BIG DATA and Machine Learning

Christine zu Eulenburg
Medical Statistics and Decision Making
UMCG

09.10.2018

Note

For the machine learning part, graphs are „borrowed“ from the great book of

G. James, D. Witten, T. Hastie, and R. Tibshirani: *An Introduction to Statistical Learning, with Applications in R.* Springer, 2013

Content

- What is BIG DATA?
- From Big Data to machine learning
- Example 1: K nearest neighbors
- Example 2: Support vector machines
- Validation methods
  - Cross-validation
  - The Bootstrap
- Example 3: K-means cluster analysis
- Summary

What is BIG DATA?

https://twitter.com/devops_borat

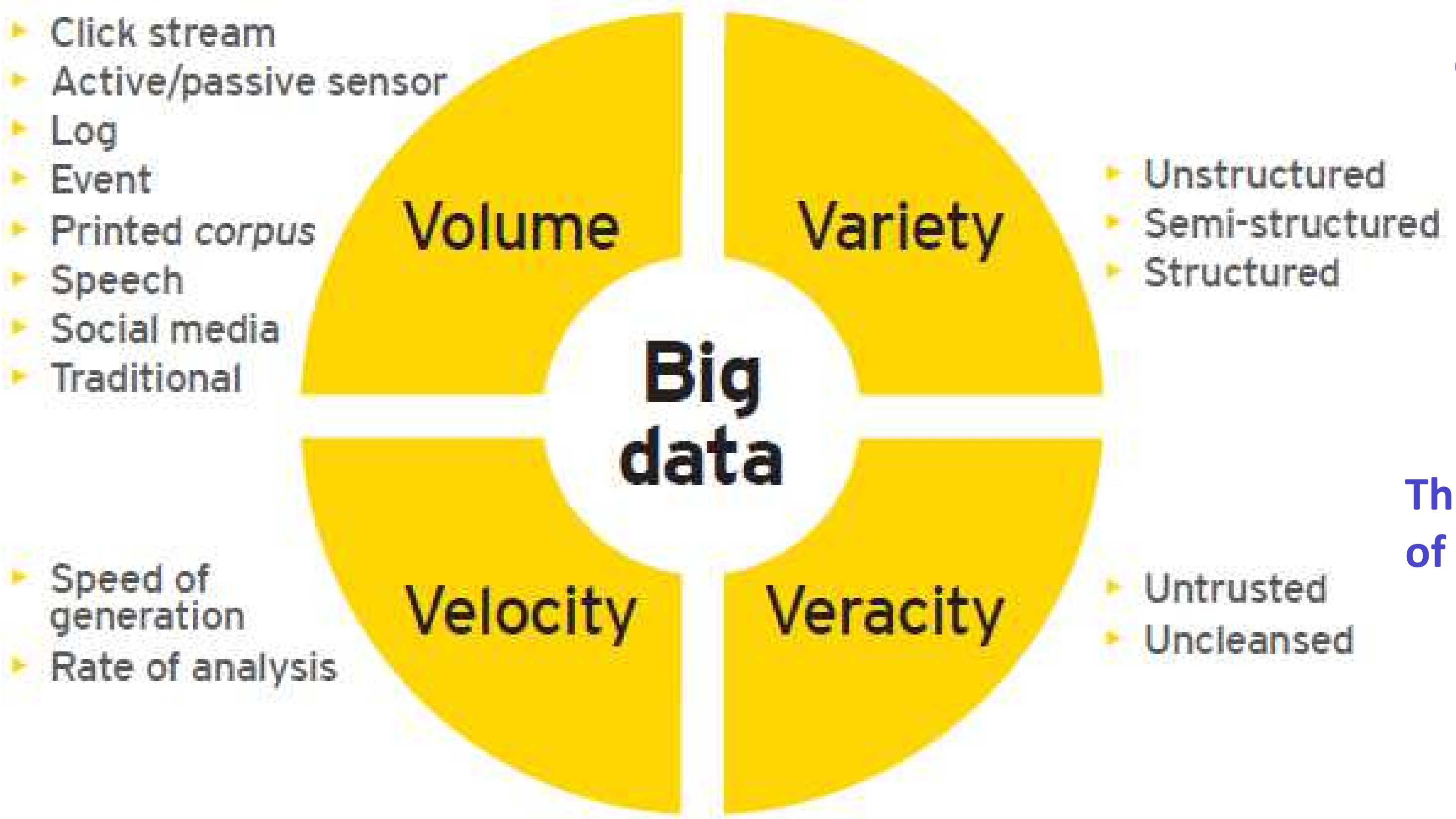
Or, in other words, Big Data is data in volumes too great to process by traditional methods.

Big data

…defined through the 4 v's

The **amount** of data

The **types** of data



The **frequency** of data

The **quality** of data

Volume: scale of data

Unit	Value	Size	
bit (b)	0 or 1	1/8 of a byte	
byte (B)	8 bits	1 byte	~ a single character
kilobyte (KB)	1000^1 bytes	1,000 bytes	~ half page of written text
megabyte (MB)	1000^2 bytes	1,000,000 bytes	~ one typical sized photograph
gigabyte (GB)	1000^3 bytes	1,000,000.000 bytes	~ average size of a DIVX movie, or a pickup full of pape
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes	~ 50% of all english Wikipedia files in May 2012
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes	~ 50% of all US academic research libaries
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes	5 EBs ~ all words ever spoken by human being
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes	
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes	

Complete Works of Shakespeare: 5 Megabyte
Smallest Iphone 7: 32 Gigabyte
Internet traffic by 2016: 1,3 Zettabytes

Volume: scale of data

- 90% of today's data has been created in just the last 2 years
- Every day we create 2.5 quintillion bytes of data or enough to fill 10 million Blu-ray discs
- 40 zettabytes (40 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005, and the equivalent of 5,200 gigabytes of data for every man, woman and child on Earth

Source: NIH U.S. National Library of Medicine
National Network of Libraries of Medicine
Pacific Northwest Region

Variety: different forms of data

- Data heterogeneity
- Structured (numbers) and unstructured data (images, text, spoken language)

Velocity: analysis of streaming data

- High speed of data flow, change and processing
- Real-time data

Veracity: various levels of data uncertainty and reliability

- Different level of data quality of structured data (precise numbers) and unstructured data (fuzzy interpreting of images, free text, spoken words,…)
- Technical data quality issues. Various formats, updates)

Challenges in processing and managing Big Data

Before it comes to the analysis of Big Data, there are a lot of practical issues to solve:

- Technically: ingesting the data
- Synchronize data from different sources
- Updating derived data
- Storing large amounts of data
- Data quality issues
- Dealing with unstructured data (eg extracting information from texts/ images)
- Importing of raw data and selection of useful inormation is an important process. Wrong data results in wrong conclusions (garbage in – garbage out)

From Big Data to machine learning

Data collection

Synchronization and storage

hadoop

Data cleaning, converting

Speech Recognition

Natural Language Processing

Data analysis

Machine Learning

Data mining

Interpretation, Decision support

Human-Computer Interaction

…that's where statisticians come into the play

Machine Learning Examples

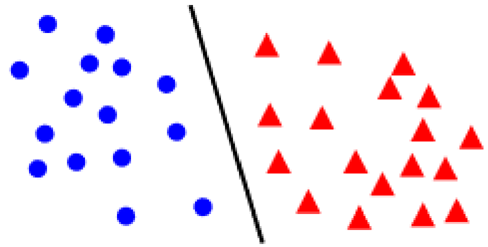
- Search and recommendation (e.g. Google, Amazon, Netflix)
- Automatic speech recognition and speaker verification
- Text parsing
- Face identification
- Financial prediction, fraud detection (e.g. credit cards)
- Medical diagnosis

Two kinds of learning

- Supervised learning (‘ machine learning’)
  - Use training data to optimize the algorithm
  - Building a statistical model on an outcome
  - Apply algorithms to test data
- Unsupervised learning (‘ data mining’)
  - No outcome variable
  - Clustering and factoring, finding patterns in data
  - Dimension reduction

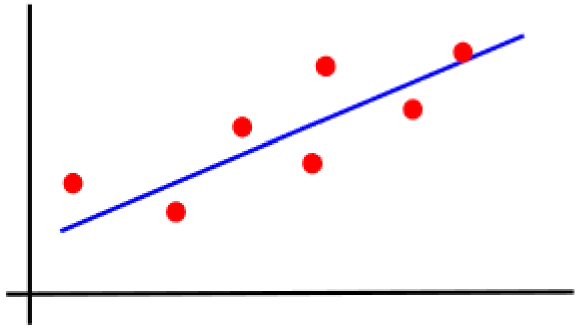
Supervised learning

– a categorical outcome = classification



e.g. predicting therapy response from clinical markers

– a continuous outcome = regression



e.g. predicting weight from height

A selection of ML-techniques

Supervised learning		Unsupervised learning
Classification	**Regression**	
Logistic regression	Linear regression	Principal Component Analysis
Linear discriminant analysis		Partial Least Squares Regression
K Nearest Neighbors		Cluster Analysis
Naive Bayes		
Support Vector Machines		
Tree-based methods (Classification / Regression Trees, Random Forests)		
Neural Networks		

Supervised learning

Goal: defining an association between a set of predictors X and an outcome Y with minimized error

$$Y = f(X) + \epsilon$$

With $\min(MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f(x_i)})^2)$

To answer the questions

- Which predictors are associated with the response?
- What is the relationship between the response and the predictor?
- If we knew our X, how would we estimate Y?

Supervised learning

Standard approach:

1. split data into training and test set (e.g. 80/20).
2. Use only the training set to adapt a statistical model so that $f(X) \approx Y$ („learning“).
3. To check the performance of the model, validate it on the test set. If our goal is prediction, the method with the highest accuracy / lowest MSE wins!

Supervised learning

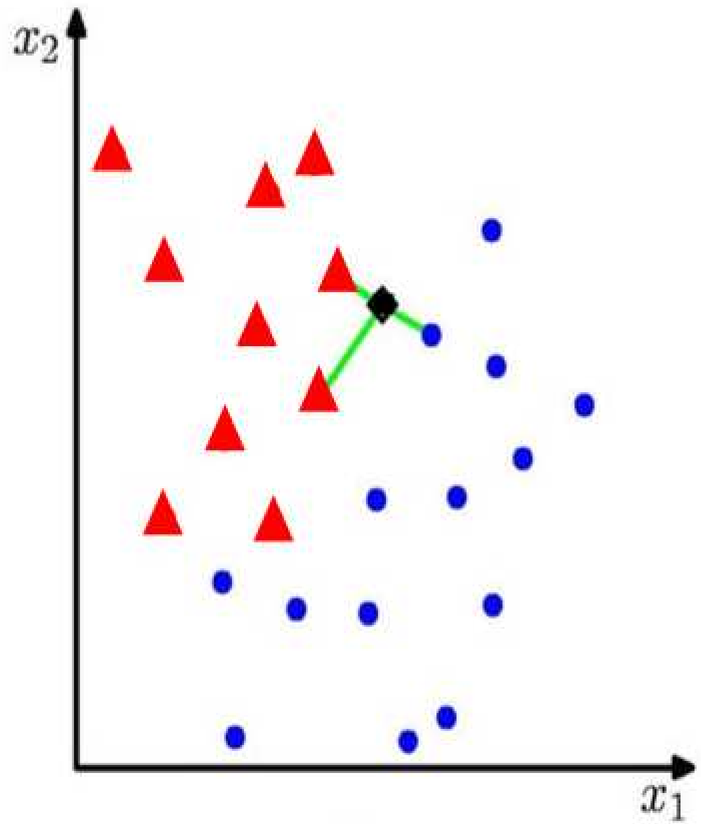
Goal: Predicting and estimating

Standard approach:

1. split data into training and test set (e.g. 80/20).
2. Use only the training set to adapt a statistical model (e.g. logistic regression)
3. To check the performance of the model, validate it on the test set.

Example: The K-NN classifier

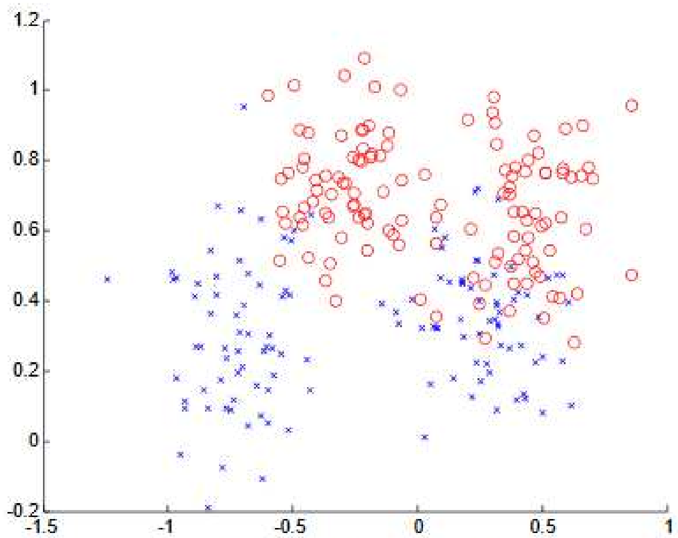
(k nearest neighbours)

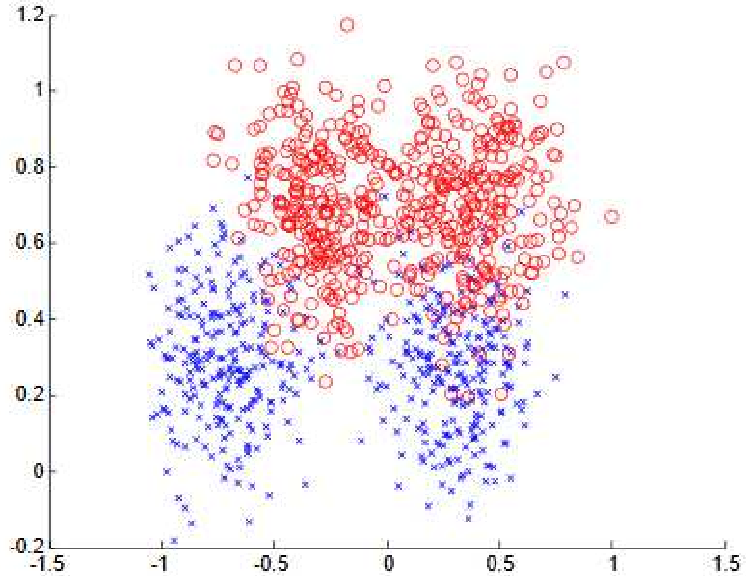


Algorithm, k=3

1. For each point in the test-dataset, find the 3 nearest neighbours (euclidean distance).
2. Classify the point as the majority of neighbours.

Assume that a test sample is drawn randomly from the data set. Then you would expect the same pattern in the test sample and in the remaining data.

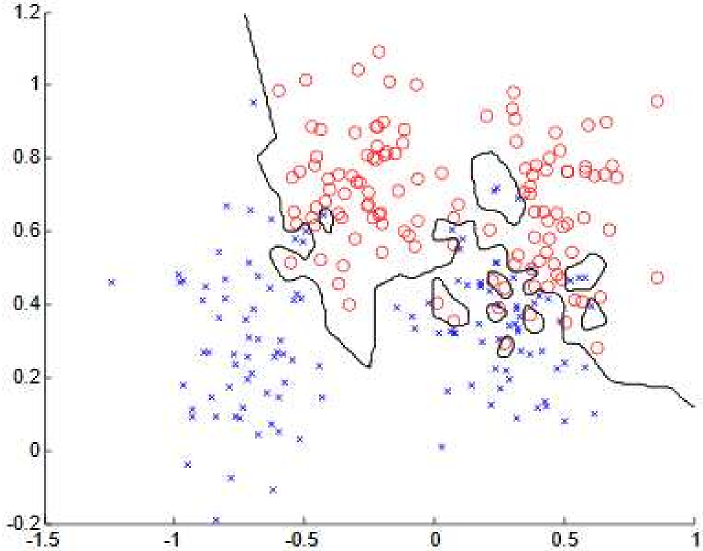




The classification error can be quantified as $\frac{1}{N}\sum_{i=1}^{N} 1_{[yi \neq f(xi)]}$

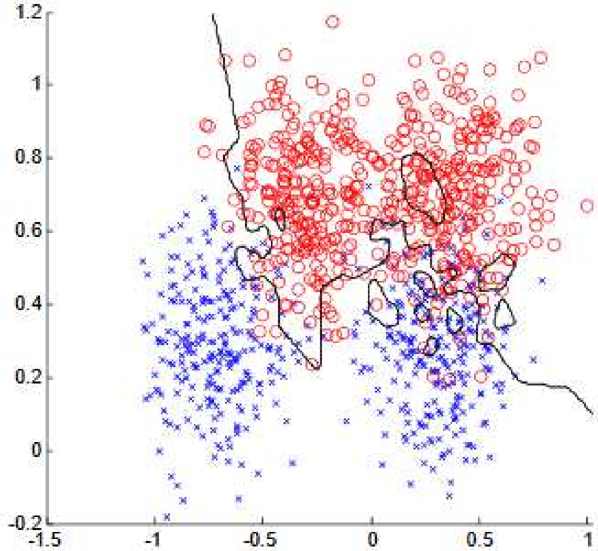
K = 1

Training data



error = 0.0

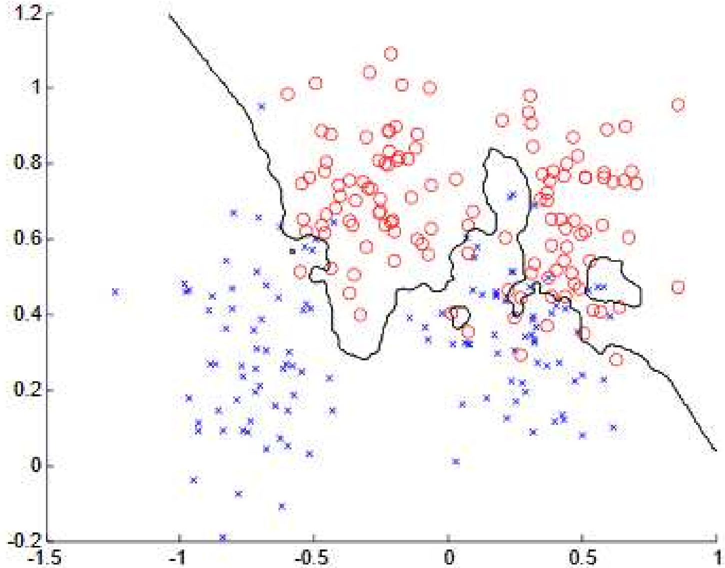
Testing data



error = 0.15

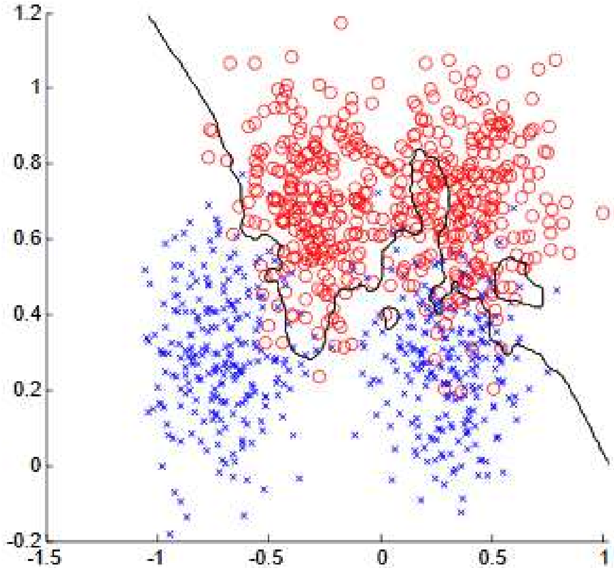
K = 3

Training data



error = 0.0760

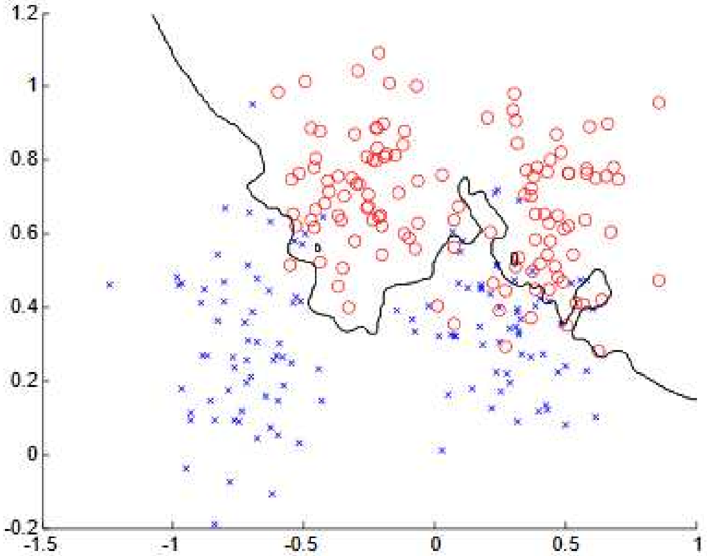
Testing data



error = 0.1340

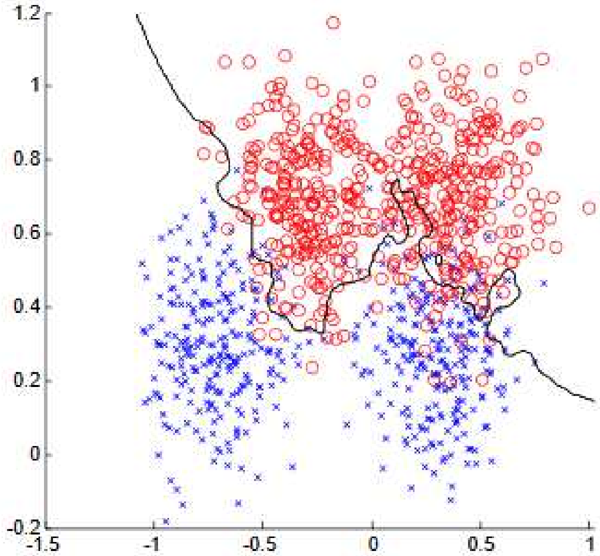
K = 7

Training data



error = 0.1320

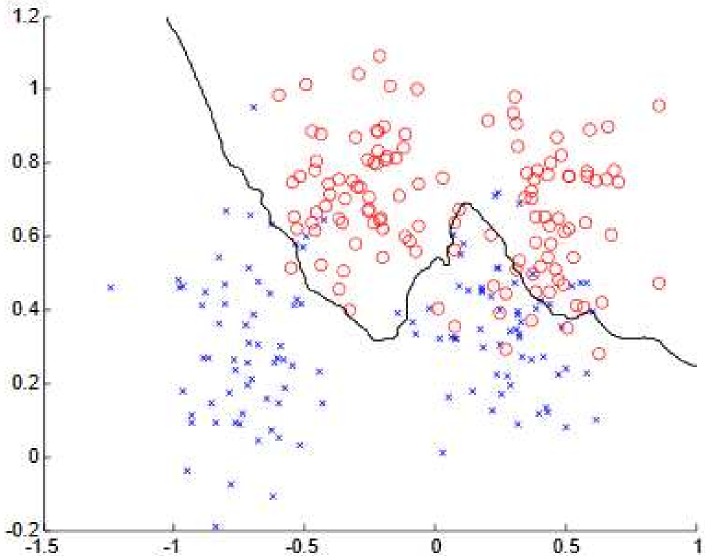
Testing data



error = 0.1110

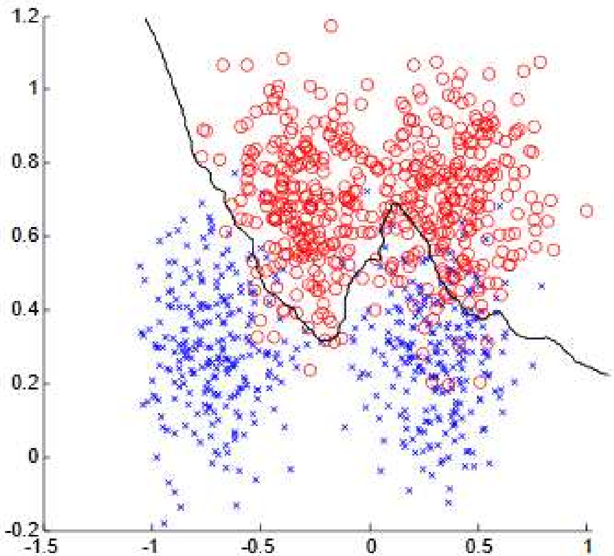
K = 21

Training data



error = 0.1120

Testing data



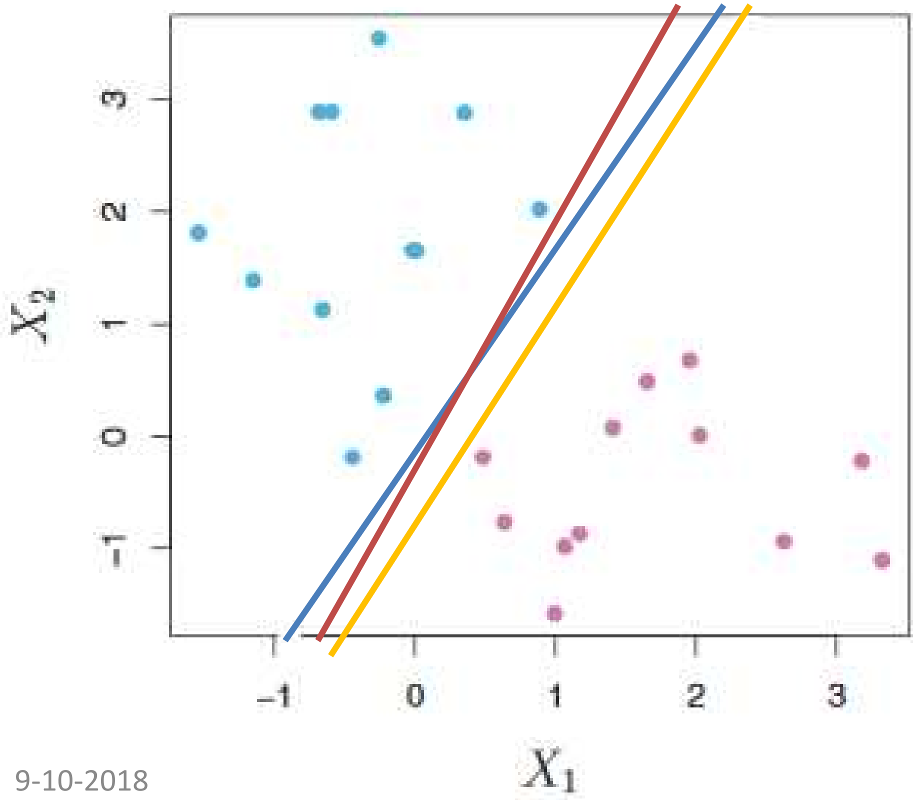
error = 0.0920

Summary k-NN

- Boundaries become smoother with increasing k
- Trade-off between overfitting (k=1) and generalization (k large)
- Training errors increase, but test errors might decrease
- Rule of thumb: choose k=sqrt(N)
- Non-linear
- Only one parameter k

Example: support vector machines

Goal: to optimally discriminate between blue and red using a (linear) classifier (hyperplane):



$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

blue:

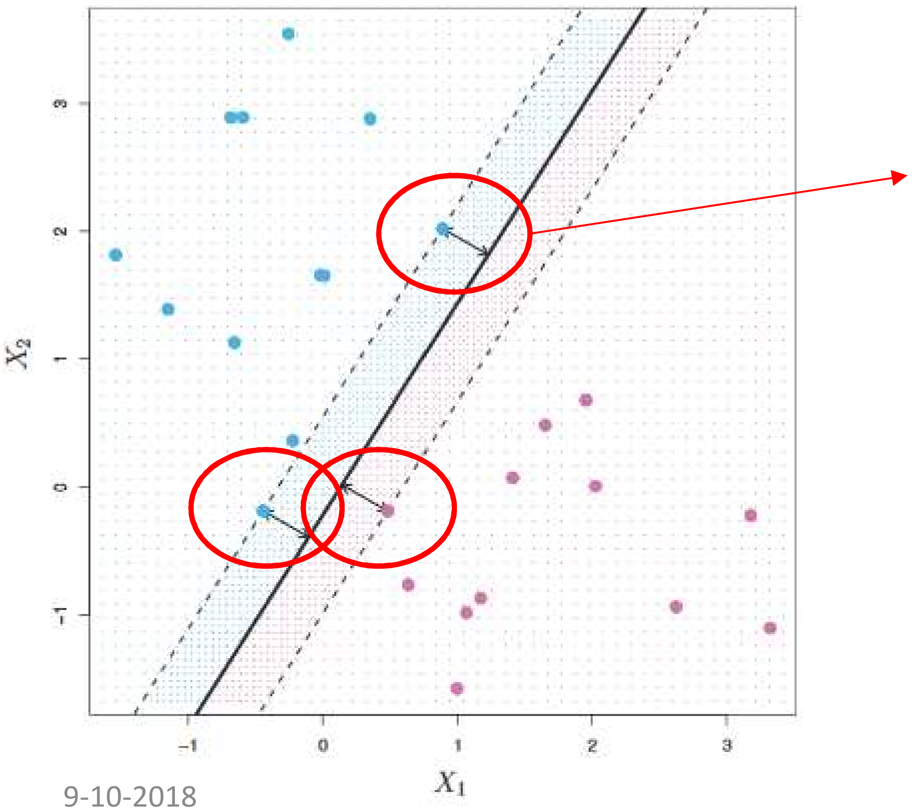
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$$

red:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 < 0$$

The maximal margin hyperplane

The hyperplane that is farthest from the training observations is the best classifier:



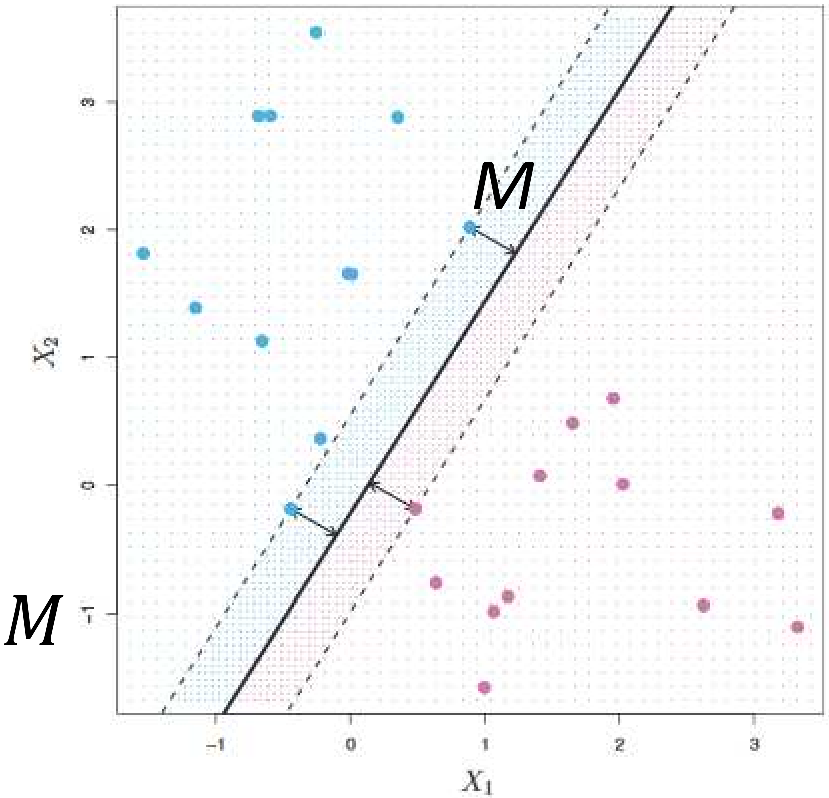
The three points are closest to the maximal margin hyperplane. They are called *support vectors.*

The optimal classifier is described only through these points!

The maximal margin hyperplane

Consider a set of observations $x_1, \ldots, x_n \in R^p$ with class labels $y_1, \ldots, y \in \{-1, 1\}$.

The maximal margin hyperplane solves the problem

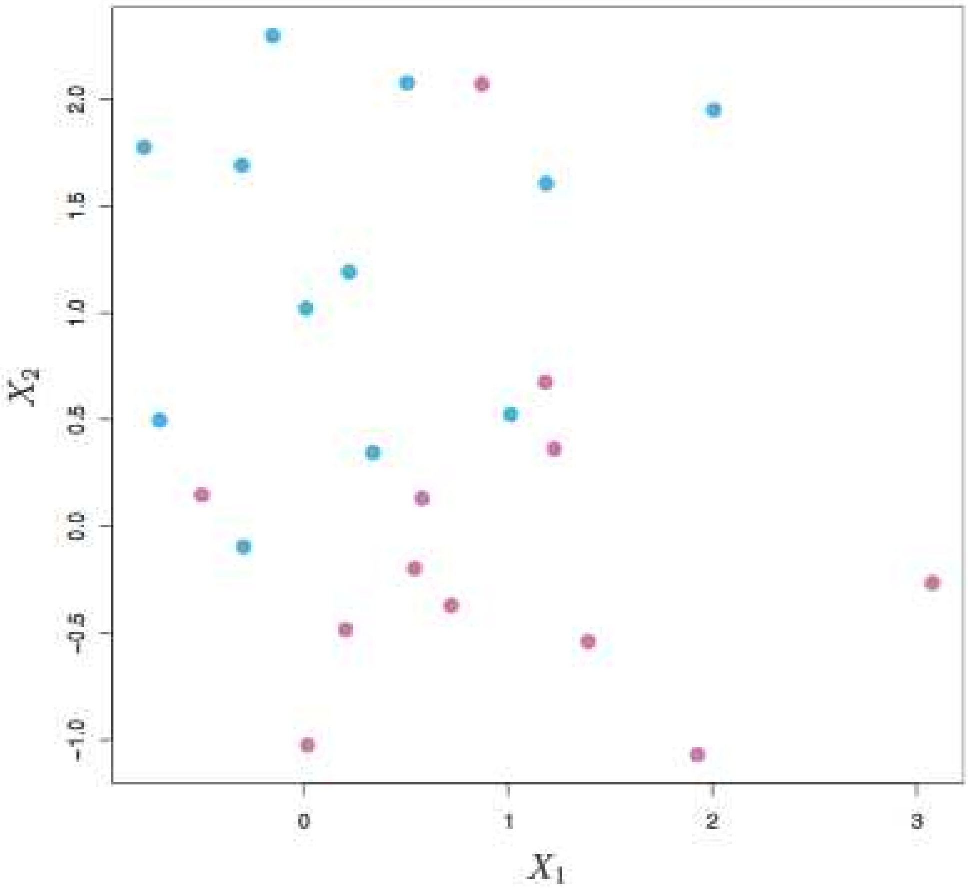


Maximize $M_{\beta_0, \beta_1, \ldots, \beta_p, M}$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1$,

$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p \, x_{ip}) \geq M$
$\forall \text{ i} = 1, \ldots, \text{n}$

The non-seperable case



The maximal margin classifier cannot be used here.
Solution: the support vector classifier (soft margin classifier)!
It classifies **most** of the observations correctly.

Maximization problem)

The support-vector classifier

Maximize $M_{\beta_0,\beta_1,\ldots,\beta_p,M}$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1$,

$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M\boxed{(1 - \epsilon_i),}$

$\epsilon_i > 0, \boxed{\sum_{i=1}^{n} \epsilon_i \leq C}$

This allows the observations to be on the wrong side of the hyperplane!

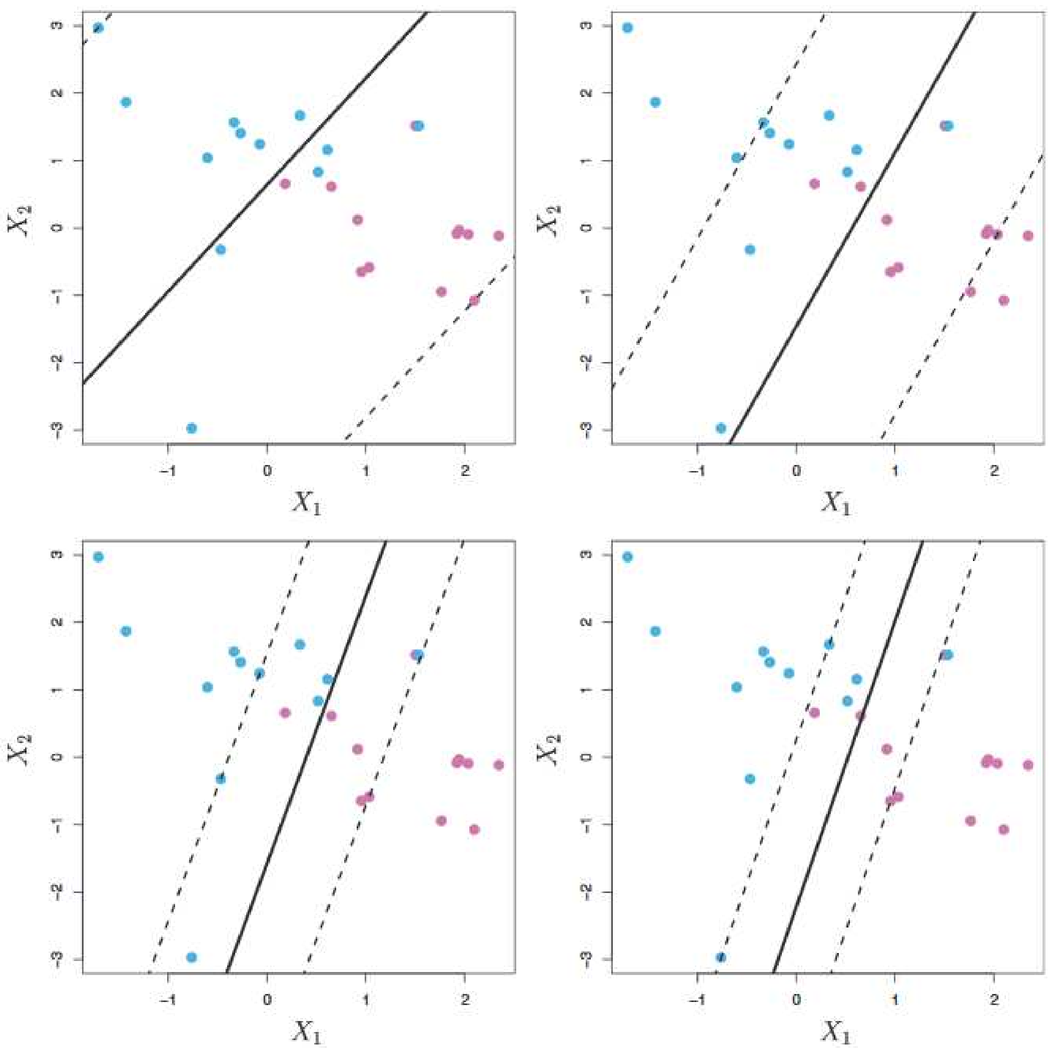
With tuning parameter $C \geq 0$

$\epsilon_i > 0$ →wrong side of the margin

$\epsilon_i > 1$ →wrong side of the hyperplane

C = # observations that can be on the wrong side of the hyperplane

The support-vector classifier

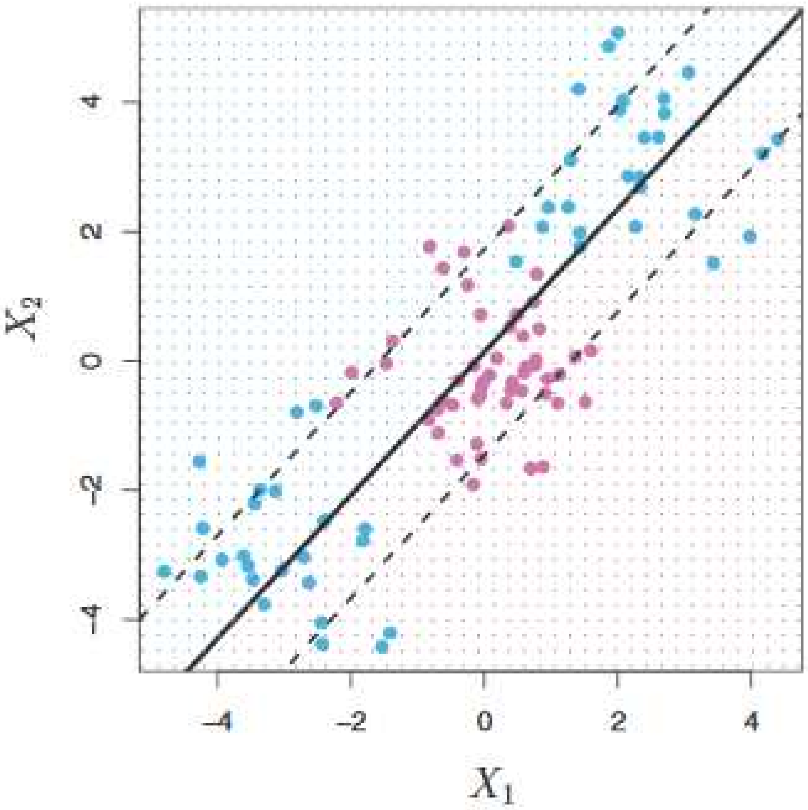


Support vector classifier on the same data, with different tuning parameters C from top left (biggest C) to bottom right (smalles C).

The **smaller** C, the narrower the margin M.

The **bigger** C, the more **robust** the solution.

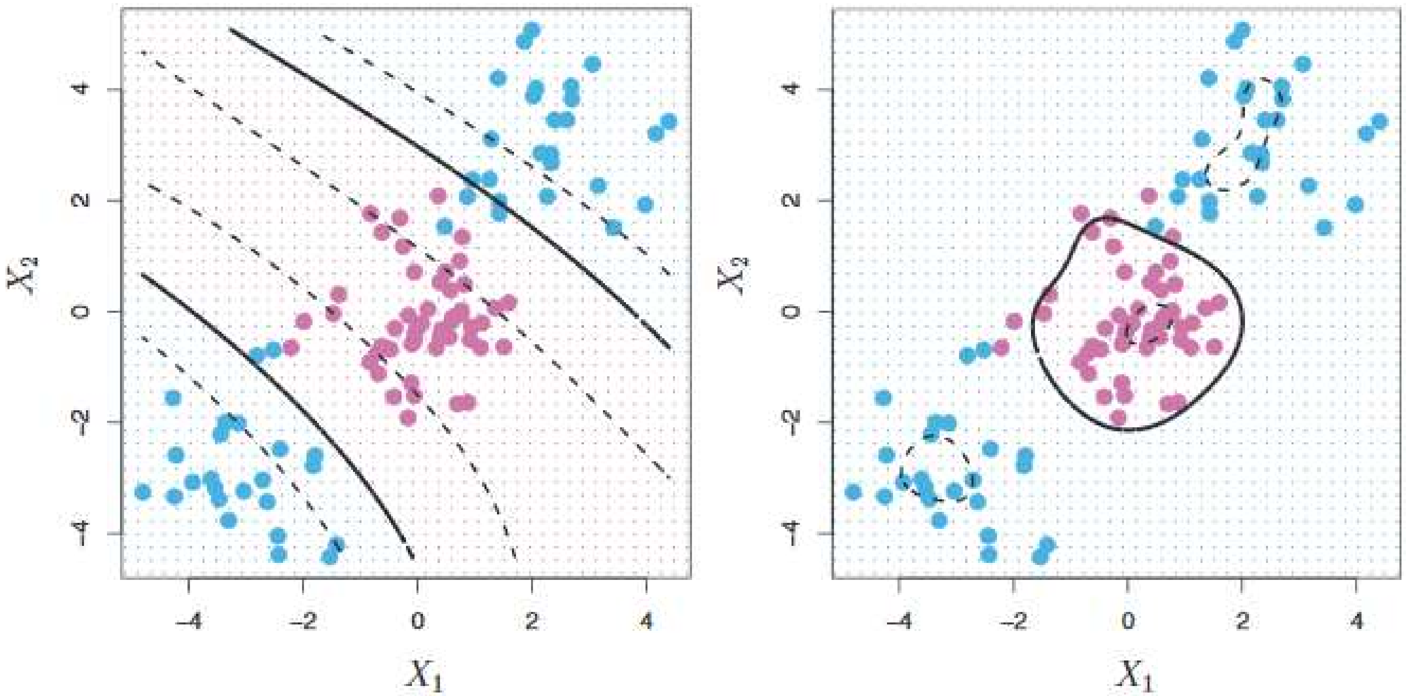
Support Vector Machines (SVM)



In some cases, linear classifiers perform very poorly.

The *support vector machine (svm)* uses so called *kernels* to address the non-linearity of the used boundary.
A kernel quantifies the similarity of two observations.

Support Vector Machines (SVM)



Polynomial kernel of degree 3

Radial kernel

Support Vector Machines (SVM)

- Computations are still very complex, therefore not outlied here.
- SVM is a dimension-reduction method, since it breaks down the classification problem to the support vectors only.
- SVMs can be extended to more than 2 classes!
- SVMs can be extended to continuous outcomes: *support-vector regression!*
- Support vector classifier versus logistic regression: similar results; svc behaves better in in more separated classes, while lr better in overlapping classes.

Validation methods

- Algorithms are trained on the training set and validated on the test set.
- -> Results depend on the random choice of the training set!
- Resampling methods are highly recommended to validate the results.
- Technique: repeated random drawings of different samples for sensitivity analysis and validation
- Two most common approaches:
  - Cross-validation
  - The Bootstrap

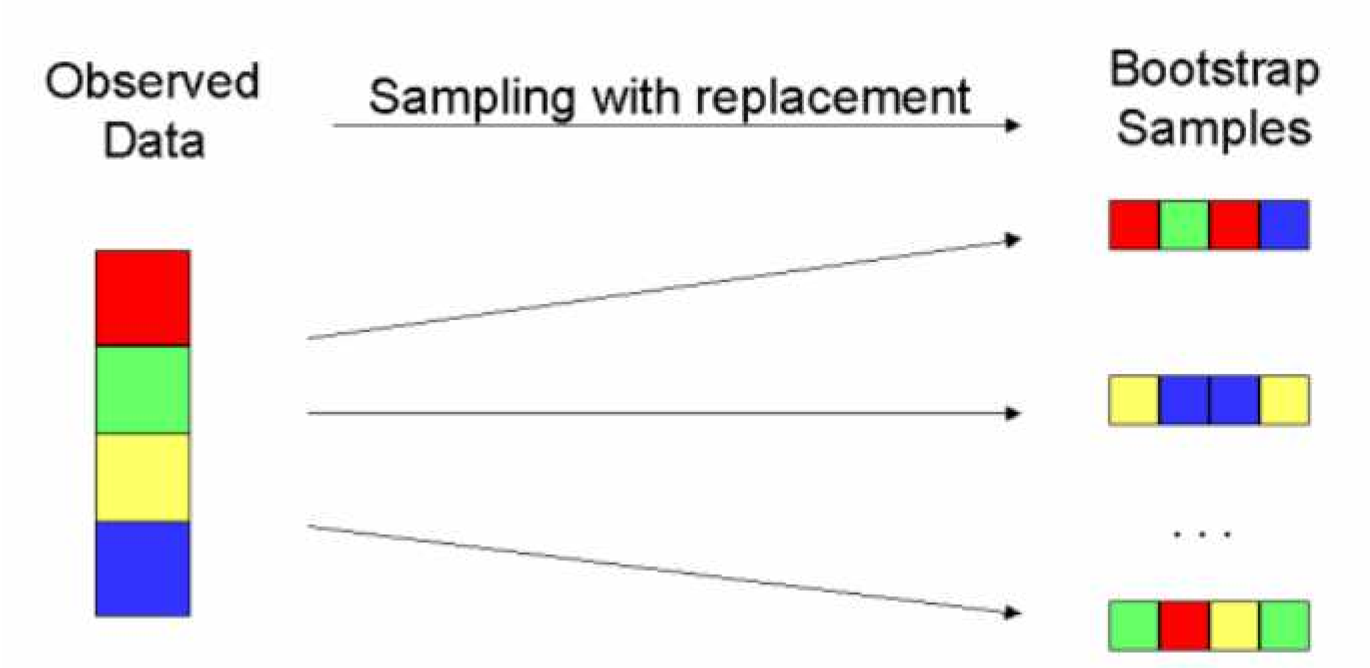
Cross-validation approach

K-fold cross validation

- Randomly divide the training set into k groups
- Train method on observations groups 2-k
- Calculate MSE_1 on group 1
- Repeat k times
- Estimate average MSE
- In practice, one performs k=5 or k=10

The Bootstrap

Generate muliple data sets by repeatedly sampling observations from the original data set.



Extremely powerful tool to quantify the uncertainty of an estimator!

The Bootstrap

- Standard errors are then calculated empirically from the n (lets say 1000) bootstrap samples.

- In supervised learning, it can be used to assess the variability of the estimates / predictions from a learning algorithm.

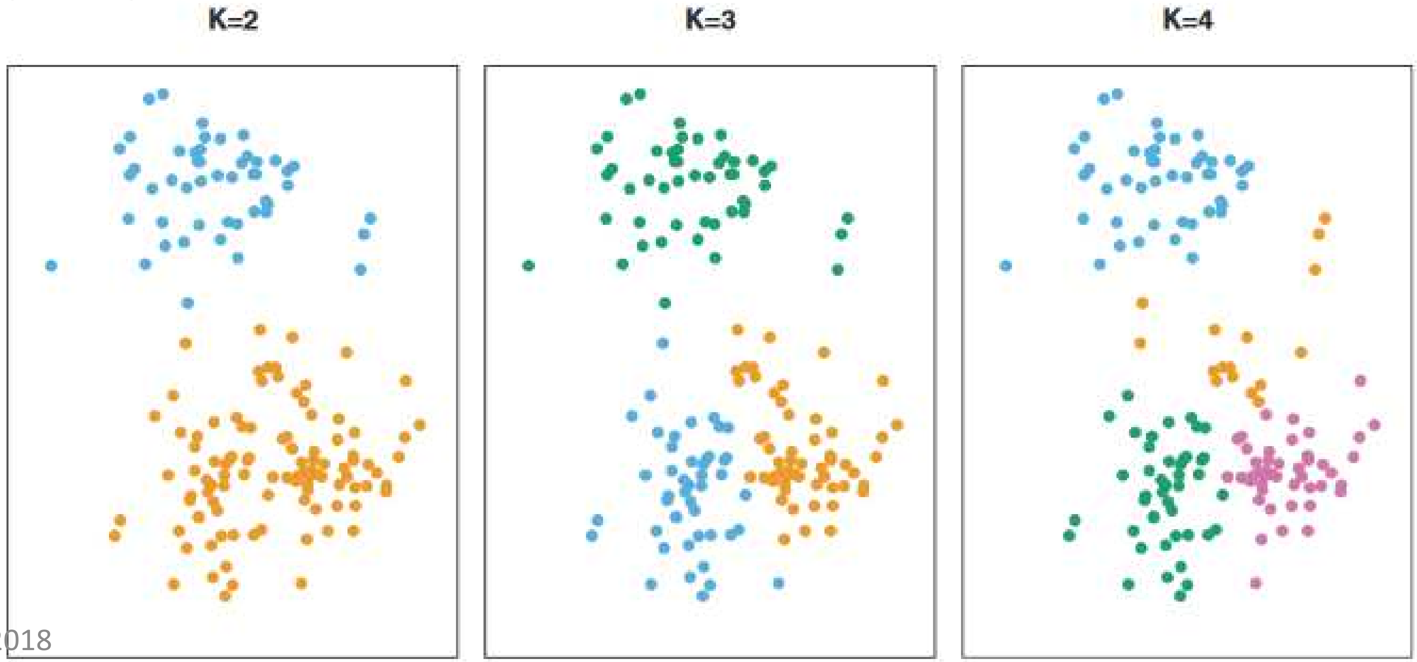
- Can be applied in almost all situations!

Two kinds of learning

- Supervised learning (‘ machine learning’)
  - Building a statistical model on an outcome
  - Use training data to optimize the algorithm
  - Apply algorithms to test data
- Unsupervised learning (‘ data mining’)
  - No outcome variable
  - Clustering and factoring, finding patterns in data
  - Dimension reduction

K-means clustering

- Goal: partitioning a data set into K distinct clusters which are most **homogeneous within** and most **heterogeneous between** each other:



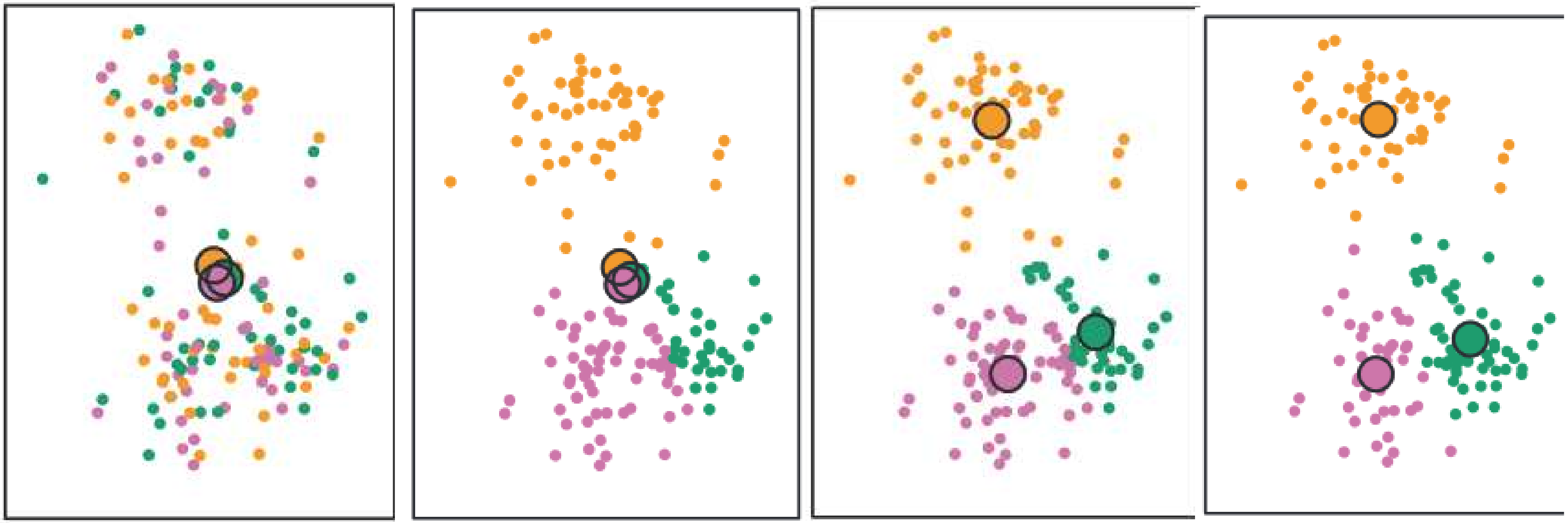
K-means clustering

Start with random cluster assignment for each observation.

2 iterative steps:

1. For each cluster, compute centroid
2. Assign each observation to the cluster whose centroid is closest (Euclidean distance).

Example K-means clustering (K=3)



random allocation, iteration 1, step 1	Iteration 1, step 2	Iteration 2, step 1	Iteration 2, step 2; Final result

More difficult: how to choose K?

Take home messages

- Big Data
  - Has to be cleaned, synchronized and processed before using
  - Human judgement is essential for this process!
  - Bear the chance to make better predictions
- Machine learning
  - constructs algorithms that can learn from data through repeated analyses on updated data.
  - Is not magic! It uses long-established statistical techniques

Take home messages

- **Supervised learning**: useful for prediction and estimation of an **outcome**
- **Unsupervised learning**: useful for general pattern recognition, without a specified outcome.
- Machine learning methods do not necessarily need big data!

Take home messages

General approach in **Supervised learning**:

- Split Data into training and test set
- Choose the optimal method for the training set
- Apply on the test set
- Apply validation techniques

General approach in **unsupervised learning**:

- No split into test- and validation set!

Literature

- *An Introduction to Statistical Learning, with Applications in R* (2013), by G. James, D. Witten, T. Hastie, and R. Tibshirani.

- *The Elements of Statistical Learning* (2009), by T. Hastie, R. Tibshirani, and J. Friedman.

Thanks for your attention ☺

When?	Where?	What?	Who?
Nov 13, 2018	Room 16	Using causal graphs to unravel statistical paradoxes	S. La Bastide
Dec 11 2018	Room 16	Non-prametrical tests	D. Postmus
Winter break January 2019			
Feb 12, 2019		Save the date!	