BIBA: Business Intelligence and Big Data

2018-10-17

Jens Ulrik Hansen

More modelling ...

Today's program

- Association rule mining
- Time series analysis
- Recap on modeling
- Work on synopsis hand-in

Association Rule Mining

Association Rule Mining

- A classical data mining task: *which items are frequently bought together?*
- Association Rule Mining is one approach
  - Find rules of the form: "customers who bought A and B, also bought C"
- Other methods for solving this and similar task: Recommender Systems or Collaborative Filtering
- Applications of association rule mining
  - Product recommendation (Amazon or Netflix)
  - Placing product next to each other in a physical store
  - Devising offers and advertisement

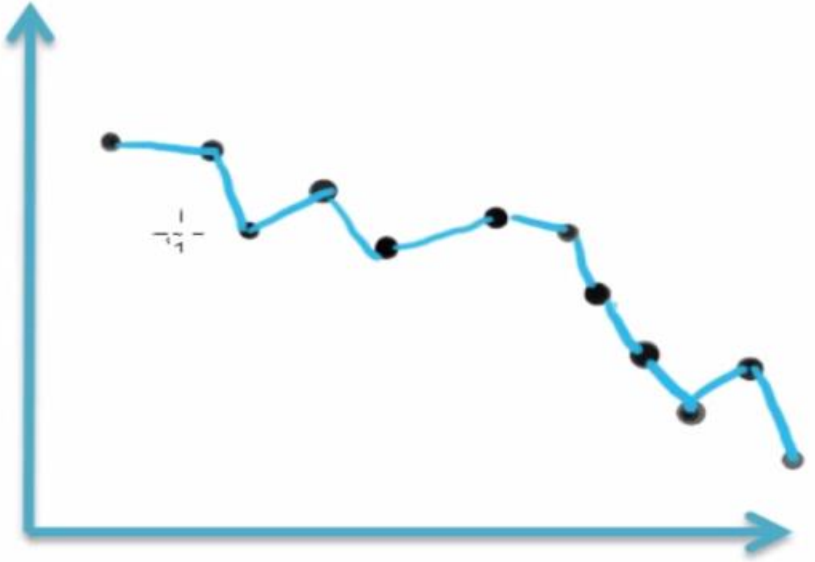
Association Rule Mining

- See the Jupyter notebook “6.1 Association Rule Mining.ipynb”

Time series analysis

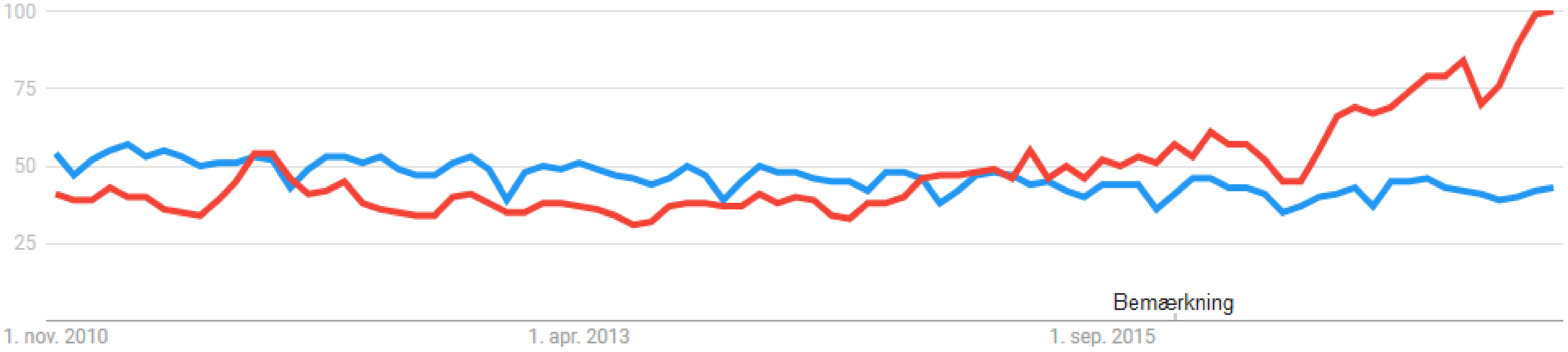
Time series and trends

- (See https://www.youtube.com/watch?v=ca0rDWo7IpI)
- **Time series data**
  - Data where each value is associated with a time stamp (year-month, year-month-day-hour-minute, …)
  - Examples
    - Number of costumers per day
    - Turnover by month
    - Average temperature by week
- **Positive and negative trend**
  - As time progress the values increase (decrease)
  - Note, a certain amount of data points are needed to properly detect a trend



Example of trends

- **Positive and negative trend example** (from Google Trend)
  - From Google Trend https://trends.google.dk/trends/explore?date=2010-10-10%202017-10-10&q=Business%20Intelligence,Artificial%20intelligence
  - Is there any trend in the timeseries?
  - Which one is BI and which is AI you think?

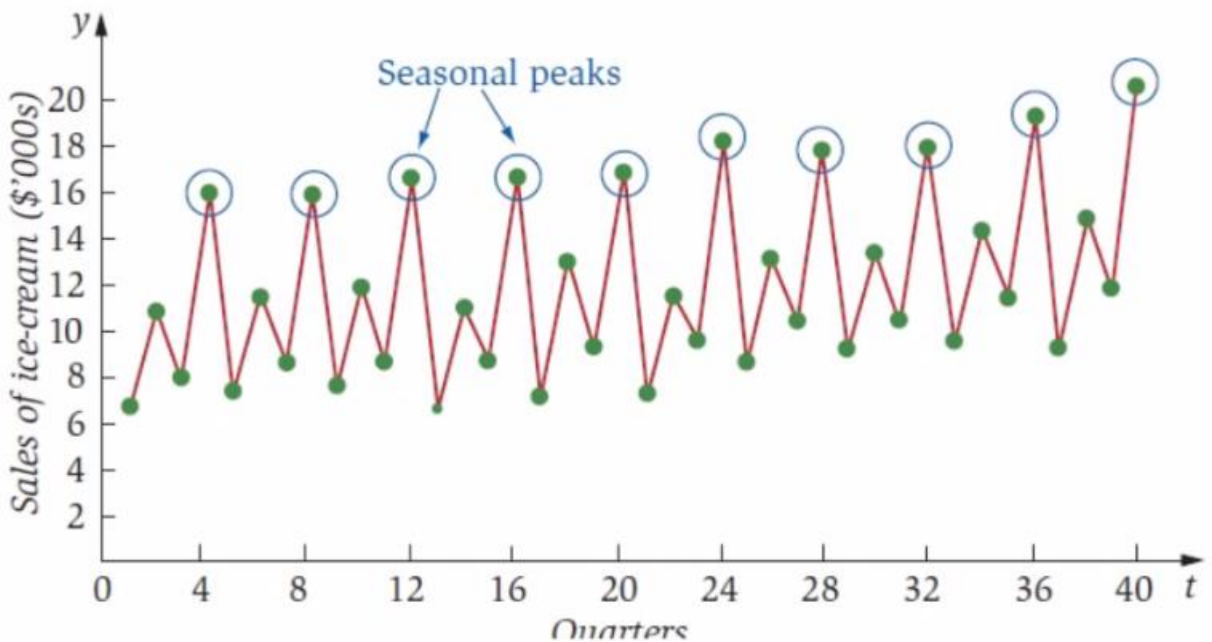


Useful insights from spotting trends

- Is there an increasing or decreasing tend in sales?
- Is there a trend in the number of new customers?
- Is there a trend to which media our primary customer segment is using?
- Has a particular vital event created a desired trend?
- ...

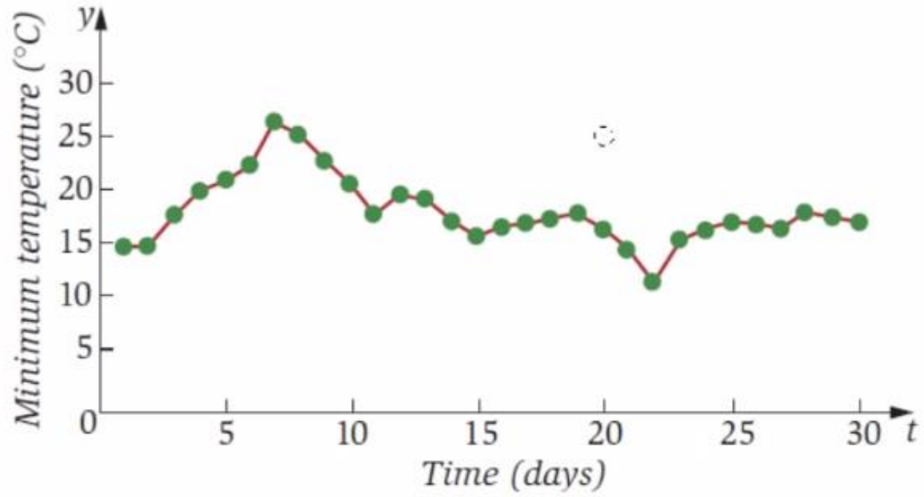
Simple variation in data

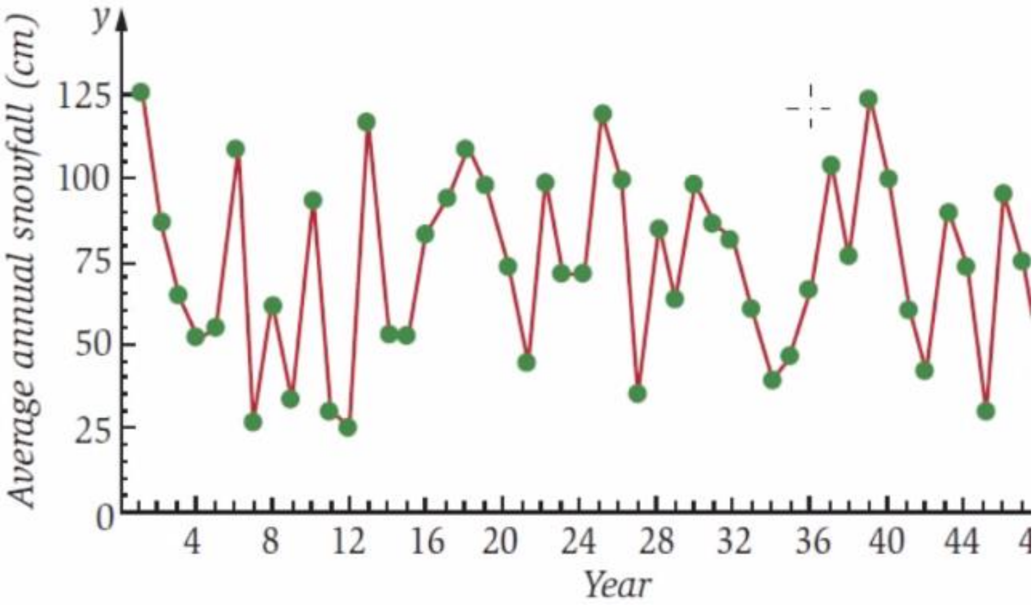
- **Variation in data**
  - Peaks and troughs in data
- **Seasonal variation in data**
  - Peaks and troughs occurs at regular (predictable) times
    - The same time every year
    - The same time every month
    - The same tree times a day
  - Examples
    - A peak in sales just after salary payment day (monthly) or just after payment of child benefit (quarterly)
    - A peak in sales of ice cream during summer
    - A trough in website visits during the weekends
  - Season is not just “spring-summer-fall-winter”



Simple variation in data

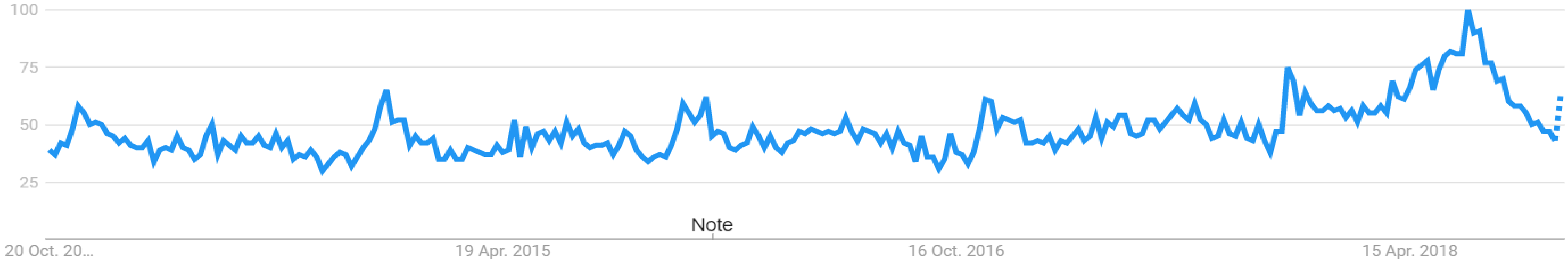
- **Cyclic variation in data**
  - Peaks and troughs occurs, but not at predictable regular times. The time periods between peaks is varying without a pattern
- **Random variation in data**
  - There is no visible pattern (There can still be a trend at the same time as random variation in data)

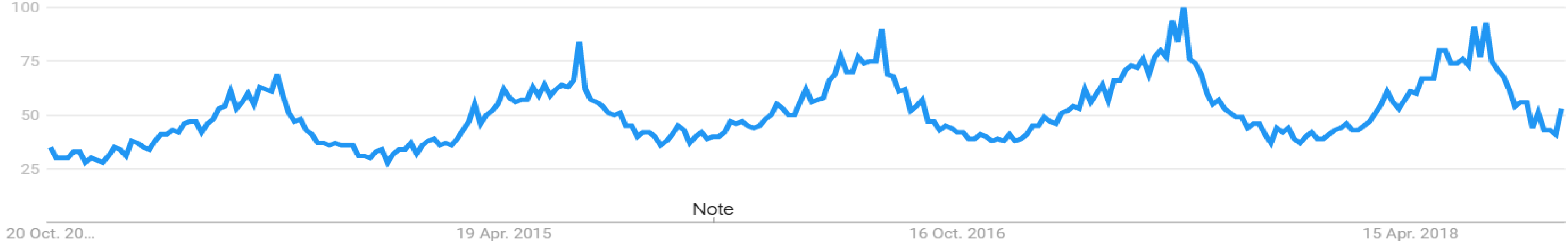




Examples of season in data

- Google trend examples again
- Can you guess the time series?





Useful insights from spotting season

- When do we sell the most?
  - When will our marketing campaigns have the biggest effect
  - Do we need to adjust production to match seasonal demand?
- When does our customers have the highest purchasing power?

Time series analysis in R

- For simple visual time series analysis in R, see the first part of the Jupyter notebook “6.2 Time series analysis.ipynb”
- For more advanced time series analysis in R, see the rest of the Jupyter notebook “6.2 Time series analysis.ipynb”

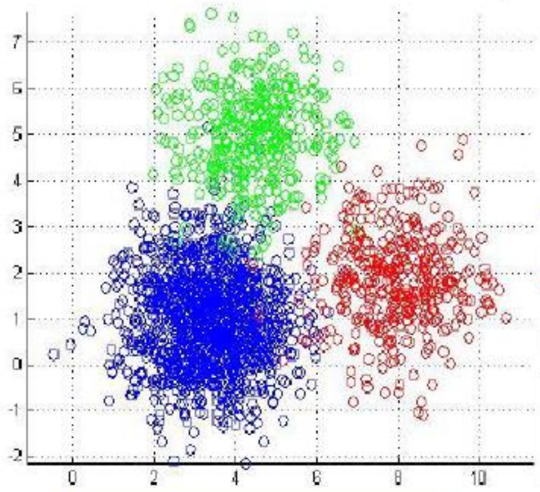
Recap on modeling

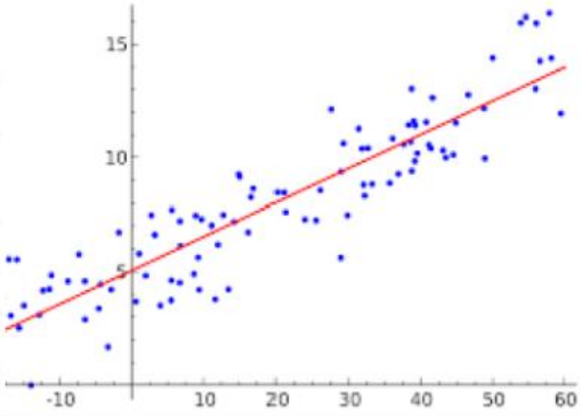
Recap on modeling

- Modeling can serve multiple purposes
  - Provide insight into particular phenomena or data
  - Discover patters in data
  - Predict future effects and event
  - . . . etc.
- Statistical/machine learning models
  - We try to learn general patterns from data
  - Three steps
    1. Define a family of potential models
    2. Select the model from this family that fits the data best
    3. Evaluate the model

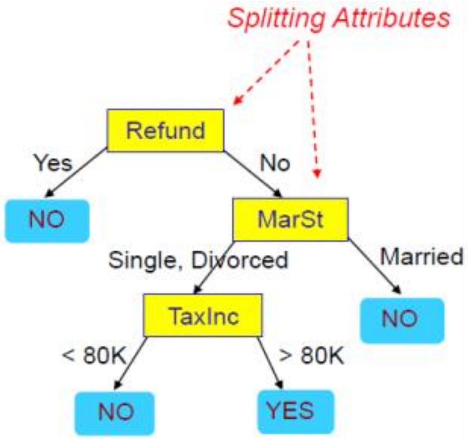
Types of machine learning

- ***Supervised learning*** – data contains values for what we want to predict
  - Regression (linear regression)
  - Classification (k nearest neighbor, decision tree)
- ***Unsupervised learning*** – data does not contain answers to what we want
  - Association Rule Mining (a prior algorithm)
  - Clustering (k-means clustering, Hierarchical clustering)
- ***Reinforcement learning***
- ***(Time series analysis?)***

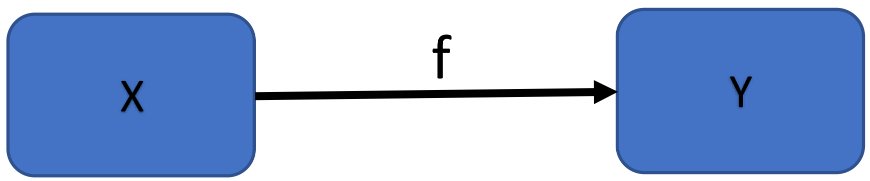




Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Supervised learning



- **We are trying to learn a function f such that Y = f(X1, X2, …, Xn), where Y is our *response/dependent variable* and X1, X2, …, Xn are our *predictor/independent/feature variables***
  - If we have such a function, we can always predict Y from any value of X1, X2, …, Xn
  - *Examples: Predicting house prices, sales, email or spam, churn, retention of employees, credit score, malignant or benign tumor, etc.*
- **Regression – when the response variable is continuous**
  - Linear regression, generalized linear models, penalized linear models
  - Decision trees, random forest, and boosting
  - Neural networks
  - Non-parametric regression
- **Classification – when the response variable is categorical**
  - k nearest neighbor,
  - Decision trees, random forest, and booting
  - Logistic regression
  - Support vector machines
  - Naïve Bayes
  - Neural networks

Evaluating supervised learning models

- Training and testing
  - The data set is split into a ***training set*** and a ***test set***
    - A 70% for training and 30% for testing is a common split
  - The *training set is used to train/fit the model*
  - The *test set is used to evaluate the model*
    - Warning! Adjusting a model based the evaluation on the test set can lead to overfitting
  - If one needs to adjust multiple parameters in the family of models, one can split the training set further into a *training set* and *a validation set*

Evaluating regression models

- Residuals
  - A model will never perfectly satisfy Y = f(X1, X2, …, Xn)
    - Y – the *true value* of the response variable (as given in the data)
    - Ŷ – the *predicted value* of Y based the predictor variables, i.e. Ŷ = f(X1, X2, …, Xn)
  - ***Residuals*** – the difference between the true value and the predicted value, i.e. Y - Ŷ
- Error measures
  - Mean Absolute Error (MAE): $1/n * \sum_i |y_i - \hat{y}_i|$
  - Mean Squared Error (MSE) : $1/n * \sum_i (y_i - \hat{y}_i)^2$
  - Root Mean Squared Error (RMSE): $\sqrt{(1/n * \sum_i (y_i - \hat{y}_i)^2)}$
- *R-square – the fraction of variability in the response variable explained by the model*

Evaluating classification models

- We only looked at the number of correct and in-correct predictions
- There is much more to it though!
- Check out concepts such as:
    - Confusion Matrix
    - False positives and false negatives, type I and type II errors
    - Accuracy and Precision
    - Sensitivity and Recall
    - Specificity
    - F1 score
    - Receiver Operating Characteristic (ROC) Curves and Area Under the Curve (AUC)

Feature selection/model selection

- ***Feature selection*** – selecting which predictor variable to use in a regression model (for instance)
- Feature selection is a trade-off
  - Including every feature can make the model overfit, insensible, computationally hard
  - Including to few make the model predict less well
- How do we compare two models with different features?
  - RMSE, (adjusted) R-square, Information criteria, etc. etc.
- Which models should we compare?
  - Exhaustive subset selection – go through all possible subsets
    - Computationally infeasible in many cases
    - Potential overfitting
  - There are numerous alternative approaches
- ***Model selection*** – we also compare different classes of models, not only different sets of features

Unsupervised learning

- There are no response variable (Y), i.e. true labels
- Less clear what types of unsupervised learning there is
- Examples of unsupervised learning
    - Clustering (k-means clustering, Hierarchical clustering, Graph-based models, etc.)
    - Association rule mining
    - Principal component analysis
    - Anomaly detection
- As we have no true labels, it is much harder to evaluate unsupervised learning models

Work on synopsis hand-in