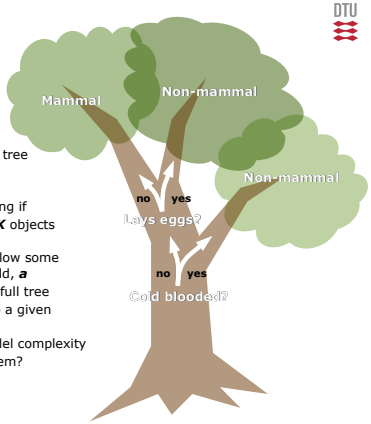
$$f(x) = w_0 + w_1 x + \cdots + w_8 x^8$$

  - **Control complexity**: Choose a suitable value for *K*



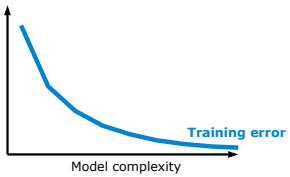
Decision trees

- Hunts algorithm
  - Continue splitting until each node is pure
  - Results in a very complex tree (overfitting)
- **Control complexity**
  - **Pre-pruning**: Stop splitting if
    - There is less than ***K*** objects on the branch
    - Impurity gain is below some predefined threshold, ***a***
  - **Post-pruning**: Generate full tree
    - Cut off branches to a given pruning level, ***c***
- ***K***, ***a***, and/or ***c*** determine model complexity
  - How should we choose them?



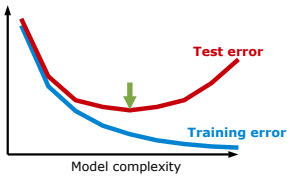
Cross-validation

- Partition the data into
  - **Training** and **test** set
- Using the **training set**
  - Train the model for a range of different complexities
    - From too simple to too complex
  - Compute the **training error**
    - It will always be lowest for the most complex model



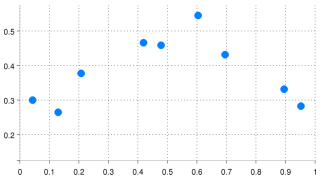
Cross-validation

- Using the **test set**
  - For each model complexity, compute the **test error**
  - Choose the model complexity that gives the lowest test error
    (The **generalization error** is the expected value of the test error.)



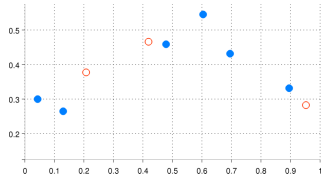
Holdout method

- Randomly choose a subset of data points to be in a **test set**
  - For example choose 1/3 of the points
- The rest is the **training set**



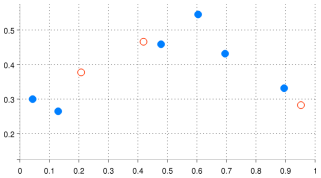
Holdout method

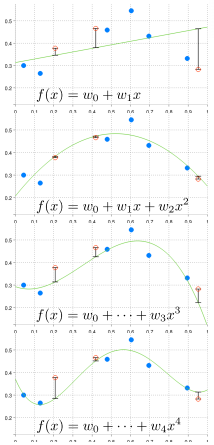
- Randomly choose a subset of data point to be in a **test set**
  - For example choose 1/3 of the points
- The rest is the **training set**



Holdout method

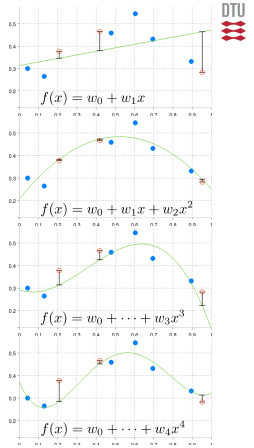
- Using the **training set**
  - Train the model for different complexities
- Using the **test set**
  - Compute the test error
- Choose the model with lowest **test error**





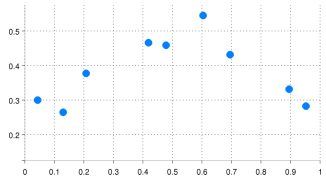
Holdout method

- Using the **training set**
  - Train the model for different complexities
- Using the **test set**
  - Compute the test error
- Choose the model with lowest **test error**



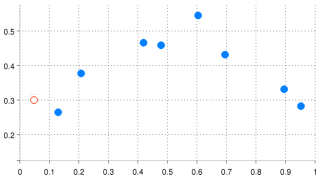
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**



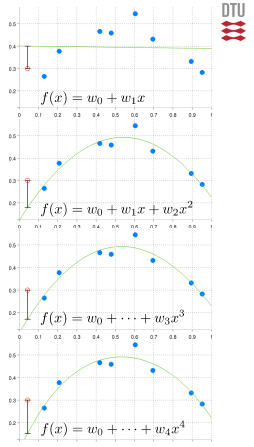
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**



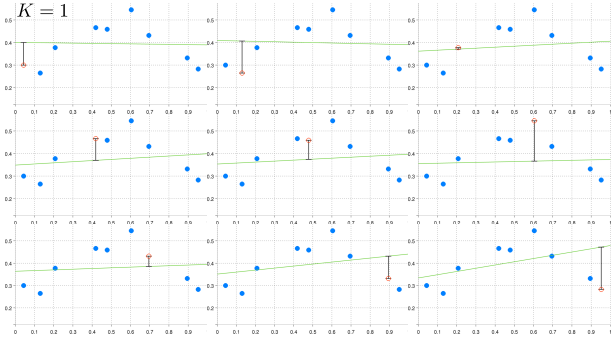
Leave-one-out

- Using the **training set**
  - Train the model for different complexities
- Using the **test set**
  - Compute the test error
- **Repeat for all data points**
  - All data points get to be test set
  - Compute **average test error**



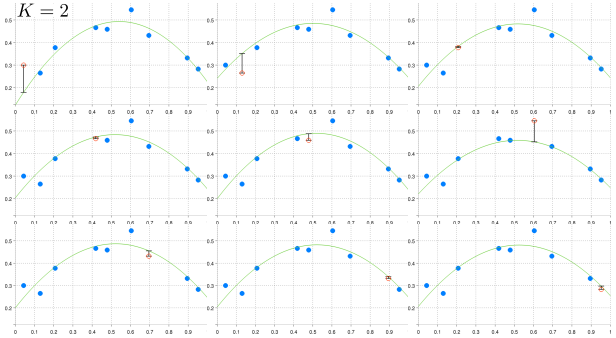
Leave-one-out

$K = 1$



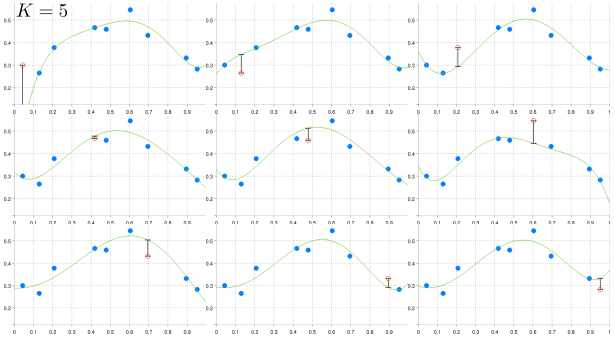
Leave-one-out

$K = 2$



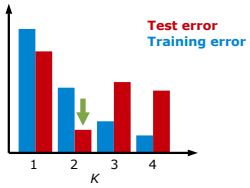
Leave-one-out cross-validation

$K = 5$

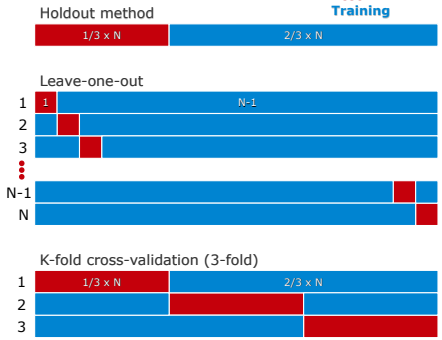


Leave-one-out

- Using the **training set**
  - Train the model for different complexities
- Using the **test set**
  - Compute the test error
- **Repeat for all data points**
  - All data points get to be test set
  - Compute **average test error**

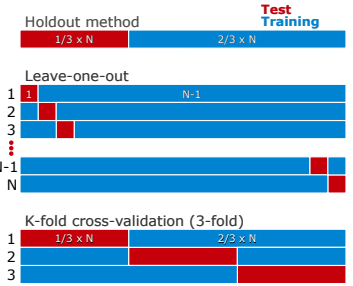


Cross-validation methods

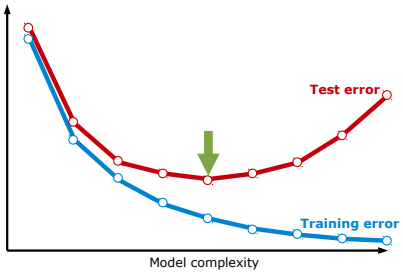


Cross-validation methods

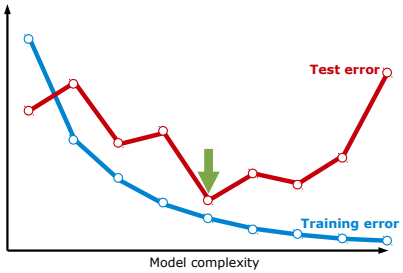
- Compare these three methods
  - What are their pros and cons?
- 10-fold cross-validation is very often used in pratice
  - Why do you think?



Training and test error



Training and test error



Feature subset selection

- Let's say we want to do linear regression
  - We have a large number of attributes

$$x_1, x_2, \ldots, x_M$$

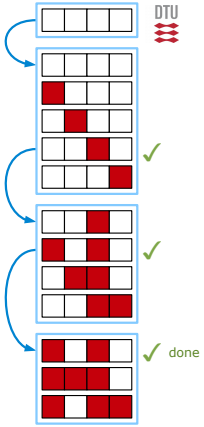
- Using all attributes results in a too complex model
  - **Control complexity**: Choose a subset of attributes
    - Small subset = Simple model
    - Large subset = Complex model

- **How many different ways can we choose a subset?**
  - How many models must be compared for
    - M=4
    - M=10
    - M=100

$$f(x) = w_0$$
$$f(x) = w_0 + w_1 x_1 + w_2 x_{27} + w_3 x_{88}$$
$$f(x) = w_0 + w_1 x_{19} + w_2 x_{76}$$
$$f(x) = w_0 + w_1 x_{19} + w_2 x_{76} + w_3 x_{88}$$
$$f(x) = w_0 + w_1 x_1 + w_2 x_{27} + w_3 x_{19}$$
$$f(x) = w_0 + w_1 x_{27} + w_2 x_{88}$$

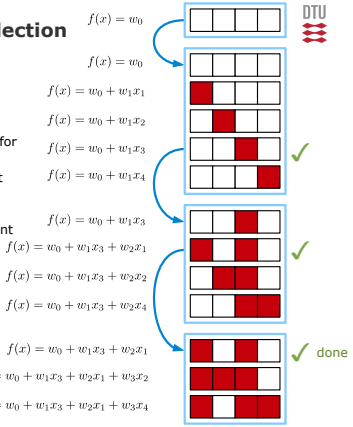
Sequential feature selection

Forward selection
- Start with no features
- Compute **cross-validation error** for
  - Current feature subset
  - All subsets equal to the current + one added feature
- Choose best subset
- Repeat until no further improvement



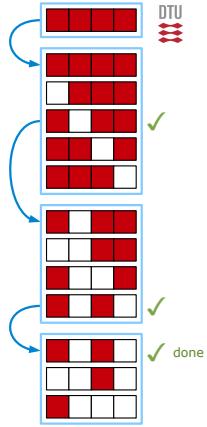
Sequential feature selection

Forward selection
- Start with no features
- Compute **cross-validation error** for
  - Current feature subset
  - All subsets equal to the current + one added feature
- Choose best subset
- Repeat until no further improvement



Sequential feature selection

Backward selection
- Start with all features
- Compute **cross-validation error** for
  - Current feature subset
  - All subsets equal to the current - one removed feature
- Choose best subset
- Repeat until no further improvement



Feature subset selection

- **How many models do we maximally have to evaluate by forward or backward selection?**

$$x_1, x_2, \ldots, x_M$$

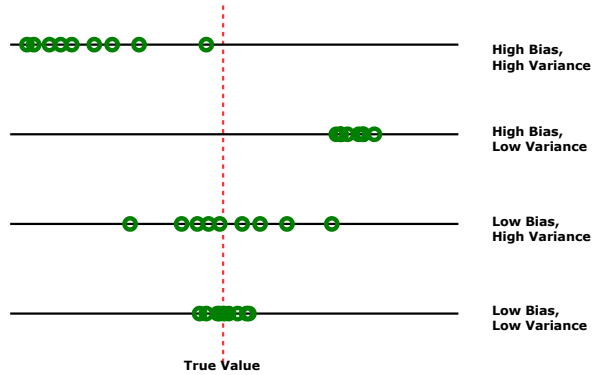
  - M=4
  - M=10
  - M=100

Bias-Variance tradeoff

Consider learning the parameter $\boldsymbol{w}$ for each of N cross-validations, denoted $\boldsymbol{w}^{(i)}$ and consider the squared error between the estimated and true parameter

$$\bar{\boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}^{(i)}$$

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{w}^{\text{true}} - \boldsymbol{w}^{(i)})^2\right] &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^{\text{true}} - \boldsymbol{w}^{(i)}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^{\text{true}} - \bar{\boldsymbol{w}} + \bar{\boldsymbol{w}} - \boldsymbol{w}^{(i)}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^{\text{true}} - \bar{\boldsymbol{w}}\right) + \frac{1}{N}\sum_{i=1}^{N}\left(\bar{\boldsymbol{w}} - \boldsymbol{w}^{(i)}\right) + \frac{1}{N}\sum_{i=1}^{N} 2\left(\boldsymbol{w}^{\text{true}} - \bar{\boldsymbol{w}}\right)\left(\bar{\boldsymbol{w}} - \boldsymbol{w}^{(i)}\right) \\
&= \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^{\text{true}} - \bar{\boldsymbol{w}}\right)}_{\text{Bias}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{\boldsymbol{w}} - \boldsymbol{w}^{(i)}\right)}_{\text{Variance}}
\end{aligned}$$



Regularized Least Squares Regression

- **Cost function:** Squared error+ridge penalty **Model:** Linear regression

$$E = \sum_{n=1}^{N}(y_n - f(\boldsymbol{x}_n))^2 + \lambda \boldsymbol{w}^T\boldsymbol{w} \qquad f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$$

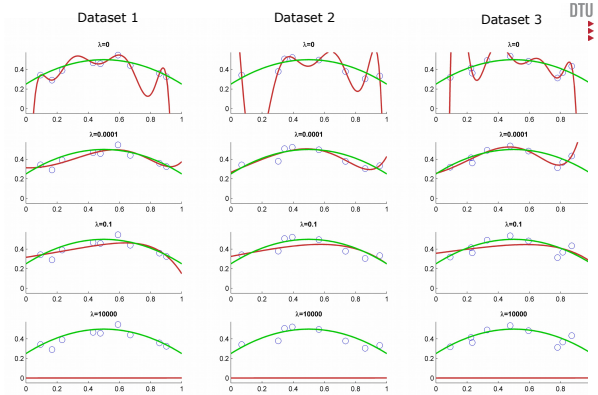
- **Parameters**

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \underbrace{\left[\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n^\top \boldsymbol{w})^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}\right]}_{E}$$

$$\begin{aligned}
\frac{\partial E}{\partial \boldsymbol{w}} &= 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X} + 2\lambda \boldsymbol{w}^\top = 0 \\
\Rightarrow 2\boldsymbol{y}^\top \boldsymbol{X} &= 2\boldsymbol{w}^\top(\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}) \\
\Rightarrow \boldsymbol{w} &= (\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{y}
\end{aligned}$$

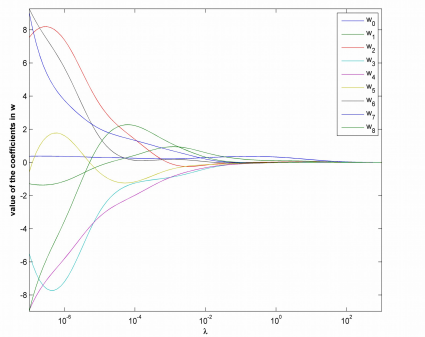
What do you think happens when $\lambda \to \infty$?
And how can we select the optimal value of λ?

$$\begin{aligned}
\boldsymbol{w} &= \arg\min_{\boldsymbol{w}} \left[\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n^\top \boldsymbol{w})^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}\right] \\
\frac{\partial E}{\partial \boldsymbol{w}} &= 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X} + 2\lambda \boldsymbol{w}^\top = 0 \\
\Rightarrow 2\boldsymbol{y}^\top \boldsymbol{X} &= 2\boldsymbol{w}^\top(\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}) \\
\Rightarrow \boldsymbol{w} &= (\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{y}
\end{aligned}$$



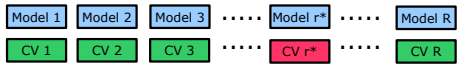
By regularization we can tradeoff bias and variance, in particular we can hope to substantially reduce variance without introducing too much bias!

$$\begin{aligned}
f(x) &= w_0 + w_1x^1 + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + w_8x^8 \\
&= \sum_{k=1}^{8} w_k x^k
\end{aligned}$$

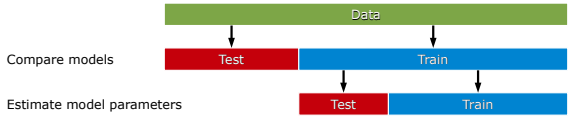


Imagine we estimate R different models and select the model r* with the lowest cross-validation error as the best model. Will the estimated cross-validation error for this model be a correct estimate of how well the model generalizes to new data?

(i.e., is the obtained cross-validation error an un-biased estimator of the generalization error?)



Multi-level cross-validation



Evaluating the performance of classifiers

- Compare classifier performance to random guessing
  - i.e., evaluate how significantly the classifier performs relative to random guessing
- Derive confidence interval for accuracy of classifiers
  - i.e., confidence intervals are used to indicate the reliability of an estimate.
- Compare the performance of two classifiers
  - i.e., is one classifier significantly better than another classifier.

Evaluating the significance of a classifier

Imagine we train a classifier on a balanced two class problem with N observations and correctly classify L of the observations. We want to know if our classifier is better than random guessing. I.e. We would like to reject the null hypothesis:

H_0: Our classifier is guessing at random

If we were random guessing the distribution of classifying L observations correctly would follow a binomial with p=½.

$$p(L|N,p) = \frac{N!}{L!(N-L)!}p^L(1-p)^{N-L}$$

The probability by random of classifying L or more observations correctly by random is then given by

$$p(l \geq L|N,p) = \sum_{l=L}^{N} \frac{N!}{l!(N-l)!}p^l(1-p)^{N-l}$$

$$= 1 - \sum_{l=0}^{L-1} \frac{N!}{l!(N-l)!}p^l(1-p)^{N-l}$$

Consider two classification problems.

Problem 1: N=10, L=8

Problem 2: N=100, L=60

```
p(l≥8|10,0.5)=0.0547
p(l≥60|100,0.5)=0.0284
```


Confidence interval for the accuracy of a classifier

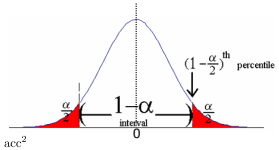
The empirical accuracy has mean p=L/N and variance p(1-p)/N. For large N we can approximate the binomial distributed by the normal distribution to obtain confidence intervals for the accuracy, i.e.

$$p\left(Z_{\alpha/2} \leq \frac{\text{acc}-p}{\sqrt{p(1-p)/N}} \leq Z_{1-\alpha/2}\right) = 1-\alpha$$

$$Z_{1-\alpha/2} = -Z_{\alpha/2}$$

$$p\left(Z_{\alpha/2} \leq \frac{\text{acc}-p}{\sqrt{p(1-p)/N}} \leq Z_{1-\alpha/2}\right) = p\left(\left[\frac{\text{acc}-p}{\sqrt{p(1-p)/N}}\right]^2 \leq Z^2_{\alpha/2}\right)$$

$$\left[\frac{\text{acc}-p}{\sqrt{p(1-p)/N}}\right]^2 = Z^2_{\alpha/2}$$



$$0 = (N + Z^2_{\alpha/2})p^2 - (2N\ \text{acc} + Z^2_{\alpha/2})p + N\ \text{acc}^2$$

$$p \in [\mu - d; \mu + d]$$

$$\mu = \frac{2N\ \text{acc} + Z^2_{\alpha/2}}{2(N + Z^2_{\alpha/2})}, \qquad d = \frac{Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + 4N\ \text{acc} - 4N\ \text{acc}^2}}{2(N + Z^2_{\alpha/2})}$$

Consider a model with accuracy 80%, the confidence interval as a function of number of observations N are

N	20	50	100	500	1000	5000
p∈	[0.584,0.919]	[0.670;0.888]	[0.711;0.867]	[0.763;0.833]	[0.774;0.824]	[0.789;0.811]

Comparing the performance of two classifiers based on k-fold cross-validation using the same splits for the two models

Let e_{1j} and e_{2j} be the error rates at cross validation split j for model 1 and model 2 respectively and $d_j = e_{1j} - e_{2j}$

$$\bar{d} = \frac{1}{k}\sum_{j=1}^{k} d_j$$

Variance of the mean value

$$\bar{\sigma}^2 = \frac{\sum_{j=1}^{k}(d_j - \bar{d}_j)^2}{k(k-1)}$$

$$d^{\text{cv}} \in [\bar{d} - t_{1-\alpha/2,k-1}\bar{\sigma}; \bar{d} + t_{1-\alpha/2,k-1}\bar{\sigma}]$$

Consider two classifiers with the following error rates across 5-cross validation splits.

Classifier 1: $\boldsymbol{e}_1$=[0.2 0.1 0.3 0.2 0.4]

Classifier 2: $\boldsymbol{e}_2$=[0.1 0.1 0.2 0.2 0.1]

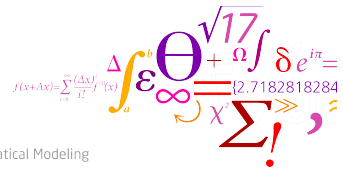
$\boldsymbol{d} = \boldsymbol{e}_1 - \boldsymbol{e}_2$=[0.1 0 0.1 0 0.3]

$$\bar{d} = 0.1, \quad \hat{\sigma}^2 = 0.003,$$

$$d^{cv} \in [0.1 - 2.7764 \cdot \sqrt{0.003}; 0.1 + 2.7764 \cdot \sqrt{0.003}]$$

$$= [-0.0521; 0.2521]$$

02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 5.2-5.4

Feedback Groups of the day:

Mathias Gæde, Matthias S. Alan Larsen, Frederik Kirk
Søren Howe Gersager, Ahmet Yildirim,

Christopher Frederick Wiin Schenk

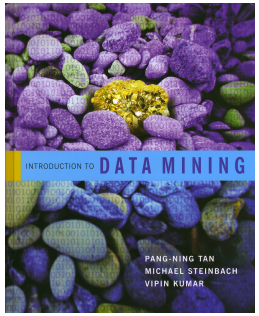
Loke Kristensen, Michael Schmidt Nielsen

Jamie Neubert Pedersen, Ulf Aslak Jensen, Daniel Heestermans Svendsen

Thomas Kristensen, Frederik Wedel-Heinen

Karol Dzitkowski, Marco Becattini
Anders Rahbek

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. **Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)**
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

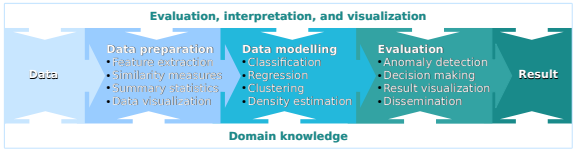
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

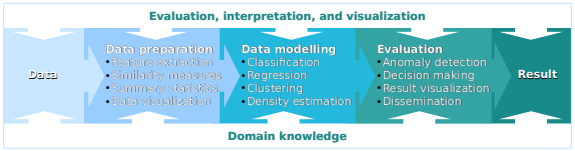
12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:
Explain how K-Nearest Neighbors can be used to classify data
Account for the assumptions made in Naïve Bayes
Apply Bayes theorem to obtain the class posterior likelihood
Understand the principle behind artificial neural networks (ANN) and how ANN can be used for classification and regression.

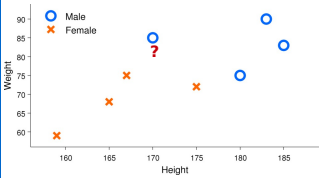
Data modeling framework



Feedback on projects:
To find out which teacher has given feedback to your project, download *"instructions for group project 1,2 and 3/02450_project_1_2014fall_students.xlsx"* from campusnet and find the teacher who has graded your project after the lecture.
Please ensure your student ID is on the list to ensure you will get credit. Otherwise contact the relevant exercise teacher or me.

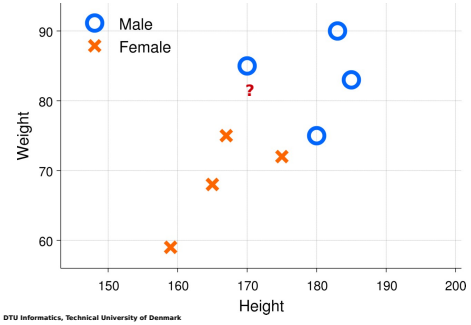
Classify gender based on height and weight

	Height	Weight	Gender
1	183	90	Male
2	180	75	Male
3	170	85	Male
4	185	83	Male
5	159	59	Female
6	167	75	Female
7	165	68	Female
8	175	72	Female
9	171	82	?



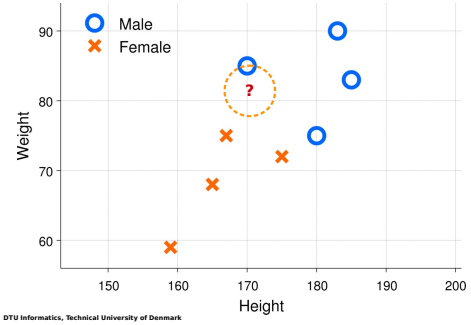
Nearest neighbor classifier

• 1 nearest neighbor



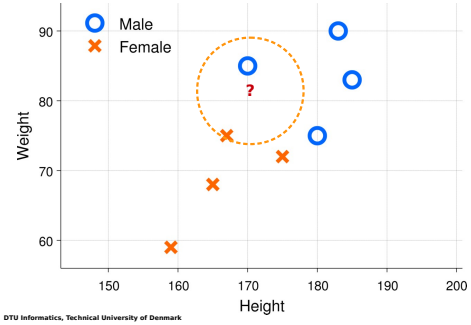
Nearest neighbor classifier

• 1 nearest neighbor



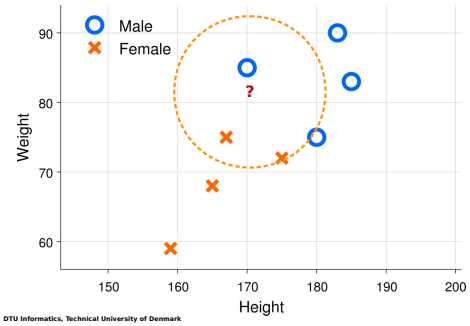
Nearest neighbor classifier

• 2 nearest neighbors



Nearest neighbor classifier

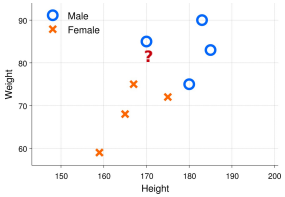
- 3 nearest neighbors



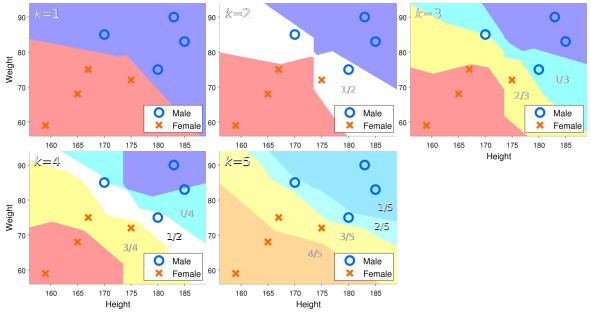
Nearest neighbor classifier

- Choose
  - The number of neighbors, *k*
  - A distance measure

1. Compute distance to all other data objects
2. Find the *k* nearest data objects
3. Classify according to majority of neighbors



Nearest neighbor decision surface



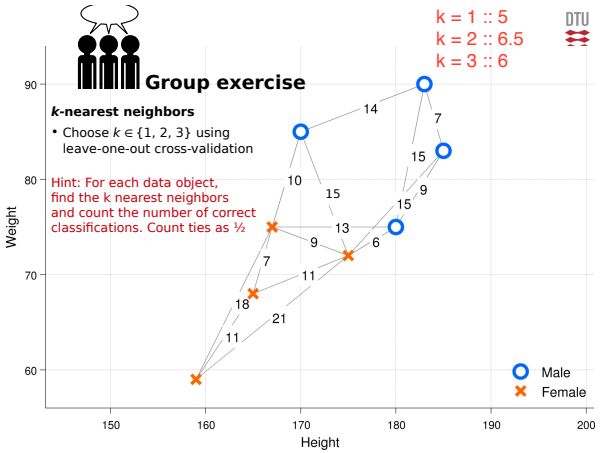
Group exercise

k = 1 :: 5
k = 2 :: 6.5
k = 3 :: 6

***k*-nearest neighbors**

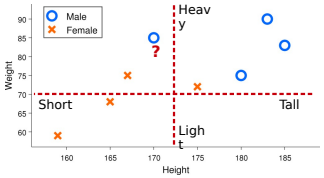
- Choose $k \in \{1, 2, 3\}$ using leave-one-out cross-validation

Hint: For each data object, find the k nearest neighbors and count the number of correct classifications. Count ties as ½



Bayesian classifiers

	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female
9	Short	Heavy	Light



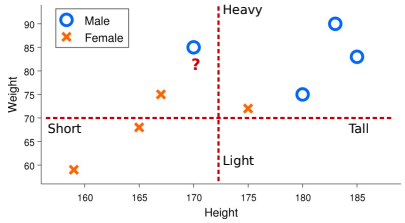
Bayesian classifiers

- What is the probability that **?** is male

$$p(\text{Gender} = \text{Male} | \text{Height} = \text{Short}, \text{Weight} = \text{Heavy})$$

- Shorthand notation:

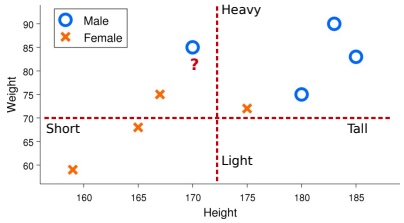
$$p(G = m | H = s, W = h) = p(m | s, h)$$



Bayesian classifiers

- Bayes rule

$$p(m|s,h) = \frac{p(s,h|m)p(m)}{\sum_{G \in \{m,f\}} p(s,h|G)p(G)} = \frac{\frac{1}{4} \cdot \frac{4}{8}}{\frac{1}{4} \cdot \frac{4}{8} + \frac{1}{4} \cdot \frac{4}{8}} = \frac{1}{2}$$



Bayesian classifiers

- **Contingency table**
  - All combinations of attribute values
  - Huge table

	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female

Gender	Height	Weight	Fraction
Male	Short	Light	0/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	3/4
Female	Short	Light	2/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	1/4

Bayesian classifiers

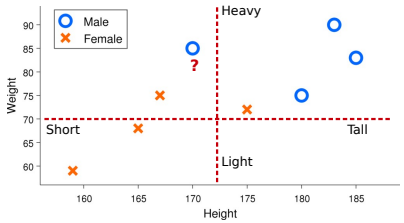
- Naïve Bayes assumption
  - Conditional probabilities of attributes are independent

$$p(\text{Height}, \text{Weight} | \text{Gender}) = p(\text{Height} | \text{Gender}) \times p(\text{Weight} | \text{Gender})$$

Bayesian classifiers

- **Naïve Bayes classifier**

$$p(m|s,h) = \frac{p(s|m)p(h|m)p(m)}{\sum_{G\in\{m,f\}} p(s|G)p(h|G)p(G)} = \frac{\frac{1}{4}\cdot\frac{4}{4}\cdot\frac{4}{8}}{\frac{1}{4}\cdot\frac{4}{4}\cdot\frac{4}{8}+\frac{3}{4}\cdot\frac{2}{4}\cdot\frac{4}{8}} = \frac{2}{5}$$

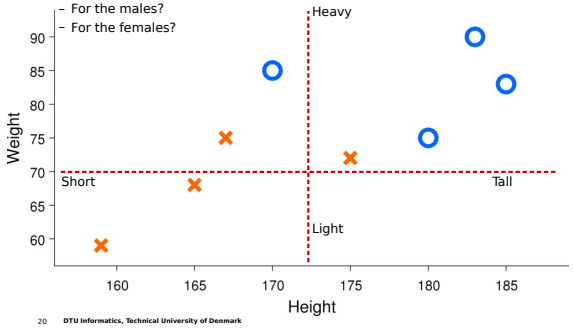


Naïve Bayes assumption

$$p(\text{Height}, \text{Weight}|\text{Gender}) = p(\text{Height}|\text{Gender}) \times p(\text{Weight}|\text{Gender})$$

Group exercise

- Does the naïve Bayes assumption hold empirically
  - For the males?
  - For the females?



Bayesian classifiers

- **Naïve Bayes contingency table**
  - Only counts for each attribute
  - Small table

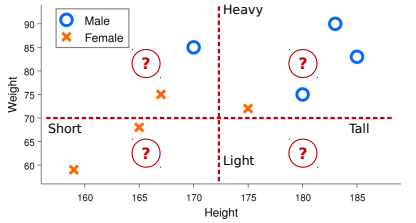
	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female

Gender	Attribute	Fraction
Male	Height=Short	1/4
	Weight=Light	0/4
Female	Height=Short	3/4
	Weight=Light	2/4

Group exercise

Bayes classifiers

- Classify (compute the posterior probability of G=m) for the four **?** using
  - Bayes classifier $p(m|s,h) = \frac{p(s,h|m)p(m)}{\sum_{G\in\{m,f\}} p(s,h|G)p(G)} = \frac{\frac{1}{4}\cdot\frac{4}{8}}{\frac{1}{4}\cdot\frac{4}{8}+\frac{1}{4}\cdot\frac{4}{8}} = \frac{1}{2}$
  - Naïve Bayes classifier $p(m|s,h) = \frac{p(s|m)p(h|m)p(m)}{\sum_{G\in\{m,f\}} p(s|G)p(h|G)p(G)} = \frac{\frac{1}{4}\cdot\frac{4}{8}\cdot\frac{4}{8}}{\frac{1}{4}\cdot\frac{4}{8}\cdot\frac{4}{8}+\frac{3}{4}\cdot\frac{2}{4}\cdot\frac{4}{8}} = \frac{2}{5}$



Gender	Attribute	Fraction
Male	Height = Short	1/4
	Weight = Light	0/4
Female	Height = Short	3/4
	Weight = Light	2/4

Gender	Height	Weight	Fraction
Male	Short	Light	0/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	3/4
Female	Short	Light	2/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	1/4

M-estimate

- Probability of class given attribute

$$p(y|x) = \frac{n_c}{n}$$

n_c: Number of objects in class y that take value x

n: Total number of objects in class y

  - Not defined when n=0

- M-estimate

$$p(y|x) = \frac{n_c + mp}{n + m}$$

p: Prior probability

m: Equivalent sample size

  - If no objects in class y take value x the probability will be p
  - Corresponds to putting m extra objects into the data set

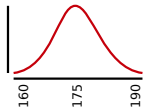
Bayesian classifiers

$p(\text{Height}|\text{Gender} = \text{Male})$

- Handling continuous attributes
  - Two way split (x<a)
    - Converts into binary attribute (We have used this in the previous example)

  - Discretize into a number of bins
    - Converts into discrete ordinal attribute

  - Probability density estimation
    - Assume attribute follows a Normal distribution
    - Use data to compute parameters (mean and variance)

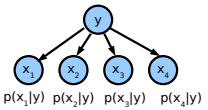


Baysian Belief Networks (BBN)

- Independence assumption may not hold for some attributes (use BBN)

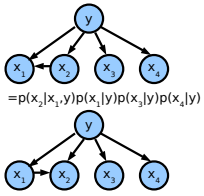
Naïve Bayes

$p(\mathbf{X}|y) = p(x_1|y)p(x_2|y)p(x_3|y)p(x_4|y)$



When x_1 and x_2 are not independent given y

$p(\mathbf{X}|y) = p(x_1,x_2|y)p(x_3|y)p(x_4|y)$
$= p(x_1|x_2,y)p(x_2|y)p(x_3|y)p(x_4|y)$



Remember basic rules of probability

- Sum rule

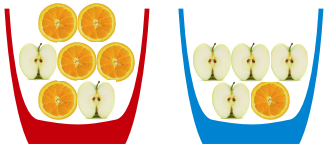
$$p(x) = \sum_y p(x,y)$$

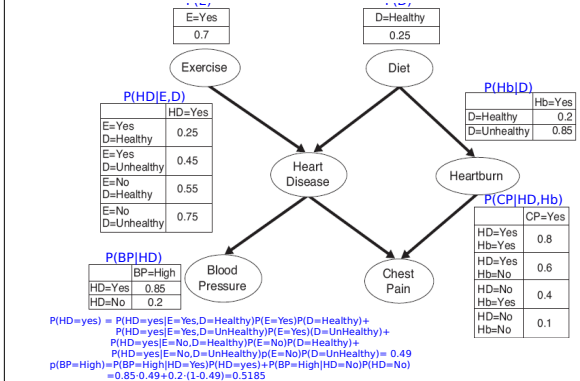
- Product rule

$$p(x,y) = p(x|y)p(y)$$

- Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$





P(HD=yes) = P(HD=yes|E=Yes,D=Healthy)P(E=Yes)P(D=Healthy)+
P(HD=yes|E=Yes,D=UnHealthy)P(E=Yes)(D=UnHealthy)+
P(HD=yes|E=No,D=Healthy)P(E=No)P(D=Healthy)+
P(HD=yes|E=No,D=UnHealthy)p(E=No)P(D=UnHealthy)= 0.49
p(BP=High)=P(BP=High|HD=Yes)P(HD=yes)+P(BP=High|HD=No)P(HD=No)
=0.85·0.49+0.2·(1-0.49)=0.5185

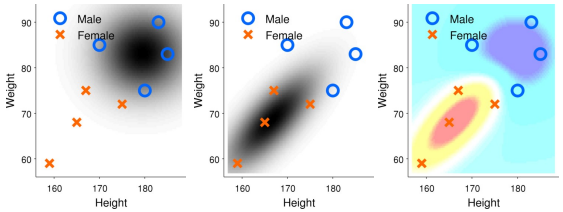
P(HD=yes|BP=High) =P(BP=High|HD=Yes)P(HD=Yes)/P(BP=High)
=0.85·0.49/0.5185=0.8033
P(HD=yes|BP=High,D=Healthy, E=yes) =P(BP=High|HD=Yes)P(HD=Yes|D=Healthy, E=yes)/P(BP=High|D=Healthy, E=yes)
=0.85·0.25/(0.85·0.25+0.2·(1-0.25))= 0.5862

Bayesian classification by the multivariate normal distribution

Continuous density estimation $P(\boldsymbol{x}|y=c) = \frac{1}{(2\pi)^{M/2} det(\boldsymbol{\Sigma}_c)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c (\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$

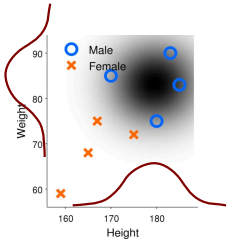
- Fit a Normal distribution to each class
  - Compute class mean and covariance
- Classify using Bayes rule as before

$$P(y=c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=c)P(y=c)}{\sum_{c'} P(\boldsymbol{x}|y=c')P(y=c')}$$



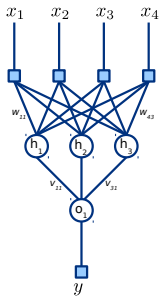
Group exercise

- What does the Naive Bayes assumption of independence of the attributes correspond to in terms of the parameters of the multivariate normal distribution?



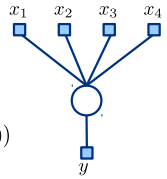
Artificial neural networks (ANN)

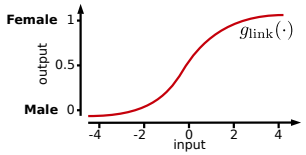
- The human brain contain in the order of 10^{11} neurons with 10^{15} connections
- Artificial Neural Networks (McCulloch & Pitts, 1943) are inspired by the architecture of the human brain



Artificial neural networks (ANN)

- **Remember the generalized linear model?**
  - Data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$
  - Model $f(\boldsymbol{x}) = g_{\text{link}}(\boldsymbol{x}^\top \boldsymbol{w})$
  - Cost function $d\big(y, f(\boldsymbol{x})\big)$
  - Parameters $\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^N d\big(y_n, f(\boldsymbol{x}_n)\big)$

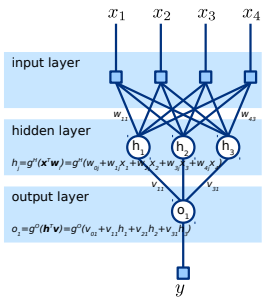




Artificial neural networks

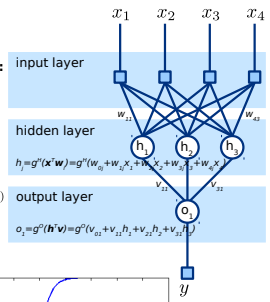
Feed forward network

- Each "neuron"
  - Computes a non-linear function of the sum of its inputs
  - Is just like a generalized linear model
  - Has its own set of parameters
- Modeling choices
  - Cost function
  - Non-linearities
  - Number of neurons and hidden layers
  - Selection of inputs
- Parameter estimation using numerical optimization methods
- Very flexible model: Can easily overfit



Artificial Neural Networks

- **The ANN we will consider in the exercises:**
  - Data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$
  - Model $f(\boldsymbol{x}_n) = g^O(v_{10} + \sum_i g^H(\boldsymbol{x}^\top \boldsymbol{w}_i))$
  - Cost function $d\big(y, f(\boldsymbol{x})\big)$
  - Parameters $\boldsymbol{W}, \boldsymbol{v} = \arg\min_{\boldsymbol{W},\boldsymbol{v}} \sum_{n=1}^N d(y_n, f(\boldsymbol{x}_n))$

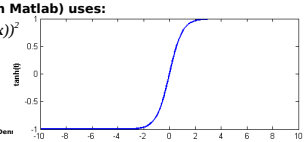


- **The implementation (in Matlab) uses:**

$d(y,f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$

$g^O(t) = t$

$g^H(t) = tanh(t)$

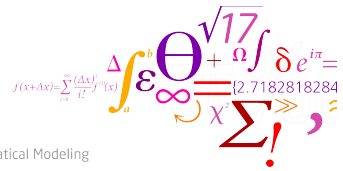


Midterm practice test

The midterm practice test is used solely for you to test your knowledge and for me to see how well you have understood the covered material so far.

The test **does not** count towards your grade for this course.

02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Question 5

Consider a data set of four features: A, B, C, and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.
We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

- C
- A and B
- A and B and C
- B and C and D
- Don't Know

Feature(s)	Error rate
A	0.40
B	0.45
C	0.33
D	0.42
A and B	0.20
A and C	0.25
A and D	0.34
B and C	0.29
B and D	0.42
C and D	0.40
A and B and C	0.13
A and B and D	0.17
B and C and D	0.10
A and C and D	0.15
A and B and C and D	0.28

Midterm test: Question 6

When training a decision tree we will use the classification error as impurity measure I(t) given by I(t) = 1-max_i[p(i|t)] where p(i|t) denotes the fraction of data objects belonging to class i at a given node t. We will use Hunt's algorithm to grow the tree and recall that the purity gain is given by:

$\Delta = I(Parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j. We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.
- After the split:
  - 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
  - 5 Iris-Setosa, 2 Iris-Viriginica and 8 Iris-Versicolor in the right node.

Which statement is correct?
1: The purity gain is 3/5.
2: The purity gain is 3/15.
3: The purity gain is 6/25.
4: The purity gain is 7/15.

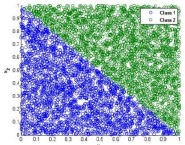
Midterm test: Question 9

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values σ_1=4, σ_2=2, σ_3=1, and σ_4=0. Which one of the following statements is wrong?

1: The first principal component accounts for more than 60% of the variation in the data.
2: The third principal component accounts for less than 5 % of the variation in the data.
3: The second principal component accounts for more than 20 % of the variation in the data.
4: The data can be perfectly represented in a three dimensional sub-space.

Midterm test: Question 12

Consider the classification problem given in Figure 5 where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0. Which one of the following statements is wrong?



1: The two classes can be perfectly separated by a logistic regression model using x1 and x2 as features.
2: A decision tree with less than five nodes can perfectly separate the classes using only x1 and x2 as features.
3: A logistic regression model can perfectly separate the two classes using only the feature t given by t = x1 + x2.
4: In logistic regression the probability that each observation belong to the two classes can be derived from the logit link function.

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 5.6-5.8

Feedback Groups of the day:

Jakob Kaufmann Jespsern

Morten Bjelbo Thomsen, Rolf s123490

Arthur Katossky, Anders Emil Nielsen

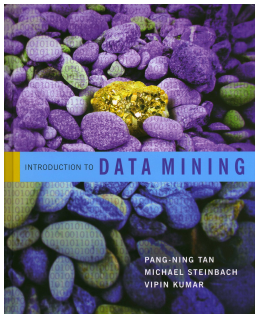
Mette Nielsen, Henrik Thistrup

Utku Norman, Darren Williams, Jose Esteves

Henry Henry, Dominik Otto, Felix Ngajieh Ekwoge

**Simon B. Hemmingsen,
Thomas Kjæller Hammerbak**

Minh Haw Truong
If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

1. Introduction
   (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
   (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics
   (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
   (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
   (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
   (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
   (Tan 5.2-5.4)
8. **Ensemble methods and multi class classifiers**
   (Tan 5.6-5.8)

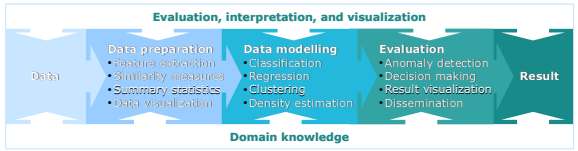
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
   (Tan 8.1-8.3)
10. Mixture models and association mining
    (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection
    (Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

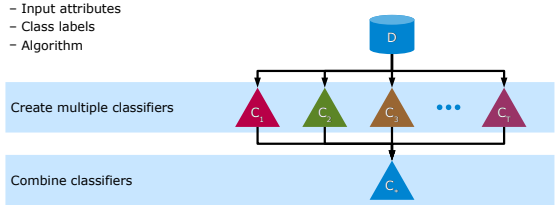
Data modeling framework



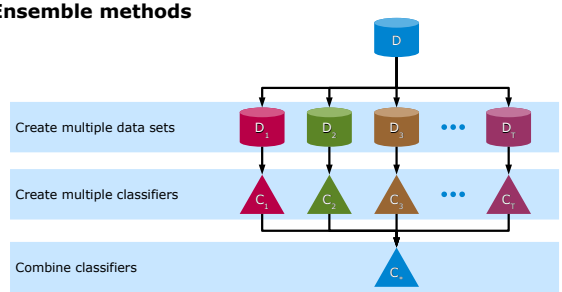
After today you should be able to:
Explain the principle behind boosting and bagging and apply it to improve classifiers
Be able to address issues of class-imbalances by resampling
Understand the definition of Precision, Recall, ROC and AUC
Be able to extend binary classifiers to multi-class classification

Ensemble methods

- Combine multiple (weak) classifiers into one (strong) classifier
- Each classifier trained using different variations of
  - Data set
  - Input attributes
  - Class labels
  - Algorithm



Ensemble methods



Why ensemble methods?

- Can improve classification algorithms in terms of
  - Better classification accuracy
  - Increased stability
  - Reduced variance
  - Less overfitting

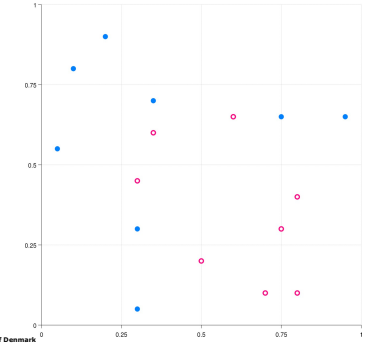
- Consider N=25 independent classifiers for binary classification, each with error rate ε=0.35

$$\epsilon_* = \sum_{n=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{n} \epsilon^n (1-\epsilon)^{N-n}$$

$$= \sum_{n=13}^{25} \binom{25}{n} 0.35^n (1-0.35)^{N-n} = 0.06$$

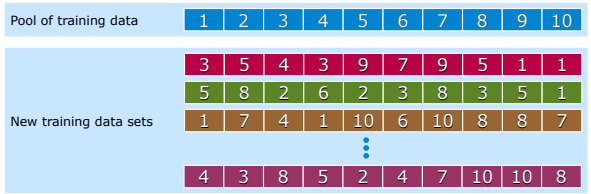
Data example

- Classification using logistic regression

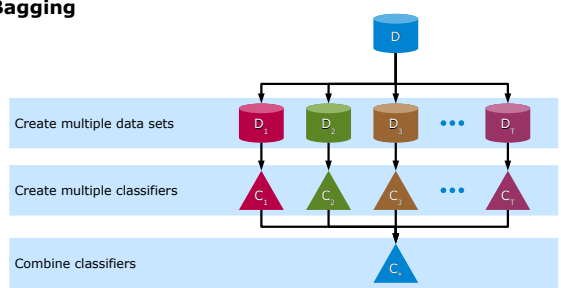


Bagging

- New training data sets drawn randomly from pool with replacement

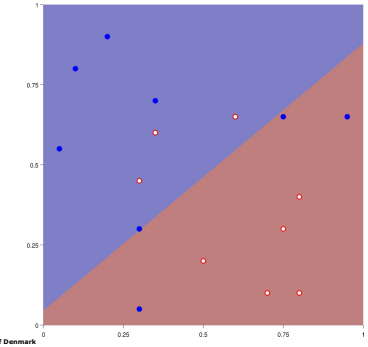


Bagging

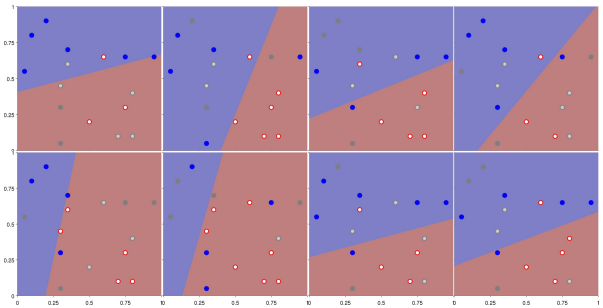


Bagging

- **Single classifier**
  - Logistic regression
  - Two features, (x,y)



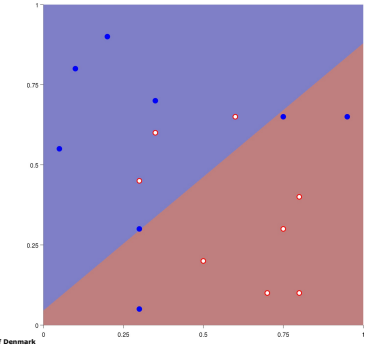
Bagging



Notice, gray dots are observations not included in bagging round

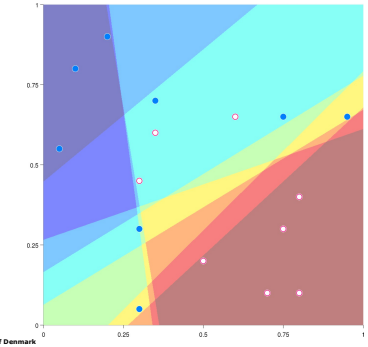
Bagging

- Single classifier



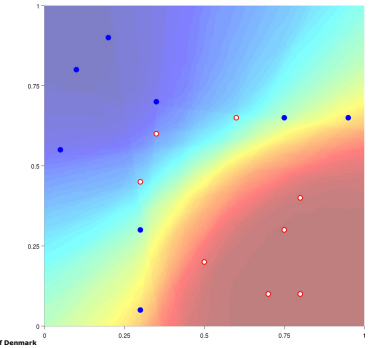
Bagging

- Majority voting of 10 bagged classifiers



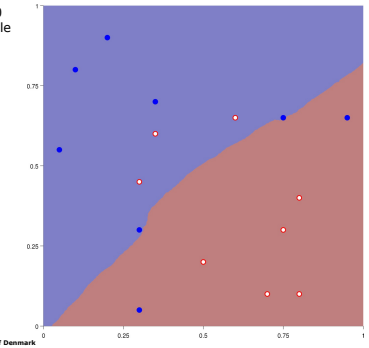
Bagging

- Majority voting of 100 bagged classifiers



Bagging

- Decision boundary of 100 bagged classifier ensemble



Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier						C_1				

Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier						C_1				
Classify all data objects	1✓	2✗	3✓	4✗	5✓	6✗	7✓	8✓	9✓	10✓

Boosting

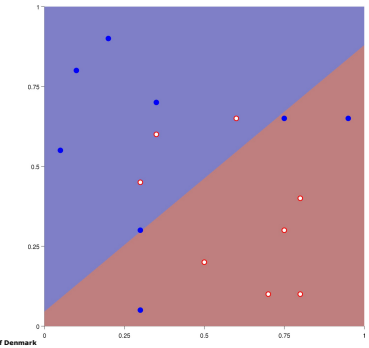
Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier						C_1				
Classify all data objects	1✓	2✗	3✓	4✗	5✓	6✗	7✓	8✓	9✓	10✓
Update weights	.07	.17	.07	.17	.07	.17	.07	.07	.07	.07

Boosting

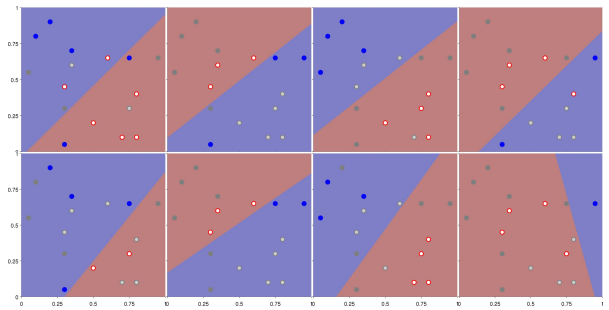
Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3	5	4	3	9	7	9	5	1	1
Train classifier						C_1				
Classify all data objects	1✓	2✗	3✓	4✗	5✓	6✗	7✓	8✓	9✓	10✓
Update weights	.07	.17	.07	.17	.07	.17	.07	.07	.07	.07
New training data set	6	4	7	3	2	4	10	2	5	6
Train classifier						C_2				

Boosting

- Single classifier



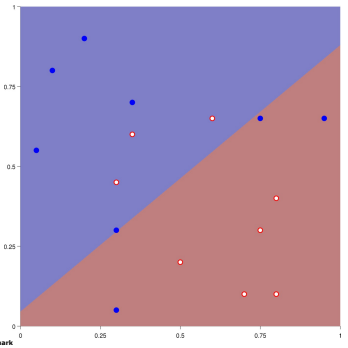
Boosting



Notice, gray dots are observations not included in boosting round

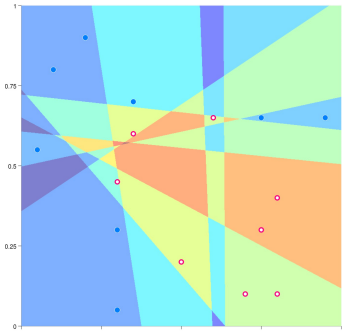
Boosting

- Single classifier



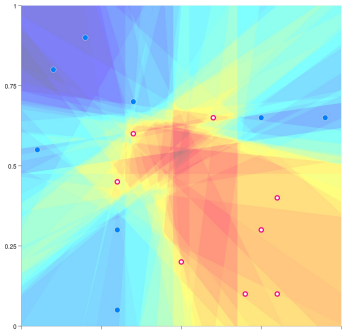
Boosting

- Majority voting of 10 boosted classifiers



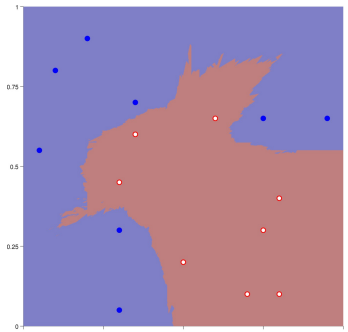
Boosting

- Majority voting of 100 boosted classifiers



Boosting

- Decision boundary of 100 boosted classifier ensemble



Class imbalance problem

- Many data sets have **imbalanced class distributions**
  - Example: Detection of defects that only occur rarely (e.g. 1/1,000,000)
  - Danger: Algorithm that says nothing is defect will be 99.999% correct
- **Solution approaches**
  - Resample to balance data sets
  - Modify existing classification algorithms
  - Measure performance in a way that takes balance into account

Resampling balanced data

- New sample has equal number of data objects from each class
- **Approaches**
  - **Undersampling** majority class: Throws out potentially useful data
  - **Oversampling** minority class: Increase data size and computational burden
  - **Somewhere in between**...

Imbalanced training data	1	2	3	4	5	6	7	8	9	10
Oversampling	1	2	3	4	5	1	2	3	6	6
	6	6	8	8	8	8				
Undersampling	3	5	6	8						
Somewhere in between	3	5	4	3	9	6	6	8	8	8

Confusion matrix

Actual \ Predicted	positive	negative
positive	TP True Positive	FN False Negative
negative	FP False Positive	TN True Negative

Precision and recall

- **Precision**
  - Fraction of true positive among objects predicted to be positive

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- **Recall**
  - Fraction of objects predicted to be positive among all positive objects

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Actual \ Predicted	positive	negative
positive	TP True Positive	FN False Negative
negative	FP False Positive	TN True Negative

Precision ⟷ Recall

Group exercise

- You consider two different classifiers, on a test set with 20 positive objects
  - **Classifier 1** detects 54 positives of which 18 are actually positive
  - **Classifier 2** detects 16 positives of which 14 are actually positive
- Compute the **precision** and **recall** for the two classifiers
- Which classifier (if any) is the best?
- Which would you use if the objective is to detect credit card fraud
  *(consider what is most costly – **missing** or **falsely detecting** a positive)*

- **Precision**
  - Fraction of true positive among objects predicted to be positive

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

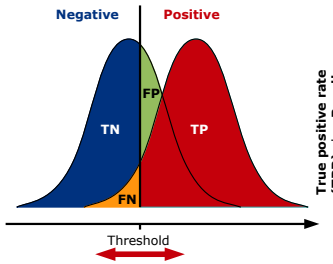
- **Recall**
  - Fraction of objects predicted to be positive among all positive objects

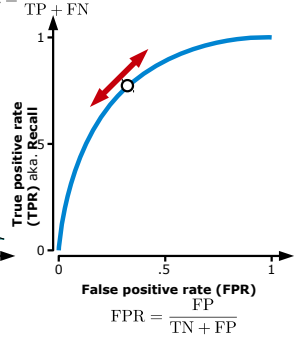
$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Actual \ Predicted	positive	negative
positive	TP True Positive	FN False Negative
negative	FP False Positive	TN True Negative

Receiver operating characteristic

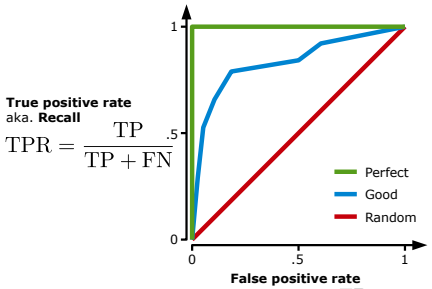
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$





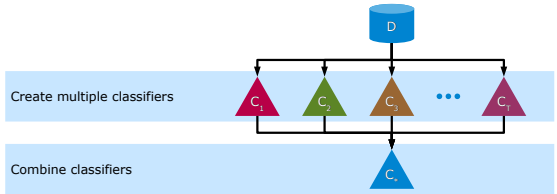
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

Receiver operating characteristic



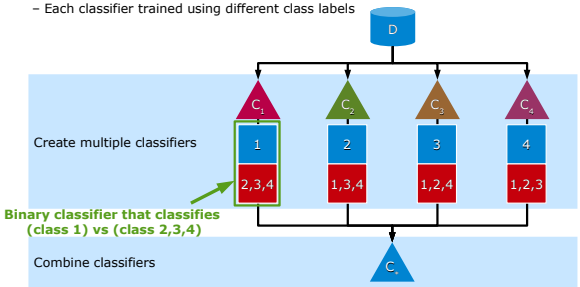
Multiclass problems

- Combine multiple **binary** classifiers into one **multiclass** classifier
  - Each classifier trained using different class labels



Multiclass problems

- Combine multiple **binary** classifiers into one **multiclass** classifier
  - Each classifier trained using different class labels

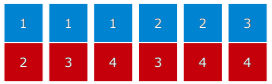


Multiclass problems

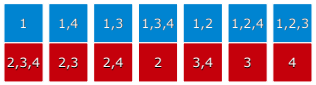
- Classification algorithms designed for **binary classification** can also be applied to **multiclass** problems
  - Train a number of classifiers and make final classification by **majority voting**

- **1-against-rest**
  - K classifiers

- **1-against-1**
  - K·(K-1)/2 classifiers



- **Error correcting output coding**
  - Robustness against errors



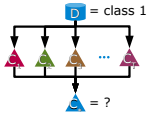
Group exercise

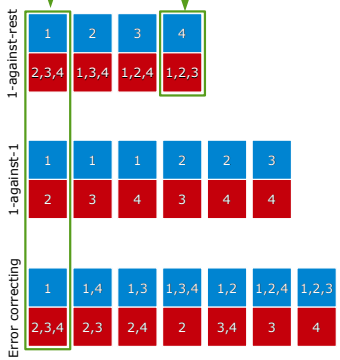
Possibly broken classifier

Binary classifier that classifies (class 4) vs (class 1,2,3)

- Assuming that **all classifiers are perfect**
  - Use the three multiclass methods to classify a data object belonging to class 1

  Hint: Go through the classifiers and add up the votes, giving one vote to each of the winning classes. (In 1-against-1 give 0.5 to each class when class 1 is not included.)
- Assuming that **the first classifier is broken** and makes a wrong classification
  - Repeat the exercise





Remember one-out-of-K coding

One-out-of-K coding

Nationality

TXT=

	Denmark	Norway	Sweden
'Sweden'	0	0	1
'Sweden'	0	0	1
'Sweden'	0	0	1
'Sweden'	0	0	1
'Norway'	0	1	0
'Norway'	0	1	0
'Norway'	0	1	0
'Norway'	0	1	0
'Norway'	0	1	0
'Sweden'	0	0	1
'Norway'	0	1	0
'Denmark'	1	0	0
'Denmark'	1	0	0
'Sweden'	0	0	1
'Sweden'	0	0	1
'Sweden'	0	0	1
'Denmark'	1	0	0
'Sweden'	0	0	1
'Norway'	0	1	0
'Denmark'	1	0	0

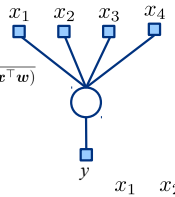
X_tmp=

Extending classifiers to handle multi-class problems

- Logistic regression

$$\{x_n, y_n\}_{n=1}^{N}$$

$$f(\boldsymbol{x}) = logit(\boldsymbol{x}^\top \boldsymbol{w}) = \frac{1}{1+\exp(-\boldsymbol{x}^\top \boldsymbol{w})}$$

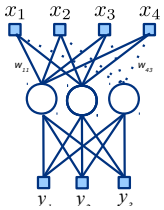


- Multinomial Regression

$$\{x_n, y_n\}_{n=1}^{N}$$

$$f_c(\boldsymbol{x}) = softmax(\boldsymbol{x}^\top \boldsymbol{W}) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{w}_c)}{\sum_{c'} \exp(\boldsymbol{x}^\top \boldsymbol{w}_{c'})}$$

$$\boldsymbol{w}_C = \boldsymbol{0}$$



Group exercise

- Show that for a binary (i.e, two class problem) multinomial regression is equivalent to logistic regression if $\mathbf{w}_0=\mathbf{0}$.

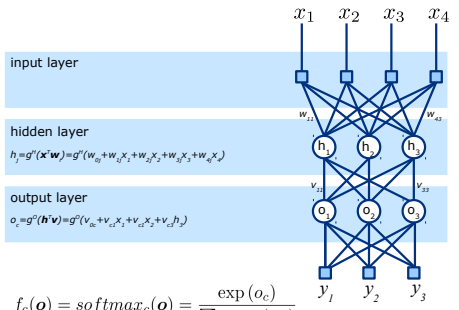
Hint: consider the softmax function for two classes:

$$f_c(\boldsymbol{x}) = softmax_c(\boldsymbol{x}^\top \boldsymbol{W}) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{w}_c)}{\sum_{c'} \exp(\boldsymbol{x}^\top \boldsymbol{w}_{c'})}$$

and show that this is equivalent to the logit function:

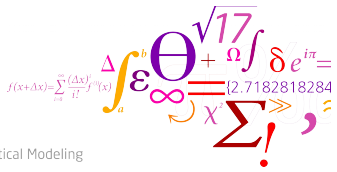
$$f(\boldsymbol{x}) = logit(\boldsymbol{x}^\top \boldsymbol{w}) = \frac{1}{1+\exp(-\boldsymbol{x}^\top \boldsymbol{w})}$$

Artificial Neural Networks for multiclass classification



$$f_c(\boldsymbol{o}) = softmax_c(\boldsymbol{o}) = \frac{\exp(o_c)}{\sum_{c'} \exp(o_{c'})}$$

02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 8.1-8.3+8.5.7

Feedback Groups of the day:

Carlos Roncal, Antonio Martin Rodriguez

Jullian Bellino, Simon Thordal

Siham Barghadouch, Ali Ilhan, Abiramy Mohanarajah

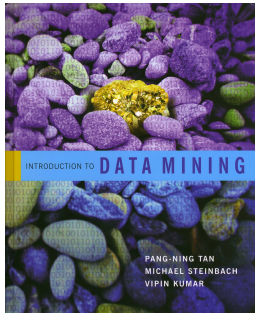
Kristian K. Riisgaard, Tryfon Tzanetis Ioanna Psylla

Michael Raagaard, Mathias Kaas-Olsen

Joachim Blom Hansen, Roman Piper

Karol Dzitkowski, Marco Becattini

Thorbjørn Wolf, Mick Neupart
If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

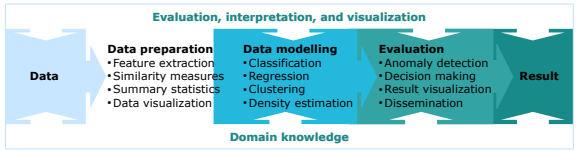
Unsupervised learning: Clustering and density est.

9. **K-means and hierarchical clustering (Tan 8.1-8.3+8.5.7)**
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

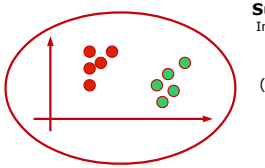
12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



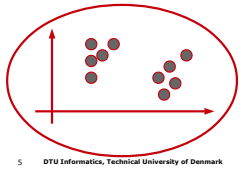
After today you should be able to:
Discuss the aims of unsupervised learning
Explain the principles behind hierarchical clustering and k-means clustering
Apply hierarchical clustering and k-means clustering
Evaluate the quality of the clustering using supervised measures of cluster validity

Supervised and Unsupervised learning



Supervised Learning
Input data $\mathbf{x}_n$ and output y_n

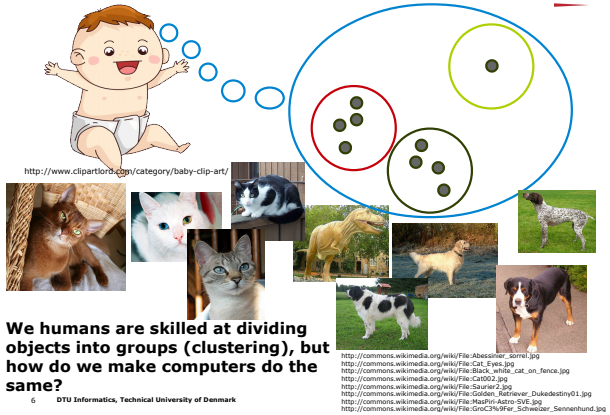
(Classification and Regression)



Unsupervised Learning
Input data $\mathbf{x}_n$ alone

(Exploratory analysis)

Imagine you observe the world for the first time!



http://www.clipartlord.com/category/baby-clip-art/

We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
http://commons.wikimedia.org/wiki/File:Cat002.jpg
http://commons.wikimedia.org/wiki/File:Saurier2.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
http://commons.wikimedia.org/wiki/File:MaiPiri-Astro-SVE.jpg
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg

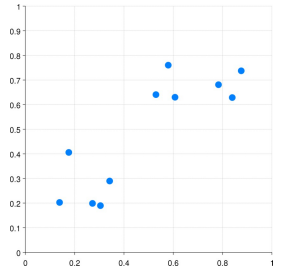
Unsupervised learning

- **Supervised learning**
  - Use the data to learn the output values
- **Unsupervised learning**
  - No output variables available
  - Sometimes called exploratory analysis
  - What learn from the data?
    - Structure
    - Regularities
    - Hidden information
    - Etc.

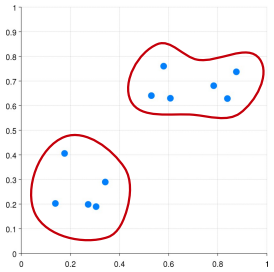
Clustering

- Divide data into groups (subsets/clusters) that are
  - **Meaningful**: Capture the natural structure of the data
  - **Useful**: Depends on purpose
- Observations in the same cluster are **similar in some sense**
- Unsupervised classification

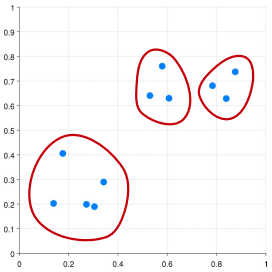
Clustering



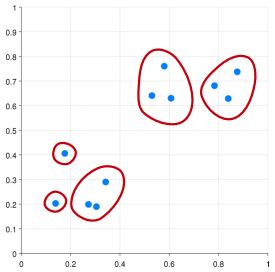
Clustering



Clustering

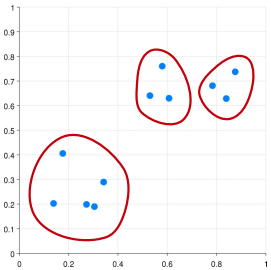


Clustering

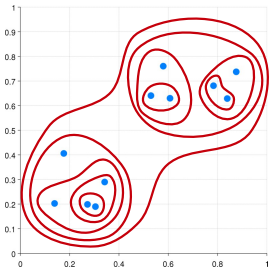


Partitional / hierarchical clustering

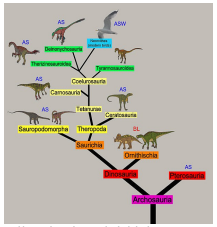
Partitional



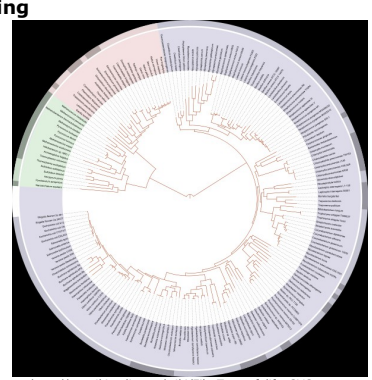
Hierarchical



Phylogenetic trees may be considered a type of hierarchical clustering



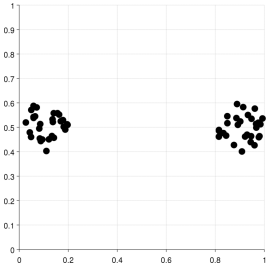
http://en.wikipedia.org/wiki/Phylogenetic_tree



Types of clustering

Well-separated

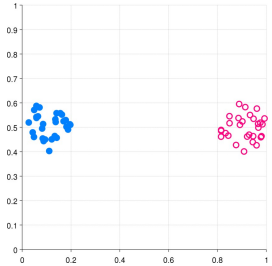
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Well-separated

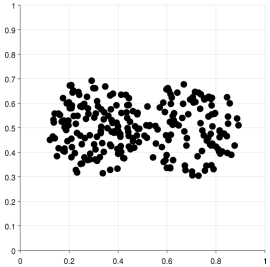
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Center-based

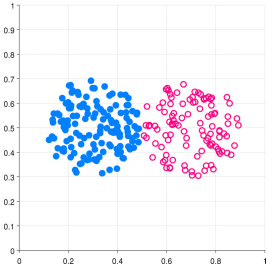
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Center-based

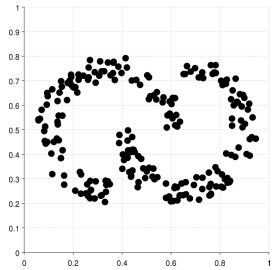
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Contiguity-based

- Each point is closer to at least one point in its cluster than to any point in another cluster

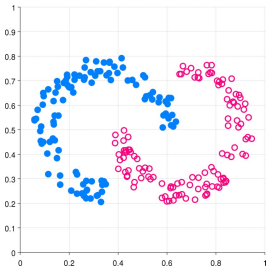


Types of clustering

Contiguity-based

- Each point is closer to at least one point in its cluster than to any point in another cluster

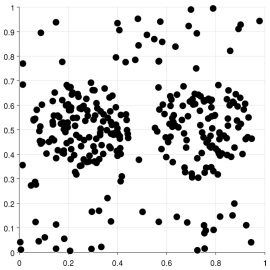
Clustering based on neighbor distance



Types of clustering

Density-based

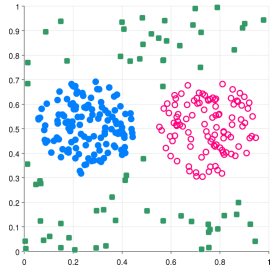
- Clusters are regions of high density separated by regions of low density



Types of clustering

Density-based

- Clusters are regions of high density separated by regions of low density

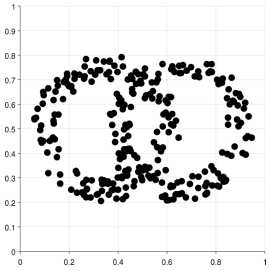


Types of clustering

Conceptual clusters

- Points in a cluster share some general property that derives from the entire set of points

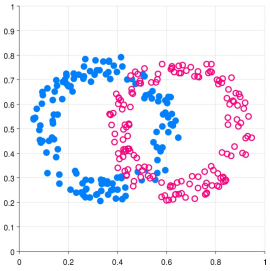
Requires some intelligence to parse



Types of clustering

Conceptual clusters

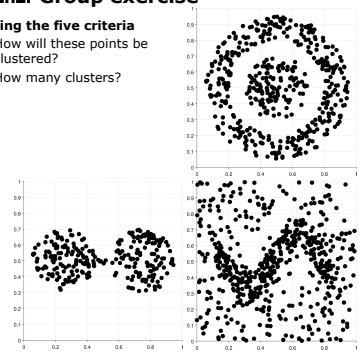
- Points in a cluster share some general property that derives from the entire set of points



Group exercise

Using the five criteria

- How will these points be clustered?
- How many clusters?



Well-separated
- Each point is closer to all points in its cluster than any point in another cluster

Center-based
- Each point is closer to the center of its cluster than to the center of any other cluster

Contiguity-based
- Each point is closer to at least one point in its cluster than to any point in another cluster

Density-based
- Clusters are regions of high density separated by regions of low density

Conceptual clusters
- Points in a cluster share some general property that derives from the entire set of points

K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change

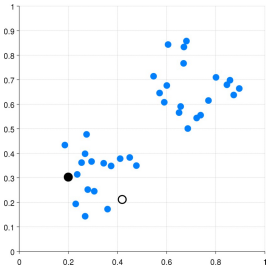
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

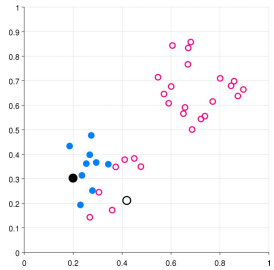
Until centroids do not change



K-means clustering

Select K points as initial centroids
Repeat
- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

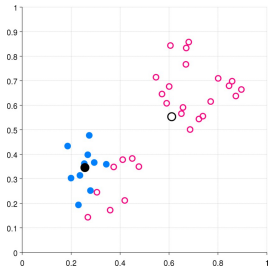
Until centroids do not change



K-means clustering

Select K points as initial centroids
Repeat
- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

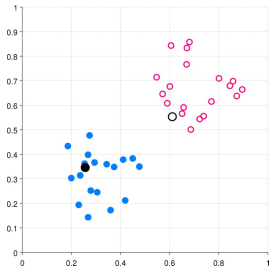
Until centroids do not change



K-means clustering

Select K points as initial centroids
Repeat
- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

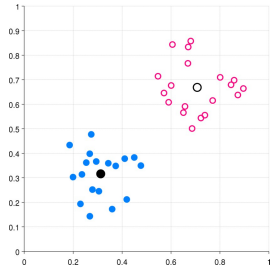
Until centroids do not change



K-means clustering

Select K points as initial centroids
Repeat
- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



K-means clustering

How do I
- Find the closest centroid?
  - Use a suitable **dissimilarity/similarity measure**
- Compute the cluster centroids
  - Depends on dissimilarity/similarity measure
  - For example, for Euclidean distance the mean is optimal (See Section 8.2.6 in Tan et al.)

To find the center of a cluster we minimize the L1 error.
L1 is minimal when c1 = mean of X1-c1 distances

Group exercise

Using pen-and-paper k-means, cluster the following data objects
- Number of clusters
  - K=2
- Distance measure
  - Euclidean
- Computation of centroid
  - Mean of cluster members
- Initial centroids
  - For example the first two data objects
- In case of any ties, flip a coin to decide

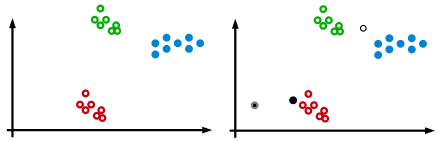
- **Data objects**

$$x = \{42, 60, 17, 48, 12\}$$

Select K points as initial centroids
Repeat
- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

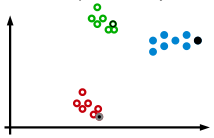
Until centroids do not change

How will the data (top-left diagram) be clustered given the initialization of the three centroids shown at the right and at the bottom?



- What could we do if we have an empty cluster?
- What could be a good initialization procedure? (Farthest First)

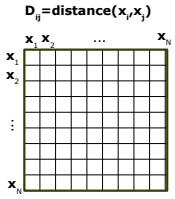
1. random point
2. farthest away point
3. farthest away point



Agglomerative hierarchical clustering

Compute the proximity matrix
Repeat
- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

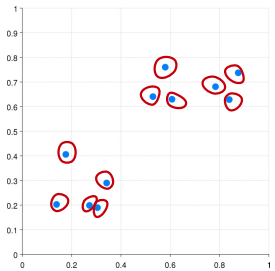
Until only one cluster remains



Agglomerative hierarchical clustering

Compute the proximity matrix
Repeat
- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



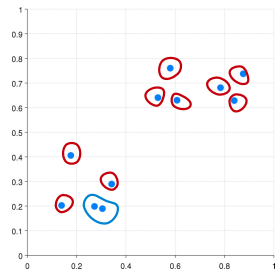
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



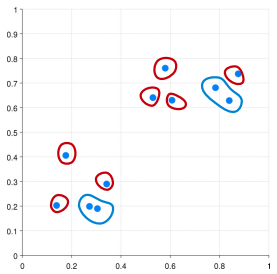
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



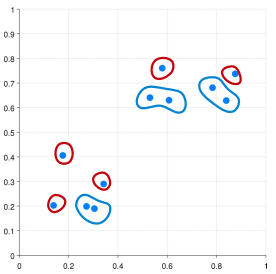
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



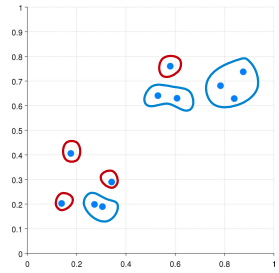
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



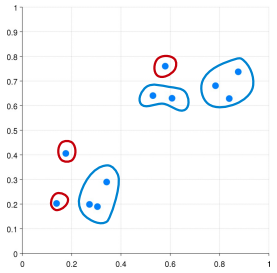
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



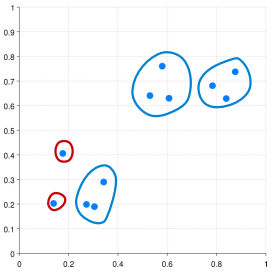
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



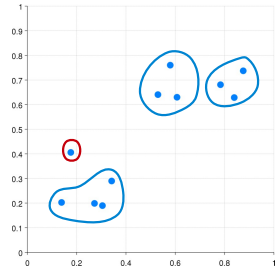
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



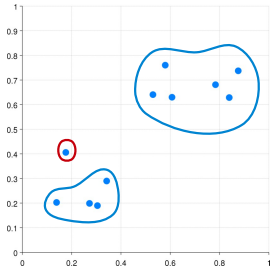
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



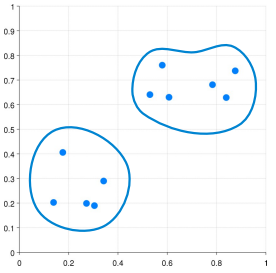
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



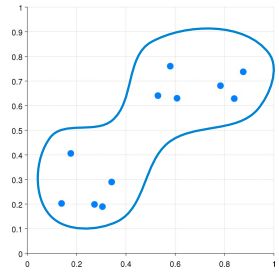
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

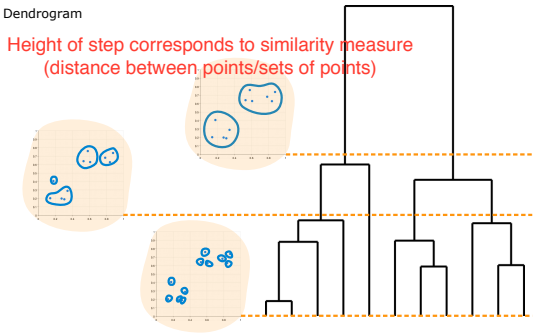
Until only one cluster remains



Agglomerative hierarchical clustering

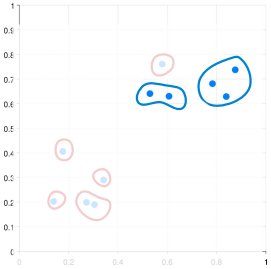
- Dendrogram

Height of step corresponds to similarity measure (distance between points/sets of points)



Similarity between clusters

- The **key operation** in agglomerative hierarchical clustering is measuring **similarity between clusters**



Proximity between clusters

- Can be computed using **proximity between objects**
- **Notice we need different definition if we are given a similarity or dissimilarity measure**
- In our example before we used Euclidian distance as proximity measure; i.e. it is the first definition which is relevant (dissimilarity)

C_i: Observations in cluster i
C_j: Observations in cluster j
m_i: Number of observations in cluster i
m_j: Number of observations in cluster j

Minimum (Single linkage)

Dissimilarity $proximity(C_i, C_j) = \min_{x \in C_i, y \in C_j} proximity(x, y)$

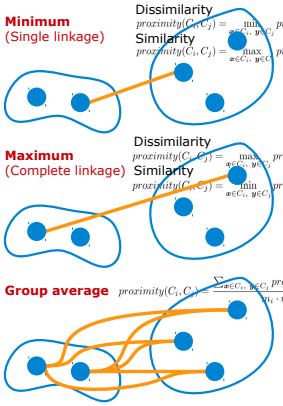
Similarity $proximity(C_i, C_j) = \max_{x \in C_i, y \in C_j} proximity(x, y)$

Maximum (Complete linkage)

Dissimilarity $proximity(C_i, C_j) = \max_{x \in C_i, y \in C_j} proximity(x, y)$

Similarity $proximity(C_i, C_j) = \min_{x \in C_i, y \in C_j} proximity(x, y)$

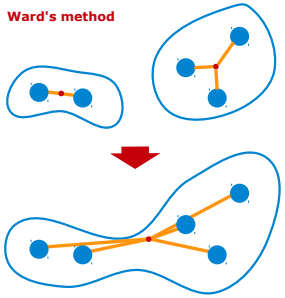
Group average $proximity(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} proximity(x, y)}{m_i \cdot m_j}$



Similarity between clusters

- Increase in sum of squared error after merging the two clusters should be as small as possible

Ward's method



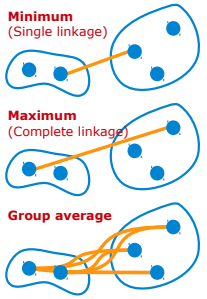
Group exercise

Can the choice of linkage be related to the notion of what constitutes clusters?

Minimum (Single linkage)

Maximum (Complete linkage)

Group average



Well-separated
- Each point is closer to all points in its cluster than any point in another cluster

Center-based
- Each point is closer to the center of its cluster than to the center of any other cluster

Contiguity-based
- Each point is closer to at least one point in its cluster than to any point in another cluster

Density-based
- Clusters are regions of high density separated by regions of low density

Conceptual clusters
- Points in a cluster share some general property that derives from the entire set of points

Group exercise

Using pen-and-paper agglomerative hierarchical clustering, **cluster** the following data objects and draw the **dendrogram**

- Distance measure
  - Euclidean
- Similarity between clusters
  - Minimum (Single linkage)
- **Data objects**

$$x = \{42, 60, 17, 48, 12\}$$

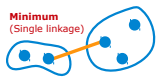
Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

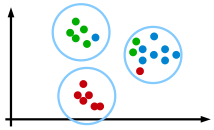
Until only one cluster remains

Minimum (Single linkage)



Cluster Purity measures when class labels are available

Motivation: Evaluate the extent to which manual classification process can be automatically produced by cluster analysis by comparing clustering to "ground truth"



Supervised measures of cluster validity

If we have (supervised) class labels, they can be used to measure the quality of the clustering

- **Cluster purity measures** (recall these from decision trees)
  - Entropy
  - Gini
  - Class error
- **Class label accuracy measures**
  - Precision
  - Recall
  - F-measure
- **Binary similarity measures**
  - Simple matching coefficient
  - Jaccard coefficient

Supervised measures of cluster validity

Cluster purity measures

- **Entropy**

$$Entropy(i) = -\sum_j p_{ij} \log_2 p_{ij}$$

- **Gini**

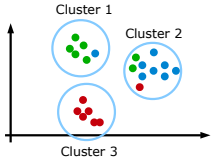
$$Gini(i) = 1 - \sum_j p_{ij}^2$$

- **Class error**

$$ClassError(i) = 1 - \max_j p_{ij}$$

- **Purity**

$$Purity(i) = \max_j p_{ij}$$



P

p_{ij} : Probability that member of cluster i belongs to class j $\quad p_{ij} = \frac{m_{ij}}{m_i}$

m_{ij}: Number of objects of class j in cluster i

m_i: Total number of objects in cluster i

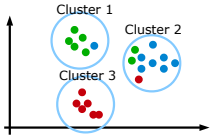
What is the Entropy, Gini, Class. Error and Purity for each of the three clusters given above?

	Entropy	Gini	Class. Error	Purity
Cluster 1	0.65	5/18	1/6	5/6
Cluster 2	1.16	23/50	3/10	7/10
Cluster 3	0	0	0	1

Notice that $0 \cdot log(0) = 0$

Supervised measures of cluster validity

Class label accuracy measures



- **Precision**

$$\text{Precision}(i,j) = \frac{m_{ij}}{m_i} = p_{ij}$$

- **Recall**

$$\text{Recall}(i,j) = \frac{m_{ij}}{m_j}$$

- **F-measure**

$$\text{F}(i,j) = \frac{2 \cdot \text{Precision}(i,j) \cdot \text{Recall}(i,j)}{\text{Precision}(i,j) + \text{Recall}(i,j)} = \frac{2m_{ij}}{m_i + m_j}$$

m_{ij}: Number of objects of class j in cluster i

m_i: Total number of objects in cluster i

m_j: Total number of objects in class j

p_{ij} : Fraction of objects of class j in cluster i

What is the Precision, recall and F-measure of the blue class in cluster 2?

Precision(blue,cluster 2)=7/10
Recall(blue,cluster 2)=7/8
F(blue,cluster 2)=2·7/(8+10)=7/9

Supervised measures of cluster validity

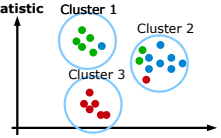
Binary similarity measures

- **Simple matching coefficient (SMC)/Rand statistic**

$$\text{SMC}(x,y) = \frac{f_{00} + f_{11}}{K}$$

- **Jaccard coefficient**

$$\text{J}(x,y) = \frac{f_{11}}{K - f_{00}}$$



K : Total number of **pairs of objects**, N·(N-1)/2

f_{00} : Number of object pairs in **different class** assigned to **different clusters**

f_{11} : Number of objects pairs in **same class** assigned to **same cluster**

In our example we find:

K=22·(22-1)/2=231

f_{11}=(5·(5-1)/2+1·(1-1)/2)$_{c1}$+(7·(7-1)/2+2·(2-1)/2+1·(1-1)/2)$_{c2}$+(6·(6-1)/2)$_{c3}$=10+22+15=47

F_{00}=(5·(7+1)+1·(2+1)+0·(2+7))$_{c1->c3}$+(5·6+1·6+0·0)$_{c1->c2}$+(2·6+7·6+1·0)$_{c2->c3}$=43+36+54=133

SMC=(47+133)/231=180/231

Jaccard=47/(231-133)=47/98

Exam question examples

QUESTION I:

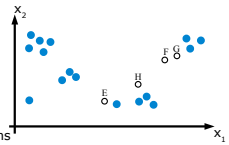
We have a one dimensional data set of size N = 5 with data examples x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 7 and x_5 = 12.

We run hierarchical clustering with a Euclidean dissimilarity between data points using group average linkage. We will use the following notation to summarize the dendrogram: (x y) means that x and y are joined in the binary tree. x and y can themselves be binary trees. What is the order we build the tree?

A. 12(34)5 → (12)(34)5 → ((12)(34))5 → (((12)(34))5).
B. 12(34)5 → 12((34)5) → (12)((34)5) → (12((34)5)).
C. 12(34)5 → (12)(34)5 → (12)((34)5) → ((12)((34)5)).
D. 12(34)5 → 12((34)5) → 1(2((34)5)) → (1(2((34)5))).

QUESTION II:

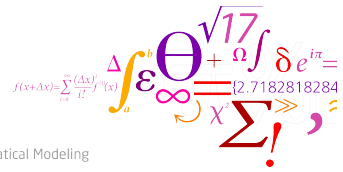
Consider the clustering problem given to the right where blue dots are observations and black circles are the initial position of four centroids denoted E,F,G and H used to cluster the data by k-means using Euclidean distances as dissimilarity. Upon convergence of the k-means algorithm which one of the following statements is wrong?



A. Cluster formed by centroid F will be empty
B. Cluster formed by centroid E will contain 10 observations
C. Clusters formed by centroid H will contain 4 observations
D. Cluster formed by centroid G will contain 3 observations

(Solution: QI: A, QII: B)

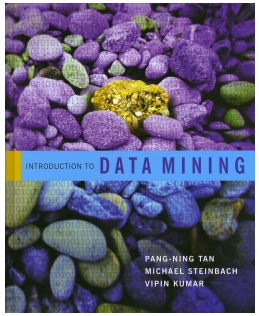
02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 9.2.2 + 6.1-6.3



Feedback Groups of the day:
Sofie Knøsgaard, Sarah Jørgensen, Sara Lolansen

Eleftherios-Marios Kotsonis-Tzannes, Mindaugas Venckus, Saad Eddin Al Orjany

Johannes Sanders, Øystein Monsen, Tomasz Kamiński

Louise Jørgensen, Ida Fillingsnes

Anders Ulrik Eliasen, Rasmus Sten Andersen, Rasmus Bendtsen

Loke Kristensen, Michael Nielsen

Søren Trads Steen, Janus Nørtoft Jensen

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.

Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

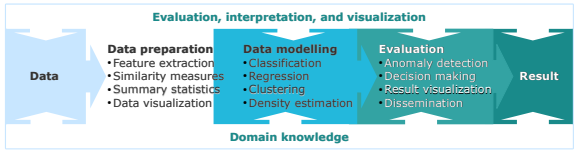
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3+8.5.7)
10. **Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)**
11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

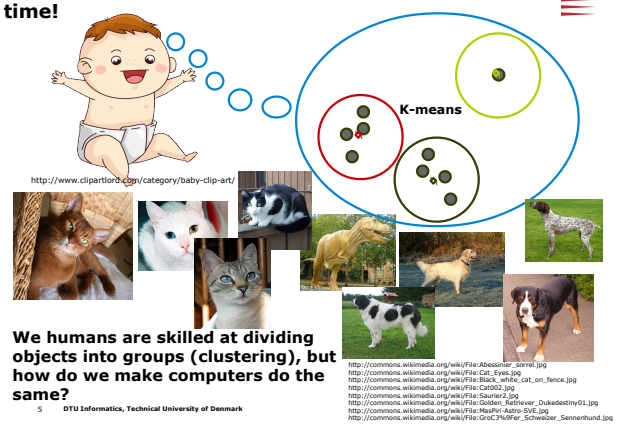
12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:
Explain the role of the parameters in the Gaussian Mixture Model (GMM) and how the parameters are updated using the EM-algorithm
Explain why cross-validation can be used for GMM
Describe the Apriori principle in association mining and explain how this can be used for efficient estimation of association rules.
Calculate support and confidence of association rules

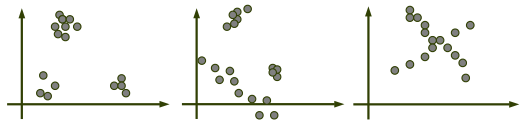
Imagine (again) you observe the world for the first time!



We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

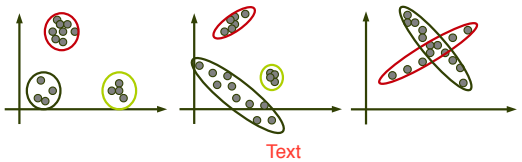
Group exercise

- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
  - Will it always find the optimum solution?
  - Can it model the sizes of the clusters?
  - Can it model the shape of the clusters?
  - How can we determine the number of clusters?



Group exercise

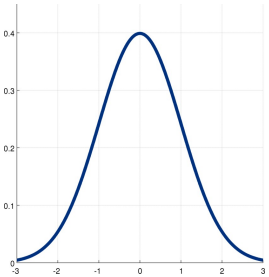
- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
  - Will it always find the optimum solution?
  - Can it model the sizes of the clusters?
  - Can it model the shape of the clusters?
  - How can we determine the number of clusters?



Normal distribution

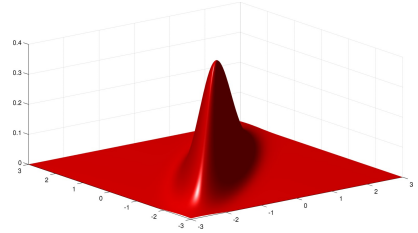
- Probability density function describes the relative chance of a given value to occur
- Normal distribution characterized by
  - Mean
  - Variance

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Multivariate Normal distribution

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$



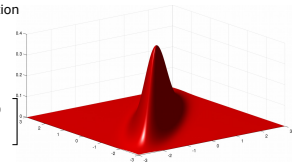
Multivariate Normal distribution

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

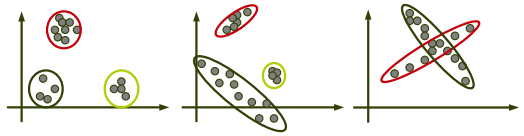
- Example: 2-dimensional Normal distribution

$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{bmatrix}$$



The Gaussian Mixture Model (GMM)



- Different locations $\boldsymbol{\mu}_{(k)}$
- Different shape $\boldsymbol{\Sigma}_{(k)}$
- Different sizes w_k

Data density

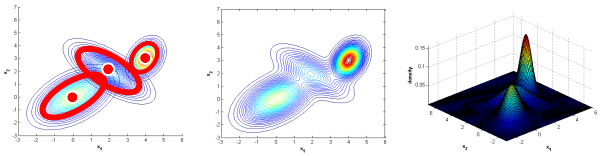
Sum of cluster specific densities assumed normal distributed

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}, \mu_{(k)}, \boldsymbol{\Sigma}_{(k)})$$

$$\text{s.t.} \sum_{k=1}^{K} w_k = 1, \; w_k \geq 0$$

GMM example

$$p(\boldsymbol{x}) = 0.5\mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}) + 0.2\mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}) + 0.3\mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.5 \end{bmatrix})$$



$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

$\boldsymbol{\Sigma}_{(k)}$: Shape of the cluster

w_k : Relative size/density of the cluster

Group exercise

- Consider the Gaussian mixture model (GMM)

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}, \mu_{(k)}, \boldsymbol{\Sigma}_{(k)}) \quad \text{s.t.} \sum_{k=1}^{K} w_k = 1, \; w_k \geq 0$$

- What is the value of the integral?

$$\int p(\mathbf{x}) \mathrm{d}\mathbf{x} \quad = 1$$

Apply sum rule!

Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

E-step

$$p(\boldsymbol{z}_n = k|\boldsymbol{x}_n) = \frac{w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}{\sum_{K=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}$$

M-step

$$N_k = \sum_{n=1}^{N} p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$

$$\boldsymbol{\mu}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} \boldsymbol{x}_n p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$

$$w_k = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})(\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})^\top p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$

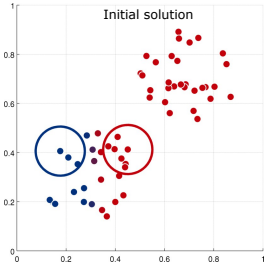
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



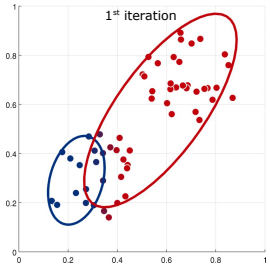
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



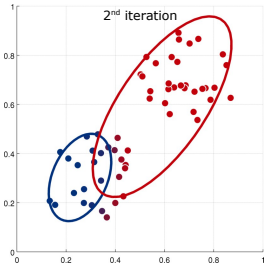
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



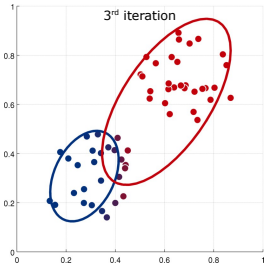
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



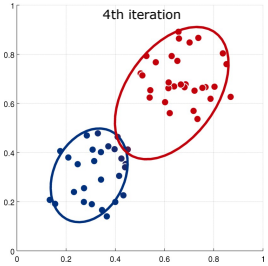
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



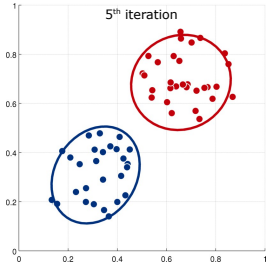
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



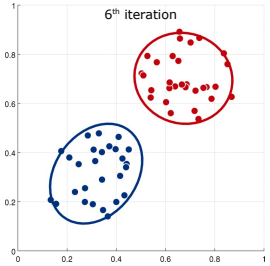
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

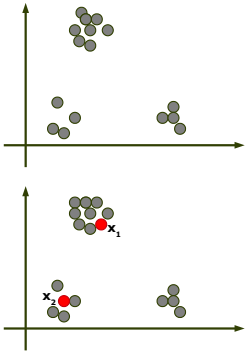
- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Group exercise

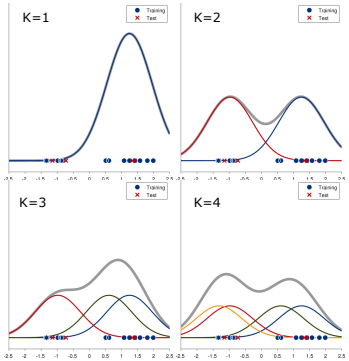
- Consider the data to the right with 16 observations.
  - What would ideally happen if we used a GMM with K=16 clusters to model the data?
- Imagine we have two **test observations** denoted $\mathbf{x}_1$ and $\mathbf{x}_2$ (red points) that are not used for training.
  - What happens to $p(\mathbf{x}_1)$ and $p(\mathbf{x}_2)$ if we use K=3 and K=14 clusters?



Mixture models

- Selecting complexity using crossvalidation

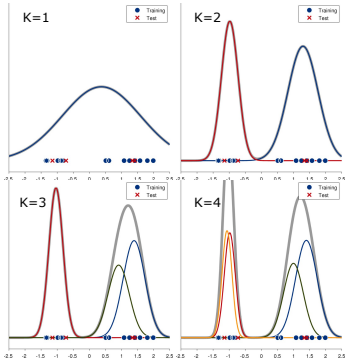
EM Initial solution



Mixture models

- Selecting complexity using crossvalidation

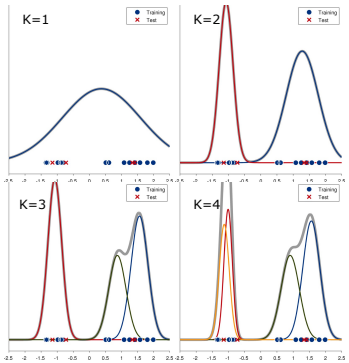
EM 1st iteration



Mixture models

- Selecting complexity using crossvalidation

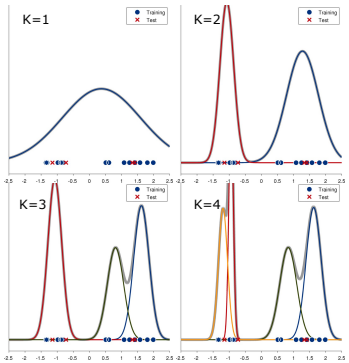
EM 2nd iteration



Mixture models

- Selecting complexity using crossvalidation

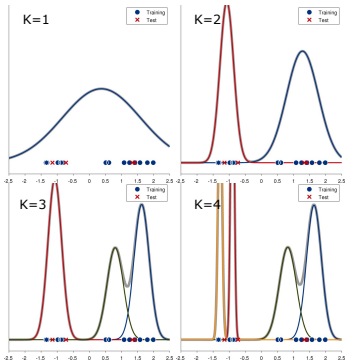
EM 3rd iteration



Mixture models

- Selecting complexity using crossvalidation

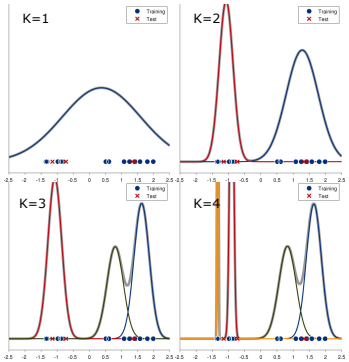
EM 4th iteration



Mixture models

- Selecting complexity using crossvalidation

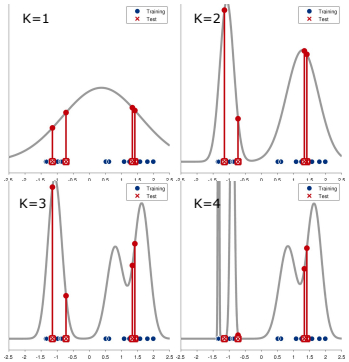
EM 5th iteration



Mixture models

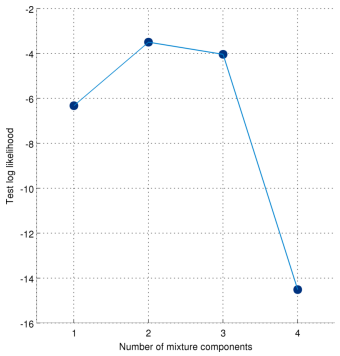
- Selecting complexity using crossvalidation

Test data evaluation



Mixture models

- Selecting complexity using crossvalidation



K-means versus GMM

K-means

- No guarantee of optimal solution
- Does not model shape of clusters
- Does not model the size of clusters
- Difficult to assess the number of clusters to use particularly when there is no ground truth

Gaussian mixture model (GMM)

- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation

http://siguealconejoblanco.es/comics/merchandising/escultura-de-superman-vs-muhammad-ali/

K-means versus GMM

K-means

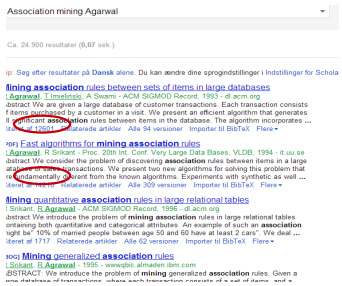
- No guarantee of optimal solution
- Does not model shape of clusters
- Does not model the size of clusters
- Difficult to assess the number of clusters to use particularly when there is no ground truth

Gaussian mixture model (GMM)

- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation

http://siguealconejoblanco.es/comics/merchandising/escultura-de-superman-vs-muhammad-ali/

Association Mining



Association rule discovery: Definition

- Given a set of **records**
  - Each containing a number of **items from a set**
- **Goal:** Produce dependency rules
  - Predict the occurence of an item based on occurences of other items

Association rule discovery: Example

Market basket analysis

Training set

1. {Bread, Soda, Milk}
2. {Beer, Bread}
3. {Beer, Soda, Diaper, Milk}
4. {Beer, Bread, Diaper, Milk}
5. {Soda, Diaper, Milk}

Rules discovered

{Milk} ► {Soda}
{Diaper, Milk} ► {Beer}

Market basket data

- Representation as

Transaction table

ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Data matrix

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Association analysis

- **Itemset**
  - For example {Bread, Soda, Milk}, {Milk, Diaper}, {}
- **Support** for an itemset **X**
  - Percentage of transactions that contain **X**
- **Association rule**
  - Expression of the form: **X ► Y** where **X** and **Y** are disjoint item sets

sigma = number of times

- **Support** for an association rule **X ► Y**
  - Percentage of transactions that contain **X ∪ Y**

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} = P(X, Y)$$

- **Confidence** for an association rule **X ► Y**
  - Percentage of transactions containing **X** that also contain **Y**

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{P(Y, X)}{P(X)} = P(Y|X)$$

Group exercise

- What is the **support** for
  - {Bread}
  - {Milk, Diaper}
- What is the **support** and **confidence** for
  - {Bread} ► {Milk}
  - {} ► {Milk}
- Find an **itemset** with
  - 0% support
  - 100% support
- Find an **association rule** with
  - 0% confidence
  - 100% confidence

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Itemset
- For example {Bread, Soda, Milk}, {Milk, Diaper}, {}

Support for an itemset **X**
- Percentage of transactions that contain **X**

Association rule
- Expression of the form: **X ► Y** where **X** and **Y** are disjoint item sets

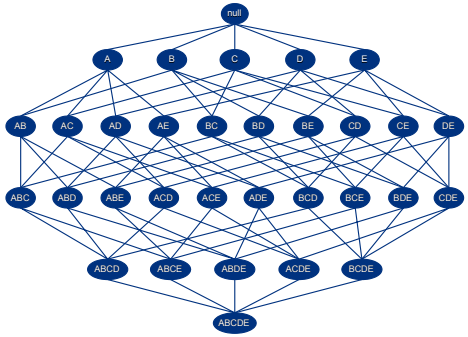
Support for an association rule **X ► Y**
- Percentage of transactions that contain **X ∪ Y**

Confidence for an association rule **X ► Y**
- Percentage of transactions containing **X** that also contain **Y**

Association rule mining

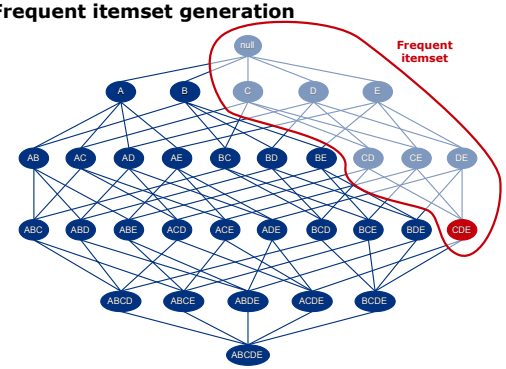
- Find all association rules that have
  - **Support ≥** *minsup*
  - **Confidence ≥** *minconf*
- Approach
  - **Frequent itemset generation**
    - Generate a list of all **itemsets** with **Support ≥** *minsup*
  - **Association rule generation**
    - Generate all **association rules** with **Confidence ≥** *minconf*

Frequent itemset generation



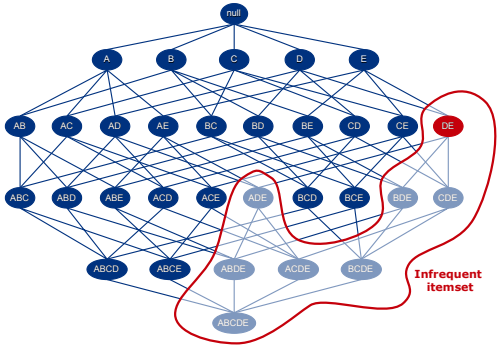
How many different itemsets can be created for a problem with a total of D items?

Frequent itemset generation



If an itemset is frequent, then all of its subsets must also be frequent

Frequent itemset generation



If an itemset is infrequent, then all of its supersets must also be infrequent

Group exercise

How many possible itemsets are there in the market basket below?
What are all the itemsets with support ≥ 35%?

	Juice	Milk	Beer	Cheese	Chocolate	Yoghurt	Sugar	Flour	Eeg	Wine
Customer 1	0	0	1	0	0	0	1	0	1	0
Customer 2	1	1	0	0	0	1	0	1	1	1
Customer 3	0	1	0	1	1	0	0	0	0	1
Customer 4	1	1	0	0	0	1	0	0	0	1
Customer 5	1	0	1	0	0	0	0	0	1	0
Customer 6	1	0	0	0	0	0	0	0	1	0
Customer 7	1	1	0	0	1	1	0	0	0	1
Customer 8	0	1	0	1	0	0	1	1	0	1
Customer 9	1	1	0	0	1	1	0	0	0	0
Customer 10	0	0	1	0	0	1	0	0	1	1

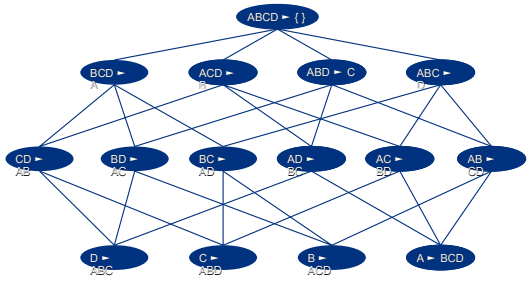
Group exercise

How many possible itemsets are there in the market basket below?
What are all the itemsets with support ≥ 35%?

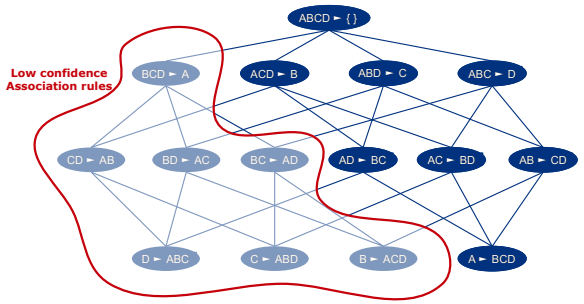
	Juice	Milk	Beer	Cheese	Chocolate	Yoghurt	Sugar	Flour	Eeg	Wine
Customer 1	0	0	1	0	0	0	1	0	1	0
Customer 2	1	1	0	0	0	1	0	1	1	1
Customer 3	0	1	0	1	1	0	0	0	0	1
Customer 4	1	1	0	0	0	1	0	0	0	1
Customer 5	1	0	1	0	0	0	0	0	1	0
Customer 6	1	0	0	0	0	0	0	0	1	0
Customer 7	1	1	0	0	1	1	0	0	0	1
Customer 8	0	1	0	1	0	0	1	1	0	1
Customer 9	1	1	0	0	1	1	0	0	0	0
Customer 10	0	0	1	0	0	1	0	0	1	1

2^{10}=1024 itemsets, itemsets with support ≥ 35% are:
{Juice}, {Milk}, {Yoghurt}, {Eeg}, {Wine}
{Juice, Milk}, {Juice, Yoghurt},{Milk, Yoghurt}, {Wine, Milk}, {Wine Yoghurt}
{Juice, Milk, Yoghurt}

Association rule generation



Association rule generation



Results for market basket example

Itemset	Support
Milk	80%
Bread	60%
Soda	60%
Beer	60%
Diaper	60%
Diaper Milk	60%
Soda Milk	60%
Bread Beer	40%
Bread Milk	40%
Soda Diaper	40%
Beer Diaper	40%
Beer Milk	40%
Soda Diaper Milk	40%
Beer Diaper Milk	40%

Association rule	Support	Confidence
{} ► Milk	80%	80%
Soda ► Milk	60%	100%
Diaper ► Milk	60%	100%
Soda, Diaper ► Milk	40%	100%
Beer, Diaper ► Milk	40%	100%
Beer, Milk ► Diaper	40%	100%

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Group exercise

- How can we do association mining for continuous data?

X=

Attribute 1	Attribute 2	Attribute 3
0.3689	0.9827	0.6999
0.4607	0.7302	0.6385
0.9816	0.3439	0.0336
0.1564	0.5841	0.0688
0.8555	0.1078	0.3196
0.6448	0.9063	0.5309
0.3763	0.8797	0.6544
0.1909	0.8178	0.4076
0.4283	0.2607	0.8200
0.4820	0.5944	0.7184
0.1206	0.0225	0.9686
0.5895	0.4253	0.5313
0.2262	0.3127	0.3251
0.3846	0.1615	0.1056
0.5830	0.1788	0.6110
0.2518	0.4229	0.7788
0.2904	0.0942	0.4235
0.6171	0.5985	0.0908
0.2653	0.4709	0.2665
0.8244	0.6959	0.1537

Binarize data according to percentiles

AttributeNames=

Attribute 1	Attribute 2	Attribute 3
0.3689	0.9827	0.6999
0.4607	0.7302	0.6385
0.9816	0.3439	0.0336
0.1564	0.5841	0.0688
0.8555	0.1078	0.3196
0.6448	0.9063	0.5309
0.3763	0.8797	0.6544
0.1909	0.8178	0.4076
0.4283	0.2607	0.8200
0.4820	0.5944	0.7184
0.1206	0.0225	0.9686
0.5895	0.4253	0.5313
0.2262	0.3127	0.3251
0.3846	0.1615	0.1056
0.5830	0.1788	0.6110
0.2518	0.4229	0.7788
0.2904	0.0942	0.4235
0.6171	0.5985	0.0908
0.2653	0.4709	0.2665
0.8244	0.6959	0.1537

X=

AttributeNamesBin=

Attribute 1 0-50 %	Attribute 1 50-100 %	Attribute 2 0-33.3 %	Attribute 2 33.3-66.7%	Attribute 2 66.7-100 %	Attribute 3 0-50%	Attribute 3 50-100%
1	0	0	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
0	1	1	0	0	1	0
0	1	0	0	1	0	1
1	0	0	0	1	0	1
1	0	0	0	1	1	0
0	1	1	0	0	0	1
0	1	0	1	0	0	1
1	0	1	0	0	0	1
0	1	0	1	0	0	1
1	0	1	0	0	1	0
1	0	1	0	0	1	0
0	1	1	0	0	0	1
1	0	0	1	0	0	1
1	0	1	0	0	1	0
0	1	0	1	0	1	0
1	0	0	1	0	1	0
0	1	0	0	1	1	0

Xbinary=

Further reading (not required)

Association Mining

- Rakesh Agrawal and Ramakrishnan Srikan "Fast Algorithms for Mining Association Rules", Proc. 20th Int. Conf. Very Large Data Bases, 1994
- Rakesh Agrawal, Tomasz Imieliński and Arun Swami "Mining association rules between sets of items in large databases" Proceedings of the 1993 ACM SIGMOD international conference on Management of data

Gaussian Mixture Model

- Christopher M. Bishop "Pattern Recognition and Machine Learning", Chapter 9, Springer 2006

http://commons.wikimedia.org/wiki/File:Old_book_bindings.jpg

Exam question examples

	Orange	Apple	Banana	Kiwi	Pear	Pineapple
Customer 1	1	0	1	0	0	1
Customer 2	1	1	0	0	1	1
Customer 3	0	1	0	1	1	0
Customer 4	1	1	0	0	0	0
Customer 5	0	0	1	0	0	0
Customer 6	1	0	1	0	0	0
Customer 7	1	1	0	0	1	1
Customer 8	0	1	0	1	0	0
Customer 9	1	1	0	0	0	1
Customer 10	0	0	1	0	0	1

QI: Consider the market basket given to the right where 10 customers have purchased various types of fruits in a grocery store. Disregarding the empty set, what are all frequent itemsets with support≥40%

A: {Orange}, {Apple}, {Banana}, {Pineapple}

B: {Orange}, {Apple}, {Banana}, {Pineapple}, {Orange, Apple}, {Orange, Pineapple}

C: {Orange}, {Apple}, {Banana}, {Pineapple}, {Orange, Apple}, {Orange, Pineapple}, {Apple, Pineapple}

D: {Orange}, {Apple}, {Banana}, {Pineapple}, {Orange, Apple}, {Orange, Pineapple}, {Apple, Pineapple},{Orange, Apple, Pineapple}

QII: What is the confidence of the decision rules {Orange, Apple}→{Pear}

A: 25%

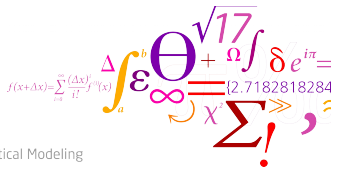
B: 50%

C: 75%

D: 100%

Correct answer QI: B, QII: B

02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 10.1-10.4

Feedback Groups of the day:

Wen Hsin Li, Robert Lyck, Viivi Halla-aho

Oliver Naaby, Rasmus Olsen

Simon Hemmingsen, Thomas Hammerbak

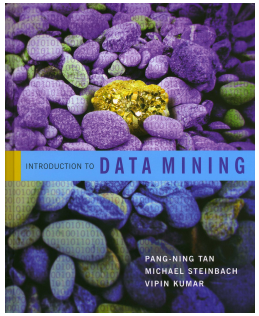
Henrik Thirstrup, Mette Marie Nielsen Thomas Kristensen

Nicolai Pedersen, Jeppe Thagaard, Mathias Brade, Christian Kragh

Oscar Petersen, Søren Vørnle Nielsen

Andreas Jacobsen, Asger Anker Sørensen, Phong Le Trung

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.



Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

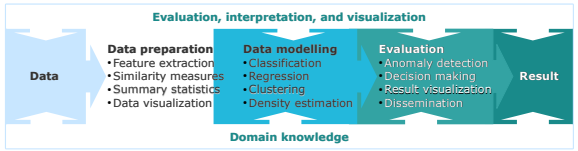
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3+8.5.7)
10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)
11. **Density estimation and anomaly detection (Tan 10.1-10.4)**

Machine learning and data modelling in practice

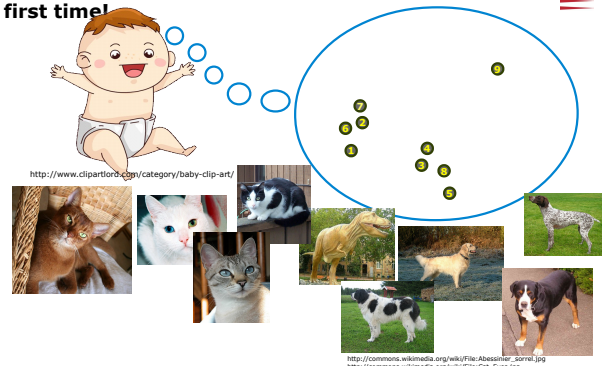
12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:
Understand and apply a variety of outlier/anomaly detection approaches, including density estimation approaches and proximity-based techniques.

Imagine (yet again) you observe the world for the first time!



http://www.clipartlord.com/category/baby-clip-art/

http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
http://commons.wikimedia.org/wiki/File:Cat002.jpg
http://commons.wikimedia.org/wiki/File:Saurier2.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
http://commons.wikimedia.org/wiki/File:MpdPei-Astro-SVE.jpg
http://commons.wikimedia.org/wiki/File:GroC3%BFer_Schweizer_Sennenhund.jpg

How do we detect anomalous objects (i.e., the dinosaur in the world of cats and dogs)

Anomaly detection: Definition

- Given a collection of data objects
  - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour

Anomaly detection: Example

- Credit card **fraud detection**
  - Recognize dubious credit card transactions based on the transaction history of the card holder
- Network **intrusion detection**
  - Detect hacker attacks, web crawlers etc.
- **Ecosystem disturbances**
  - Detect hurricanes, floods droughts, heat waves and fires
- **Healt and medicine monitoring**
  - Detect abnormal behaviour in populations and patients
- **Fault detection in industry systems**
  - Detect when a wind turbine performs poorly due to ice coating on blades
- Detection of **outliers** in data measurements
  - Remove erroneous measurements due to misreading from an instrument

Group exercise

- Come up with **your own definition** of an outlier / anomaly
- How can we detect outliers using some of the methods you have already learned in the course?

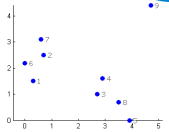
Group exercise

- Come up with **your own definition** of an outlier / anomaly
- How can we detect outliers using some of the methods you have already learned in the course?

Hawkins' definition of an outlier: *An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism*

Probabilistic definition of an outlier: *An outlier is an object that has a low probability with respect to a probability distribution model of the data.*

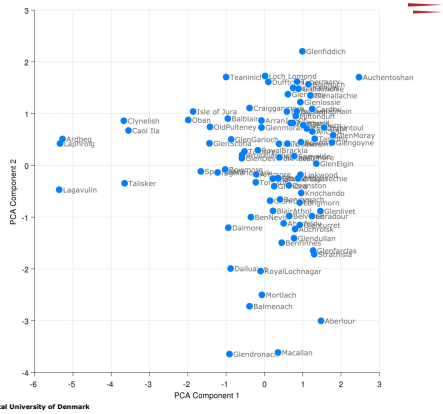
Data example I: Cats, Dogs and Dinosaurs



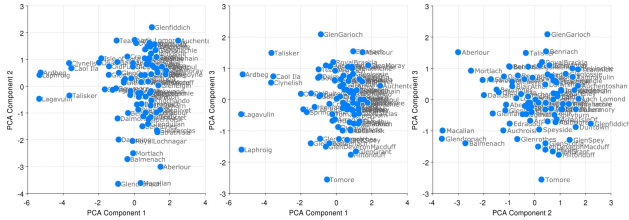
Data example II: Whisky

- 86 types of Scotch whisky
- Human ratings 1-5
- 12 taste categories
  - body, sweetness, smoky, medicinal, tobacco, honey, spicy, winey, nutty, malty, fruity, floral

PCA plot



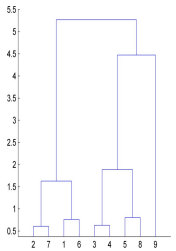
PCA plot



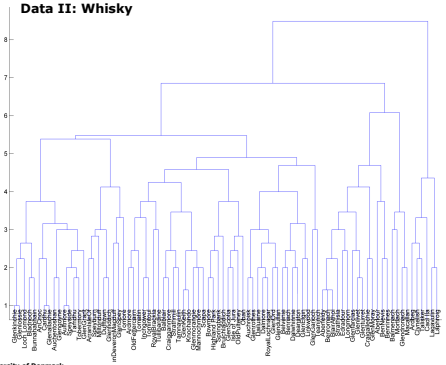
Dendrogram

- Dendrograms can be used to visualize relative distances between the observations

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



Approaches to anomaly detection

- **Density-based techniques**
  - Estimate the density of data objects
  - Outliers are:
    - Data objects in low density area

  Approaches we will consider:
  - Univariate normal distribution
  - Kernel density estimation

- **Proximity-based techniques**
  - Measure the distance between data objects
  - Outliers are
    - Data objects far from the other data objects

  Approaches we will consider:
  - Mahanalobis distance to center of data
  - Distance to K^{th} nearest neigbour
  - Inverse average distance to K nearest neighbours (KNN density)
  - Average relative KNN density

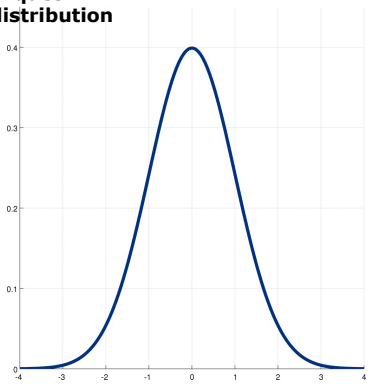
Density based techniques: Univariate normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



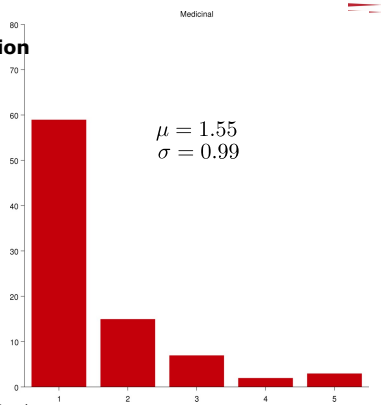
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



$$\mu = 1.55$$
$$\sigma = 0.99$$

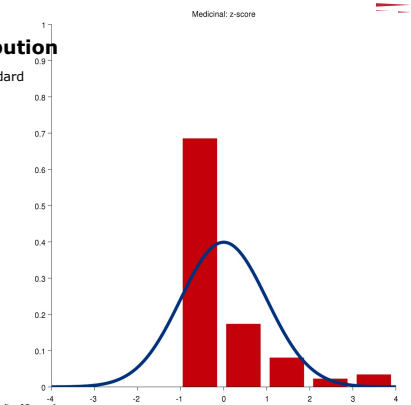
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



Normal distribution

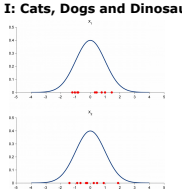
- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

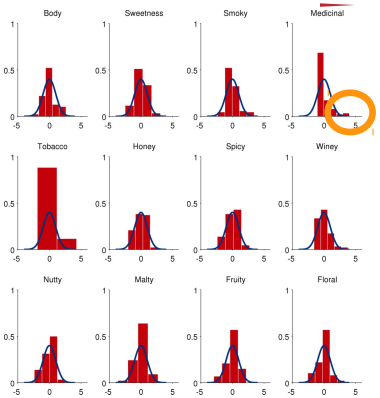
- Choose a threshold

$$z = \frac{x - \mu}{\sigma}$$
$$p(|z| > c) = 0.001$$
$$c = 3.2905$$

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



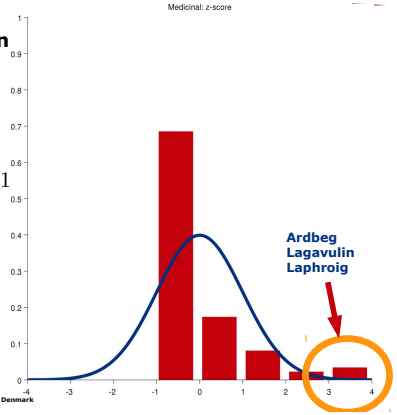
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

$$p(|z| > c) = 0.001$$
$$c = 3.2905$$

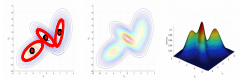


Density based techniques: Kernel Density Estimator

Rememer from Last week:
Gaussian Mixture Model (GMM)

Data density — Sum of cluster specific densities assumed normal distributed

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})$$
$$(s.t. \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0)$$

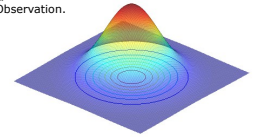


$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

$\boldsymbol{\Sigma}_{(k)}$: Shape of the cluster

w_k : Relative density of the cluster

Kernel Density estimation based on Gaussian Kernel: Consider the GMM and define a Gaussian with mean x_n and co-variance $\sigma^2\boldsymbol{I}$ around each Observation.



Let all observation weight the same, i.e. $w_n = 1/N$

$$p(\mathbf{x}) = \sum_{n=1}^{N} \frac{1}{N} \mathcal{N}(\mathbf{x}|\mathbf{x}_n, \sigma^2\mathbf{I})$$

Only free parameter σ!

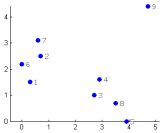
There is nothing special about the normal distribution. For a general mixture distribution p the general form of kernel density estimator is:

$$p(\mathbf{x}) = \sum_{n=1}^{N} \frac{1}{N} p(\mathbf{x}|\mathbf{x}_n, \theta)$$

This may be useful if x is discrete or non-negative.

How do we determine σ?

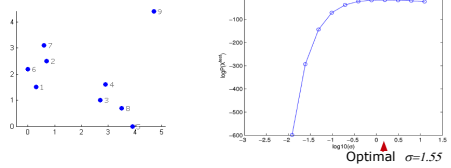
Data I: Cats, Dogs and Dinosaurs

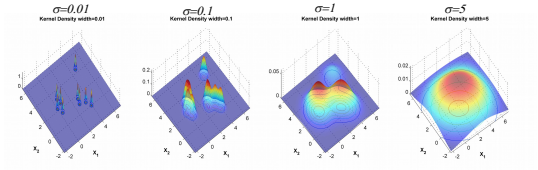


How do we determine σ?

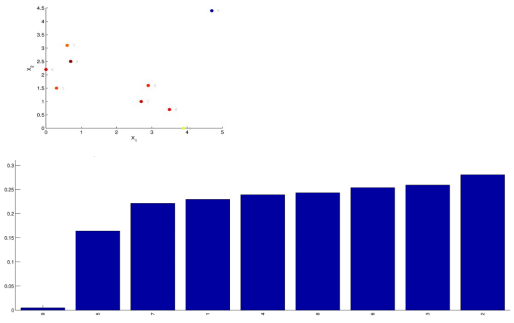
Data I: Cats, Dogs and Dinosaurs

Density of test set based on leave-one-out cross validation





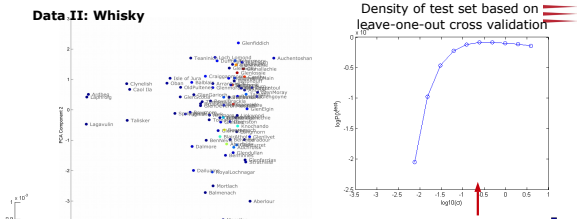
Etimated leave-one-out density evaluated at each observation



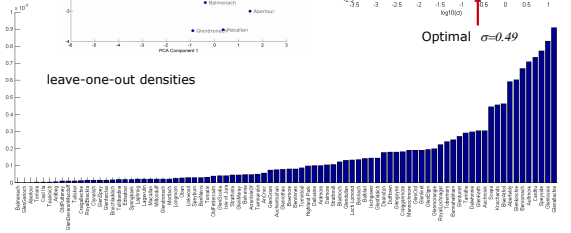
Etimated density evaluated at each observation

Data II: Whisky

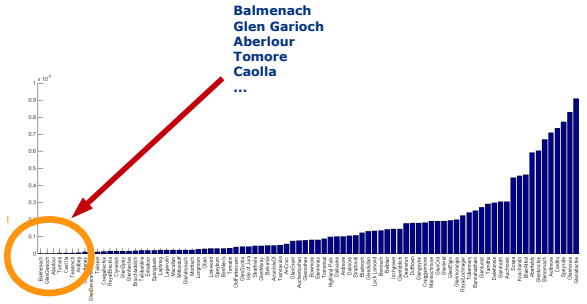
Density of test set based on leave-one-out cross validation



leave-one-out densities



Data II: Whisky



Proximity-based techniques

- Mahanalobis distance to center of data
- Distance to K[th] nearest neigbour
- Inverse average distance to KNN

Mahalanobis distance

- Distance to the mean of data, taking covariance into account

$$d_{\text{mahalanobis}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = (\boldsymbol{x} - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})$$

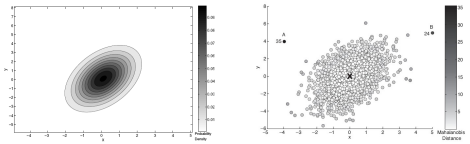


Figure 10.2. Probability density of points for the Gaussian distribution used to generate the points of Figure 10.3.

Figure 10.3. Mahalanobis distance of points from the center of a two-dimensional set of 2002 points.

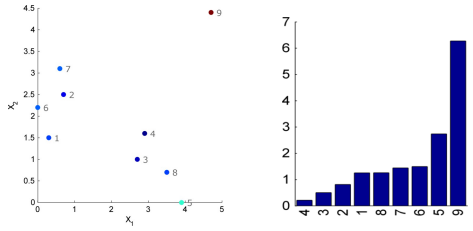
$\bar{\mathbf{x}}$ Mean vector

$\mathbf{S}$ Covariance matrix

Mahalanobis distance

- Distance to the mean of data, taking covariance into account

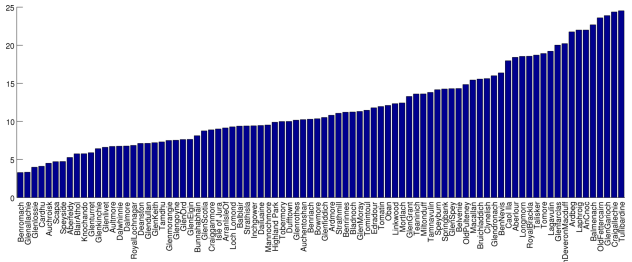
Data I: Cats , dogs and dinosaurs



Mahalanobis distance

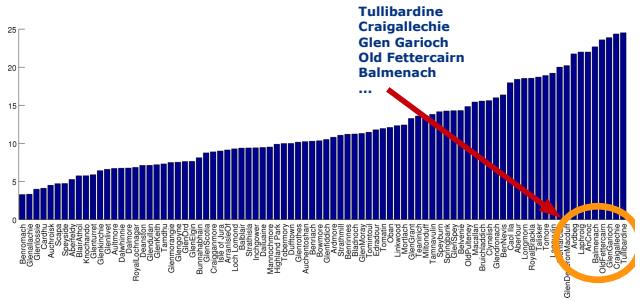
- Distance to the mean of data, taking covariance into account

Data II: Whisky



Mahalanobis distance

- Distance to the mean of data, taking covariance into account



Distance to k-nearest neighbor

- Measure the distance to the k'th nearest neighbor

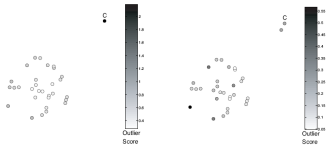


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

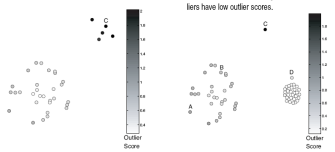


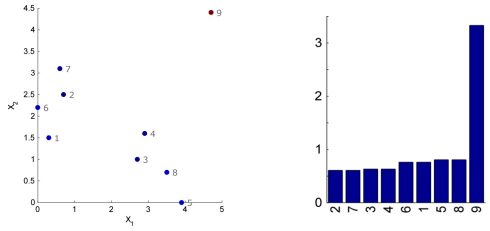
Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Distance to k-nearest neighbor

- Measure the distance to the 1st nearest neighbor

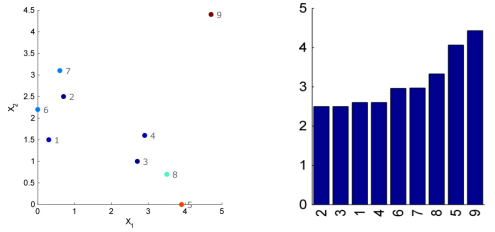
Data I: Cats , dogs and dinosaurs



Distance to kth Nearest neighbour

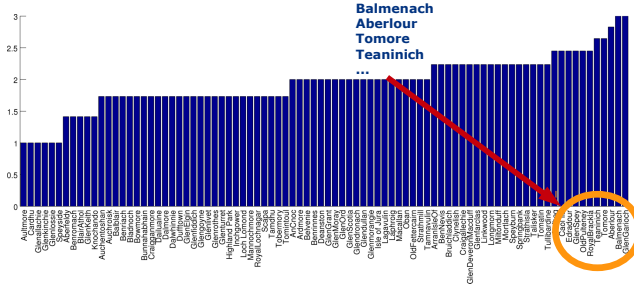
- Measure the distance to the 5th nearest neighbor

Data I: Cats , dogs and dinosaurs



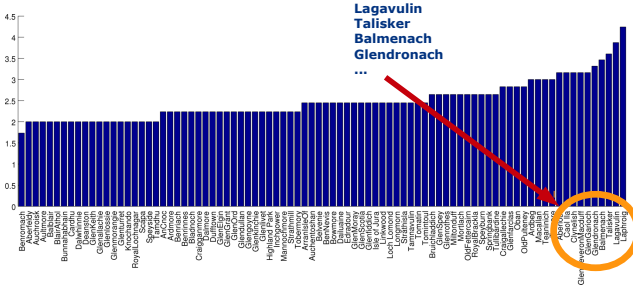
Distance to k-nearest neighbor

- Measure the distance to the 1st nearest neighbor



Distance to k-nearest neighbor

- Measure the distance to the 5th nearest neighbor



Inverse distance density estimation

- **Distance based measure of density**
  - Density is inverse proportional to avereage distance to k nearest neighbors
  - Density is low if nearest neighbors are far away

$$\text{density}(\boldsymbol{x}, k) = \left(\frac{1}{k} \sum_{\boldsymbol{y} \in N(\boldsymbol{x},k)} \text{distance}(\boldsymbol{x}, \boldsymbol{y}) \right)^{-1}$$

- **Relative density**
  - Density compared to density at nearest neighbors

$$\text{average relative density}(\boldsymbol{x}, k) = \frac{\text{density}(\boldsymbol{x}, k)}{\frac{1}{k} \sum_{\boldsymbol{y} \in N(\boldsymbol{x},k)} \text{density}(\boldsymbol{y}, k)}$$

$N(\boldsymbol{x}, k)$ The set of k nearest neighbors

Consider the pairwise distance matrix given to the left. What is the density and average relative density of the first observation for k=3?

$d(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$
$\mathbf{x}_1$	0	2.5	2.4	4.0	0.8	0.6	3.3
$\mathbf{x}_2$	2.5	0	0.6	1.6	2.9	3.0	1.1
$\mathbf{x}_3$	2.4	0.6	0	1.9	3.0	2.7	1.0
$\mathbf{x}_4$	4.0	1.6	1.9	0	4.5	4.6	3.8
$\mathbf{x}_5$	0.8	2.9	3.0	4.5	0	1.1	3.9
$\mathbf{x}_6$	0.6	3.0	2.7	4.6	1.1	0	3.8
$\mathbf{x}_7$	3.3	1.1	1.0	3.8	3.9	3.8	0

$$\text{density}(\boldsymbol{x},k)=\left(\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{distance}(\boldsymbol{x},\boldsymbol{y})\right)^{-1}$$

$$\text{average relative density}(\boldsymbol{x},k)=\frac{\text{density}(\boldsymbol{x},k)}{\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{density}(\boldsymbol{y},k)}$$

Consider the pairwise distance matrix given to the left. What is the density and average relative density of the first observation for k=3?

$d(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$
$\mathbf{x}_1$	0	2.5	2.4	4.0	0.8	0.6	3.3
$\mathbf{x}_2$	2.5	0	0.6	1.6	2.9	3.0	1.1
$\mathbf{x}_3$	2.4	0.6	0	1.9	3.0	2.7	1.0
$\mathbf{x}_4$	4.0	1.6	1.9	0	4.5	4.6	3.8
$\mathbf{x}_5$	0.8	2.9	3.0	4.5	0	1.1	3.9
$\mathbf{x}_6$	0.6	3.0	2.7	4.6	1.1	0	3.8
$\mathbf{x}_7$	3.3	1.1	1.0	3.8	3.9	3.8	0

$$\text{density}(\boldsymbol{x},k)=\left(\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{distance}(\boldsymbol{x},\boldsymbol{y})\right)^{-1}$$

$$\text{average relative density}(\boldsymbol{x},k)=\frac{\text{density}(\boldsymbol{x},k)}{\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{density}(\boldsymbol{y},k)}$$

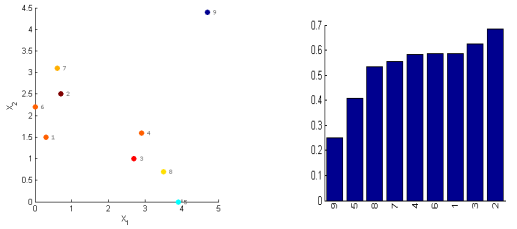
density($\mathbf{x}_1$,3)=[1/3(0.6+0.8+2.4)]$^{-1}$=3/3.8
density($\mathbf{x}_6$,3)=[1/3(0.6+1.1+2.7)]$^{-1}$=3/4.4
density($\mathbf{x}_2$,3)=[1/3(0.8+1.1+2.9)]$^{-1}$=3/4.8
density($\mathbf{x}_3$,3)=[1/3(0.6+1.0+1.9)]$^{-1}$=3/3.5

Av. Rel. Density($\mathbf{x}_1$,3)=[3/3.8]/[1/3 (3/4.4+3/4.8+3/3.5)]=1.0945

Inverse distance density estimation

- KNN density (5 nearest neighbors)

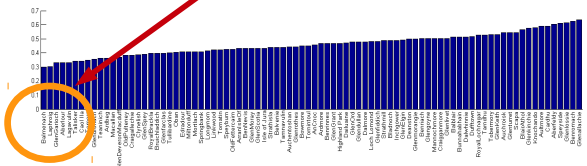
Data I: Cats , dogs and dinosaurs



Inverse distance density estimation

- KNN density (5 nearest neighbors)

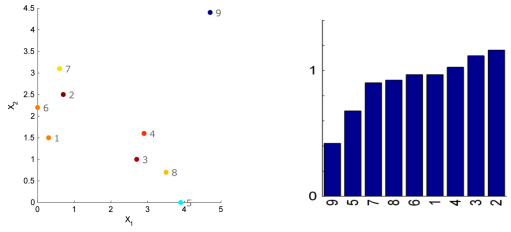
Balmenach
Laphroig
GlenGarioch
Aberlour
Lagevulin
...



Average Relative density

- Average Relative KNN density (5 nearest neighbors)

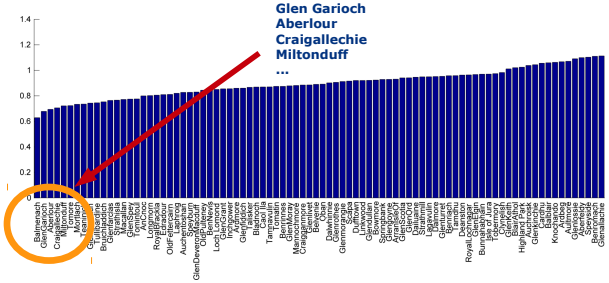
Data I: Cats , dogs and dinosaurs



Average relative density

- Average relative KNN density (5 nearest neighbors)

Balmenach
Glen Garioch
Aberlour
Craigallechie
Miltonduff
...



Results using different methods

- **Univariate Normal distribution**
  - Ardbeg
  - Lagavulin
  - Laphroig
- **Kernel Density Estimation**
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Tomore
  - Caolla
- **Mahalanobis distance**
  - Tullibardine
  - Craigallechie
  - Glen Garioch
  - Old Fettercairn
  - Balmenach
- **Distance to nearest neighbor**
  - Glen Garioch
  - Balmenach
  - Aberlour
  - Tomore
  - Teaninich
- **Distance to 5th nearest neighbor**
  - Laphroig
  - Lagavulin
  - Talisker
  - Balmenach
  - Glendronach
- **KNN density**
  - Balmenach
  - Laphroig
  - Glen Garioch
  - Aberlour
  - Lagavulin
- **KNN average relative density**
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Craigallechie
  - Miltonduff

Common: Balmenach, Glen Garioch, Laphroig, Aberlour, Tomore, Lagavulin, Craigallechie

Example of exam questions

QI: What is the average relative density for observation 2 (i.e. $\mathbf{x}_2$) for k=2 nearest neighbours?
A:1/5
B:3/10
C:7/10
D:1

$d(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{x}_1$	0	2.0	0.2	0.9	0.2
$\mathbf{x}_2$	2.0	0	1.5	0.5	2.0
$\mathbf{x}_3$	0.2	1.5	0	1.2	1.4
$\mathbf{x}_4$	0.9	0.5	1.2	0	1.0
$\mathbf{x}_5$	0.2	2.0	1.4	1.0	0

$$\text{density}(\boldsymbol{x},k)=\left(\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{distance}(\boldsymbol{x},\boldsymbol{y})\right)^{-1}$$

$$\text{avg. rel. den.}(\boldsymbol{x},k)=\frac{\text{density}(\boldsymbol{x},k)}{\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{density}(\boldsymbol{y},k)}$$

QI: C