Data Analysis and Knowledge Discovery

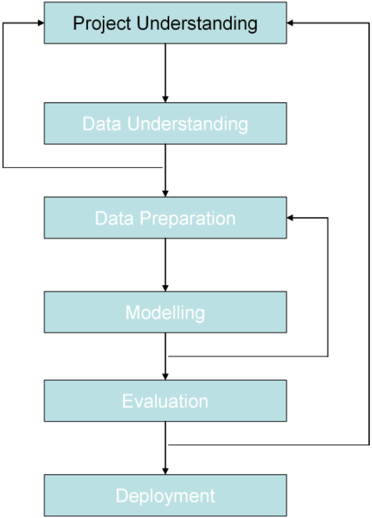
The Big Picture

Jukka Heikkonen

University of Turku
Department of Information Technology

Jukka.Heikkonen@utu.fi

Project understanding



Project understanding

- Initial phase of the data analysis project
- Problem formulation
    - Objectives
    - Potential benefits
    - Constraints and assumptions (a priori knowledge)
    - Risks
- Mapping the problem formulation to a data analysis task
- Understanding the situation (available data, suitability of the data...)
- Average time spent for project and data understanding within CRISP-DM: 20 %
- Importance for success: 80 %

Project understanding

- Project can easily fail based on misunderstandings, e.g.
    - customer does not understand what data the analyst needs, has implemented data gathering in ways that compromise the quality of data, fails to provide crucial information about the data that is obvious to them but not to the analyst...
    - analyst does not understand the domain and the data, fails to notice quality problems or inherent biases, these neglections result either in bad results, or a model that the analyst thinks works well, but does not actually work in real settings...
    - even if the analysis is correct, maybe the discovered patterns are uninteresting to the customer, or the customer does not understand how to utilize them
- Close collaboration with domain experts crucial
- Very beneficial, if you have someone who understands both the domain and data analysis
- Discuss, check assumptions, ask and encourage asking lots of "stupid questions"

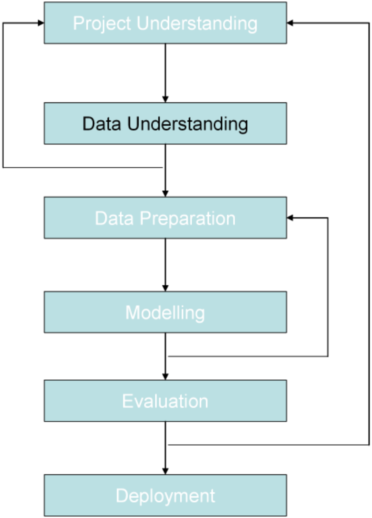
Determine analysis goal

- The primary objective must be transformed into a more technical data mining goal
- Determine data mining tasks
  - Supervised learning: you have a target to be predicted from inputs and existing (input - correct target) pairs to learn from
    - classification and regression
    - more complex prediction tasks (predict rankings, graphs, plans etc.) can be often reduced to these two basic tasks
  - Unsupervised learning: exploratory data analysis
    - clustering, association analysis, deviation analysis, visualizations...
- Consider also if the project goals could be achieved without data analysis methods
- Specify the requirements for the models that will be constructed by the data mining tasks

Technical constraints

- data set size: handling Big data requires computational and memory efficiency, parallelizations etc.
- dimensionality: certain methods are not suitable for handling very high-dimensional data (e.g. k-nearest neighbor, least-squares with no regularization etc.)
- input domain: real-valued or categorical data, structured data like images or graphs...
- computational requirements for model: if predictions need to be made fast (e.g. online advertisement), focus on models that are fast to compute (e.g. linear model needs O(d) multiplications and additions to compute, whereas k-nearest neighbor requires searching the neighbors etc.)
- Understandability: models based on simple rules, or linear models with only a few features may be interpretable by humans, more complex models usually not

Data understanding



Data Understanding

- Gain general insight on the data (independent of the project goal).
- Checking the assumptions made during the project understanding phase.
  (representativeness, informativeness, data quality, presence/absence of external factors, dependencies, ...)
- Checking the specified domain knowledge.
- Suitability of the data for the project goals.

Rule of thumb: never trust any data before some plausibility tests

Data quality

- Low data quality makes it impossible to trust analysis results: "Garbage in, garbage out"
- Mistakes made in the data are most often very difficult to recover by computational methods.

Accuracy: Closeness between the value in the data and the true value.

- Reason of low accuracy of numerical attributes: noisy measurements, limited precision, wrong measurements, transposition of digits (when entered manually).
- Reason of low accuracy of categorical attributes: erroneous entries, typos.

Syntactic and semantic accuracy

Data quality: completeness

Completeness is violated if an entry is missing.

- w.r.t. attribute values: Fraction of null entries for an attribute. Note that missing values are not always marked explicitly as missing, for instance in the case of default entries.
- w.r.t. records: Complete records might be missing because
  - three years ago SAP was introduced and not all customer data were transferred to the new system.
  - the data set is biased and non-representative.
    (A bank might have rejected customers with no income.)

Data quality: unbiased and representative

The data should always be unbiased and representative, i.e. it should contain all information about the inherent patterns and rules in the data.

In many applications we do not have that sort of data

- Machine condition monitoring: A lot of examples when machine in running normally. Sometimes not possible to obtain interesting data, such as in the case of nuclear power plant.
- Natural disasters: E.g. no earthquake data from a certain area where future earthquake probability should be estimated.
- Mortgage/insurance etc. analysis: Certain types of customers are totally missing, e.g. we may only have information about customers who have been granted a loan.

Data quality: unbalanceness and timeliness

Unbalanced data: The data set might be biased extremely to one type of records.

Example: Production line for goods including quality control. Defective goods will be a very small fraction of all records.

Timeliness: Are the available data up to date to be considered to be representative? This is related to the non-stationarity of the domain where only the recently collected data provide relevant information.

Example: Many industrial processes change dynamically and are non-stationary in nature. Hence too old data may not be much of use for analysing current and future states of the process.

Data visualisation

Data visualisation is one of the most important steps for

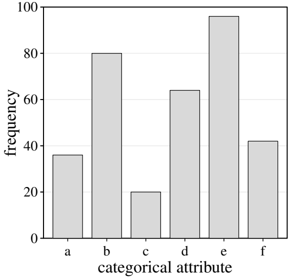
- Data understanding and preliminary data quality evaluation
- Learning from the domain

Visualisation as a test

- When visualisations reveal patterns or exceptions, then there is "something" in the data set.
- When visualisations do not indicate anything specific, there might still be patterns or structures in the data that cannot be revealed by the corresponding (simple) visualisation techniques.

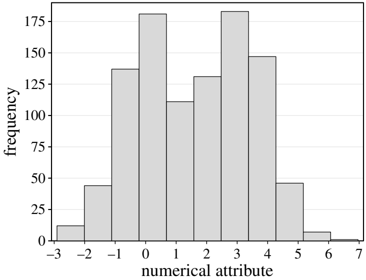
Bar charts

A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.

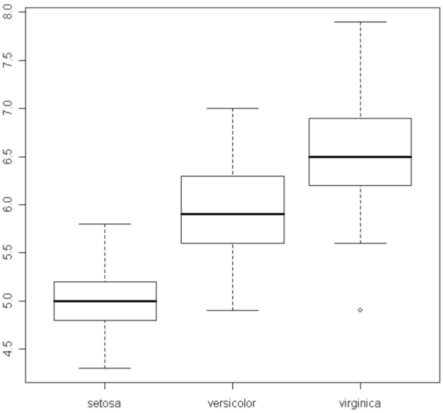


Histograms

A histogram shows the frequency distribution for a numerical attribute. The range of the numerical attribute is discretized into a fixed number of intervals (called bins), usually of equal length. For each interval the (absolute) frequency of values falling into it is indicated by the height of a bar.

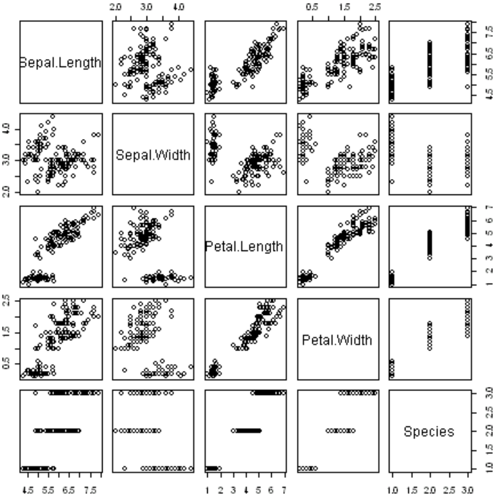


Boxplots



Scatter plots

Scatter plots visualise two variables in a two-dimensional plot. Each axes corresponds to one variable.



Methods for higher-dimensional data

Principle approach for incorporating all attributes in a plot:

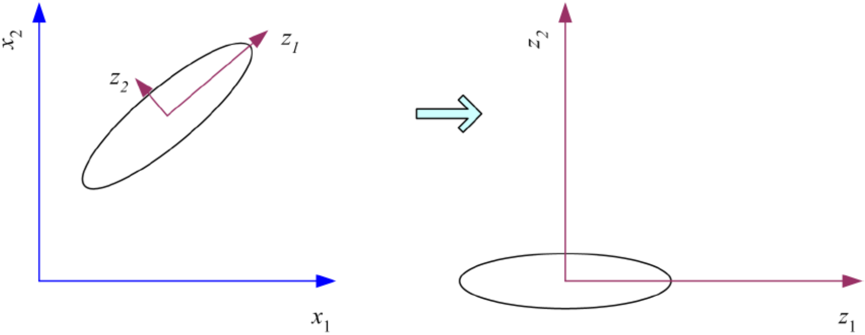
- Try to preserve as much of the “structure” of the high-dimensional data set when representing (plotting) the data in two (or three) dimensions.
- Define a measure that evaluates lower-dimensional representations (plots) of the data in terms of how well a representation preserves the original “structure” of the high-dimensional data set.
- Find the representation (plot) that gives the best value for the defined measure.

There is no unique measure for “structure” preservation.

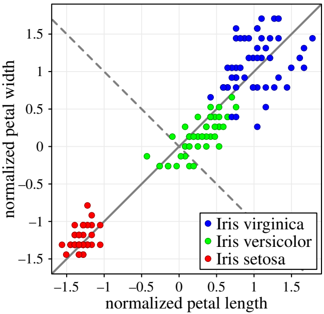
Principal component analysis

Principal component analysis (PCA) uses the variance in the data as the structure preservation criterion.

PCA tries to preserve as much of the original variance of the data when projected to a lower-dimensional space.



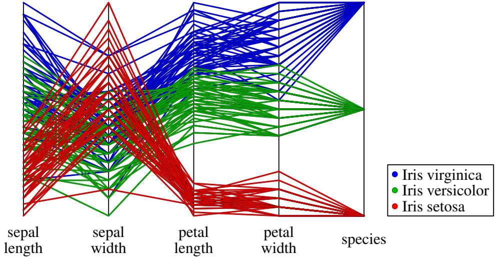
Principal component analysis

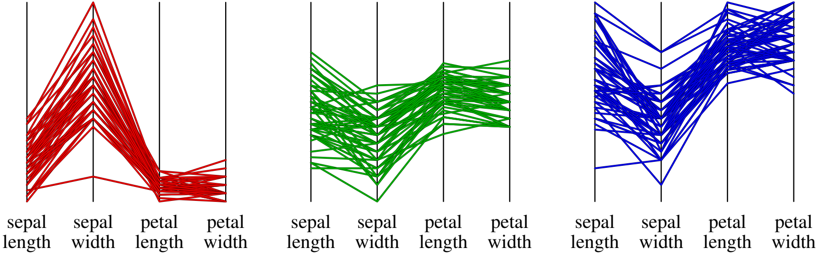


PCA applied to the Iris data set restricted to the (zscore normalised) petal length and width.

The principal components (1. solid line, 2. dashed line) are always orthogonal.

Parallel coordinates: Iris data

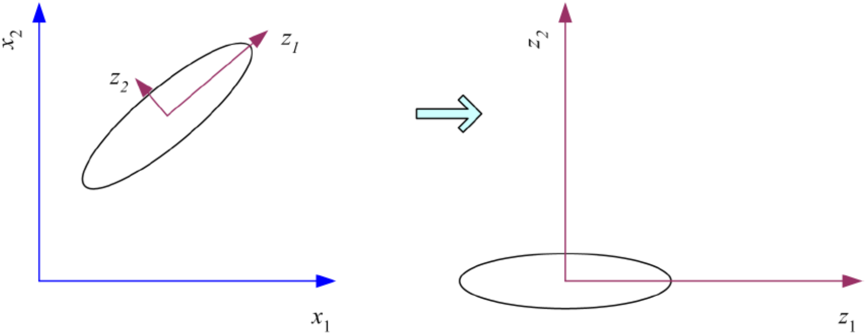




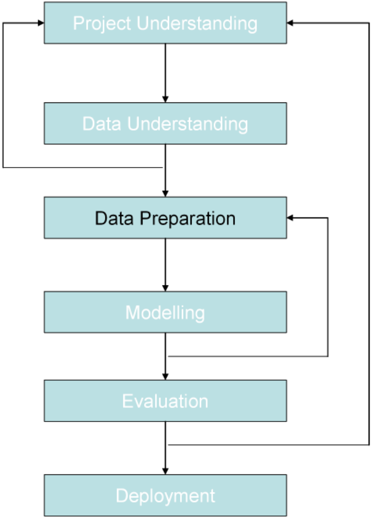
Principal component analysis

Principal component analysis (PCA) uses the variance in the data as the structure preservation criterion.

PCA tries to preserve as much of the original variance of the data when projected to a lower-dimensional space.



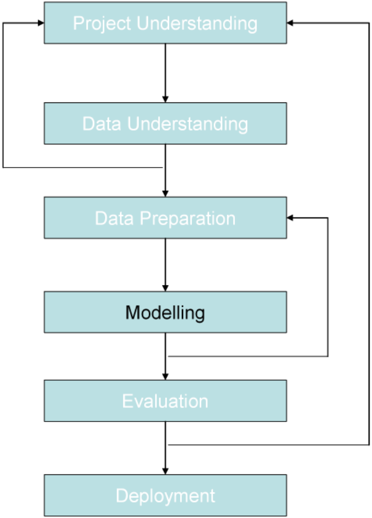
Data Preparation



Data Preparation

- already some data preparation has been needed for the tools used in the data understanding phase
- now done in a more principled manner, as a pre-processing step for the data analysis tool
- select a representative sample of instances
- generate suitable features
- solve data quality problems like missing or erraneous values
- normalization of features, possibly conversions like categorical to numerical, or vice versa depending on the method
- possibly feature selection or dimensionality reduction (e.g. PCA)

Modelling


Project Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Modelling

- what is the overall problem: classification, regression, clustering, association analysis, ...
- model structure: nearest neighbor predictor, linear model, IF-ELSE rule statemens, decision tree, neural network, clustering...
- scoring criterion: balance fit to the data with model complexity
- machine learning algorithm: tries to find a good model with respect to criterion
- model selection: different algorithms, or same algorithm with different parameters produce different models, how to choose best? Cross-validation a typical strategy

Nearest neighbour predictors

- As simple as it gets: memorize your training data, for new instances predict the majority class (or average value in regression) of its k-nearest neighbours
- Neghbours: k training instances having the smallest distance from the new instance
- No learning algorithm needed. We do not explicitly learn a model out of the training data, the data is the model.
- Surprisingly powerful, though has its limitations
- Choices needed: value of k, distance measure

Bayes' theorem, junk-mail example

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

- H: "Is junk-mail"
- E: "Words observed in the e-mail"
- $P(H|E)$: "Given the words, how likely is this to be junk-mail?"
- $P(E|H)$: "Given that a message is junk-mail, how likely it is to contain these words?"
- $P(H)$: "How likely is a randomly selected e-mail going to be junk-mail"
- $P(E)$: "How likely are these words to appear in a randomly selected e-mail"

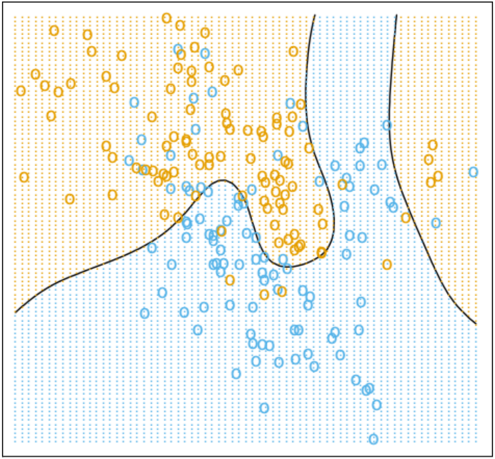
Regularized least-squares

Regularized least-squares, aka ridge regression:

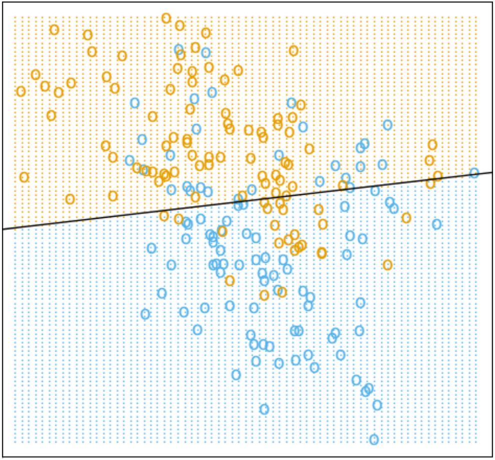
$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \left\{ \sum_{i=1}^{n} \left(\mathbf{w}^{\mathsf{T}} \mathbf{x}_i - y_i \right)^2 + \lambda \sum_{i=0}^{d} w_i^2 \right\}$$

- regularization term penalizes too complex models
- $\lambda > 0$ regularization parameter (can be chosen with cross-validation)
- unique solution, much more robust than basic least-squares fitting, especially for high-dimensional data

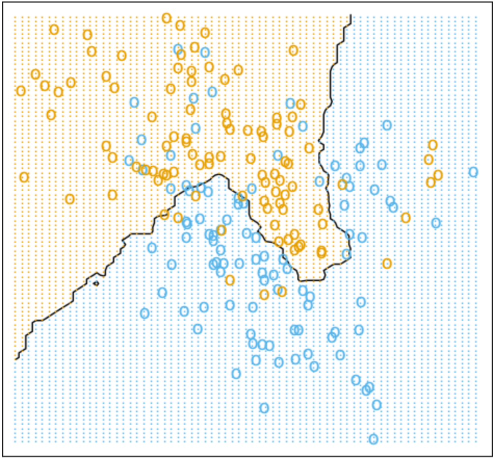
True model



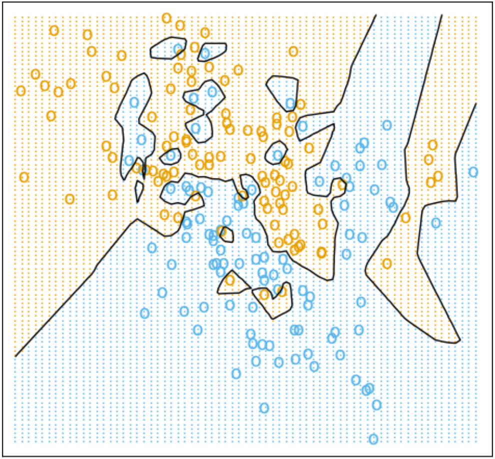
Underfitting (linear model)



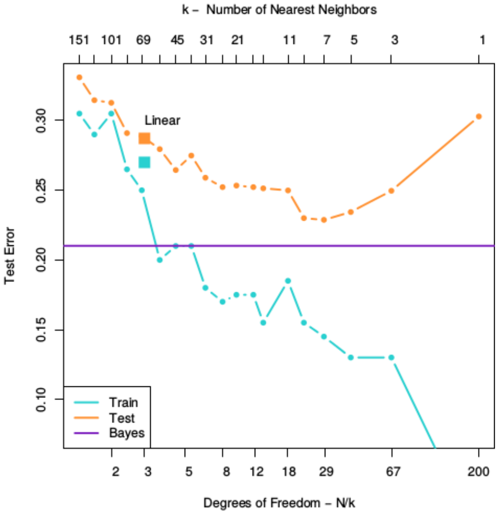
Reasonable model (15-nearest neighbour)



Overfitting (1-nearest neighbour)



k-nearest neighbour: choosing k



Moral of the story

- Two criteria need to be balanced in learning, fit to data (low error) and model complexity
- Underfitting: too simple models may not be able to capture the underlying concept well
- Overfitting: too complex models may simply model the noise in the data
- The more data you have, the more complex models you can afford to use
- Many theoretical approaches to defining what we mean by complex: regularization theory, minimum description length principle, Bayesian priors...

Controlling the complexity of $\mathcal{H}$

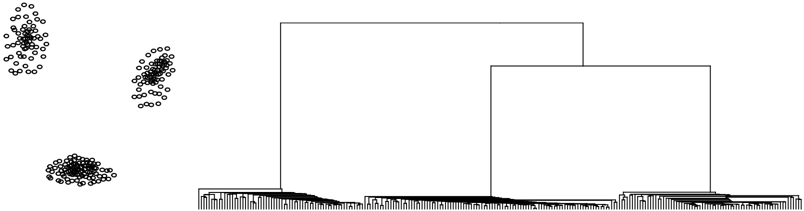
The complexity of the hypothesis set $\mathcal{H}$ is usually controlled with hyper-parameters.

- The max degree of polynomials on the above regression example
- The number of neighbours in the k-nearest neighbour method
- The regularization parameter value for many methods (regularized linear regression models, support vector machines,...)
- Depth of decision tree
- The number of layers and the number of neurons per hidden layer with neural networks
- You might also be comparing models produced by different algorithms altogether
- ...

Validation

- How to tell which model is the best?
- Validation means measuring the predictor's behavior on data points other than those in the training set (often known as the test set)
- One of the simplest and most practical approaches.
- Cross-validation: split data randomly to a training and test set, construct model on training set, compute performance (misclassification error / squared error / AUC) on test set, repeat e.g. 10 times, take average value
- Needed when you have so little data that you cannot afford a big separate test set

Hierarchical clustering



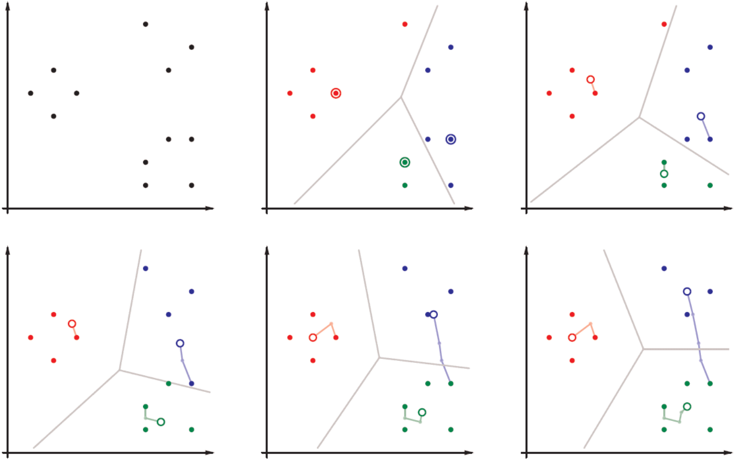
k-Means clustering

- Choose a number *k* of clusters to be found (user input).
- Initialize the cluster centres randomly
  (for instance, by randomly selecting *k* data points).
- **Data point assignment**:
  Assign each data point to the cluster centre that is closest to it (i.e. closer than any other cluster centre).
- **Cluster centre update**:
  Compute new cluster centres as the mean vectors of the assigned data points. (Intuitively: centre of gravity if each data point has unit weight.)

k-Means clustering

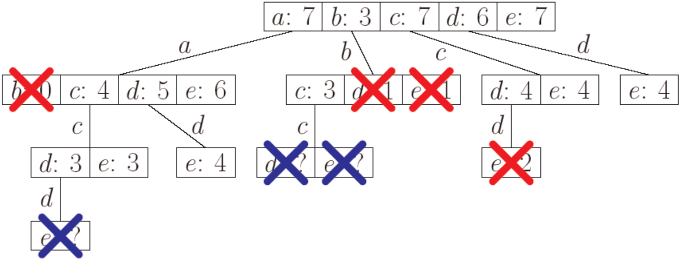
- Repeat these two steps (data point assignment and cluster centre update) until the clusters centres do not change anymore.
- It can be shown that this scheme must converge,
  i.e., the update of the cluster centres cannot go on forever.

k-Means clustering: Example

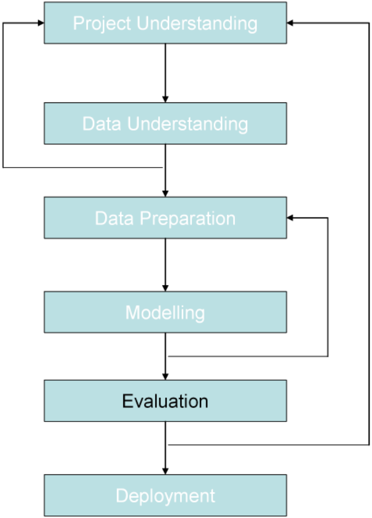


Searching for association rules

1: $\{a, d, e\}$
2: $\{b, c, d\}$
3: $\{a, c, e\}$
4: $\{a, c, d, e\}$
5: $\{a, e\}$
6: $\{a, c, d\}$
7: $\{b, c\}$
8: $\{a, c, d, e\}$
9: $\{c, b, e\}$
10: $\{a, d, e\}$



Evaluation

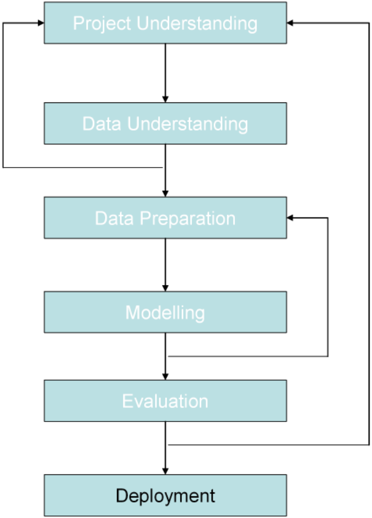


Project Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Evaluation

- Statistical evaluation for supervised learning: does the model predict well?
    - fairly straightforward for supervised learning; use cross-validation or separate test data
- Statistical evaluation more difficult for exploratory data analysis (unsupervised learning)
    - often not a well-defined criterion against which to measure
    - can measure some things like compactness of clusters, support and confidence for association rules etc.
- plausibility checks: does the model make sense to experts? (depends on the interpretability of model type whether this is feasible)
- have the business objectives been achieved, does it make sense to deploy the model?

Deployment


Project Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Deployment

- generate report, publish results, or implement model as software
- pilot project deploying the model
- monitor usage, does the model actually work as intended
  - programming bugs, or deeper flaws in the data analysis process or even the initial assumptions underlying the data may surface
  - be ready to react to these and go back to the drawing board if necessary
- consider plan for updating the model over time when the environment changes or more data is gathered

Preparing for the exam

- Electronic exam (see `tenttis.utu.fi` for details)
- Open 14.12.2015 - 22.01.2016
- Return also the second exercise set before taking the exam
- five questions

Preparing for the exam

- goal is not to memorize the material, but to understand the most central concepts
- answer in detail, explain the asked concepts and you can give simple examples where necessary
- more important to understand central concepts and methods than to memorize small technical details
  - important to understand what the regularized least-squares objective function stands for, but not the technical details of the proof on how the optimal solution is found
  - important to know what methods like k-nearest neighbour or k-means use distances for, not important to remember all the different distance measures
  - important to understand what a histogram represents and what effects choice of bin width could have, not important to remember the different formulas for determining bin widths
  - ...

Preparing for the exam

- Material on the course book
- If something in the book is not at all considered in the slides, it is not that important
- Things covered in the lectures (and slides) are important, even if they are not covered in the book
    - especially some of the classification and regression methods are not covered so well in the book
    - also, cross-validation and overfitting/underfitting are explained in my opinion quite badly in the book
    - if you missed the lectures or found some topics particularly difficult, Google for additional tutorials and explanations

Where next?

- Play around with the data analysis methods you have learned about on real data.
- This course has given you a starting point but we have just scratched the surface. One really only learns by doing.
- TKO_2096 Applications of Data Analysis, spring 2016 period III
- TKO_5519 Pattern Recognition, spring 2016 period IV