Data Analysis and Knowledge Discovery

Course summary

Antti Airola

University of Turku
Department of Information Technology

antti.airola@utu.fi

Preparing for the exam

- Electronic exam (see `tenttis.utu.fi` for details)
- Open until 31.1.2022
- You can try exam three times (as usual)
- Re-opened for March-April 2023
- Exam right requires that all exercises returned and accepted
- five questions

Preparing for the exam

- goal is not to memorize all 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 ridge regression 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
    - ...

Preparing for the exam

- Reading the course book "Guide to Intelligent Data Analysis" helpful for preparing to exam (see relevant chapters on Moodle)
- 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 the (cross-)validation lectures are not covered well in the book
  - if you found some central topics particularly difficult, Google for additional tutorials and explanations
- No questions that would require Python coding or remembering details of Python libraries
- High-level pseudocode descriptions of methods (e.g. cross-validation methods, K-nn) are relevant

Structure of the exam

Question 1: explain shortly following concepts: a) X, b) Y, c) Z,...

- six randomly chosen terms to be defined
- answer with a couple of sentences, try to demonstrate that you understand the meaning of the term (e.g. explain idea shortly in own words or give simple example)
- central basic concepts, visualization, pre-processing or analysis techniques, performance measures

Structure of the exam

Question 2 and 3: general essay questions

- demonstrate that you understand the different types of typical data analysis problems (supervised vs unsupervised, classification, regression, dimensionality reduction etc.)
- demonstrate you known the different phases of data analysis project (Crisp-DM as reference): project understanding, data understanding, data preparation, modeling, evaluation, deployment
- what are the goals of these different phases, what kinds of decisions need to be made, what kind of techniques are available, practical examples...

Structure of the exam

Question 4 technical question

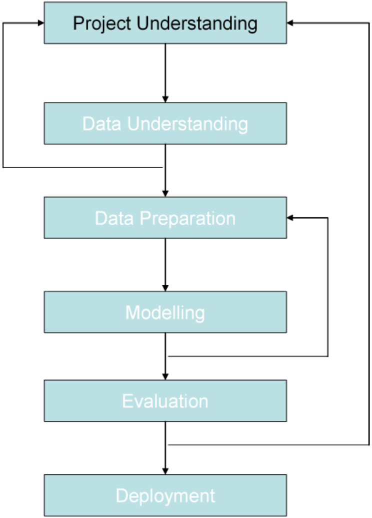
- demonstrate in-depth understanding of central data analysis methods
- principal component analysis, k-nearest neighbors, ridge regression, k-means clustering, hierarchical clustering, association rule mining
- detailed proofs not needed (though always appreciated!), but you should understand what problems the methods solve and how, have an idea of how the training algorithm works and what the method would do on simple toy examples

Structure of the exam

Question 5: validation

- how to verify, if your trained supervised learning method can predict well on new data
- do you understand what over- and underfitting are, what model selection and validation are about
- why do we need separate test data? common problems and how to overcome them?
- method of cross-validation, how is it used and why is it used?

Project understanding



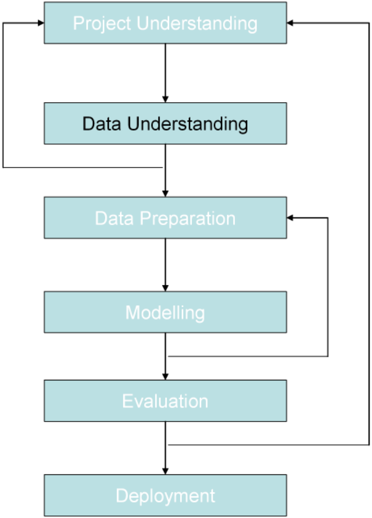
Project understanding

- 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 %

Test your understanding: Project understanding

- CRISP-DM model
- What kinds of goals could a data mining project have, examples?
- What types of methods are available for solving different types of problems?
- How can different requirements or desirable properties for the model affect our choice of method?
- How can a project go wrong from the very start?

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

Test your understanding: Data Understanding

- What is the main goal of this phase of project?
- What types of attributes are there, examples?
- How can you represent categorical variables as numerical?
- What are some typical problems with data quality?
- How do different visualization techniques work, what can they reveal about the data?

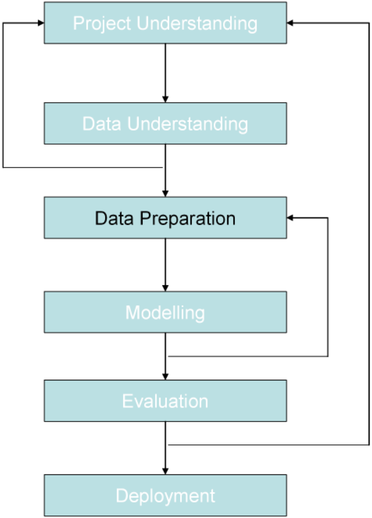
Test your understanding: Data Understanding

- What does z-score standardization do to data? Why might we need it before analysis?
- What is dimensionality reduction, how is it similar / different from feature selection?
- What structure of the data does PCA try to preserve?
- How to select the number of principal components to use?

Test your understanding: Data Understanding

- What are outliers, how can you detect them?
- What are missing values?
- Are some types of missing values more difficult to deal with than other?

Data Preparation



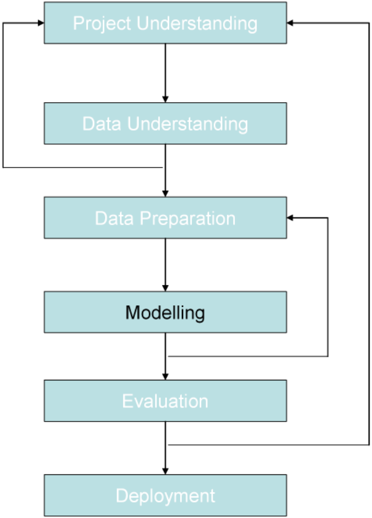
Data Preparation

- 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
- ideally, you should not use any statistics calculated from test data for pre-processing
  - sklearn: separate fit and transform functions for data dependent preprocessing methods
  - can be combined also with cross-validation, fit calculated only from training folds

Test your understanding: Data Preparation

- Overview, things typically done in the data preparation phase?
- What is feature extraction?
- Why do feature selection?
- How to deal with missing values?
- Why can it be beneficial to normalize feature values?

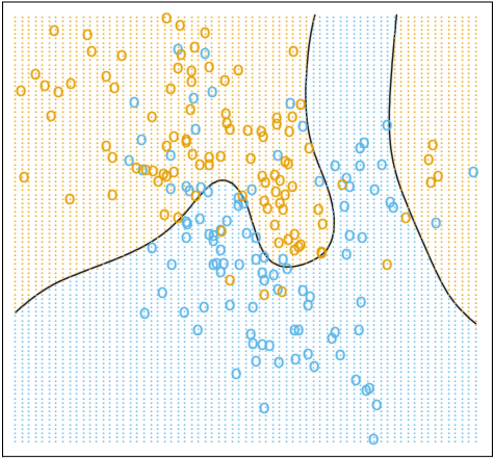
Modelling


Project Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

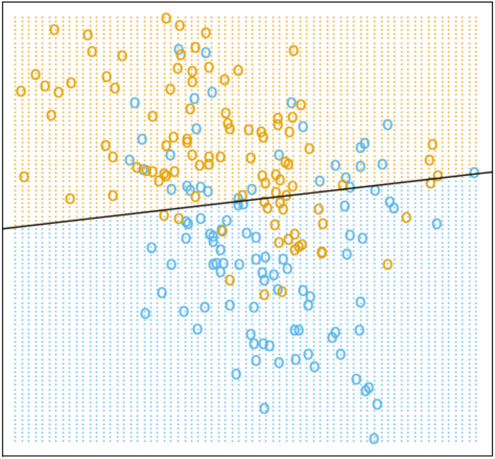
Modelling

- 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

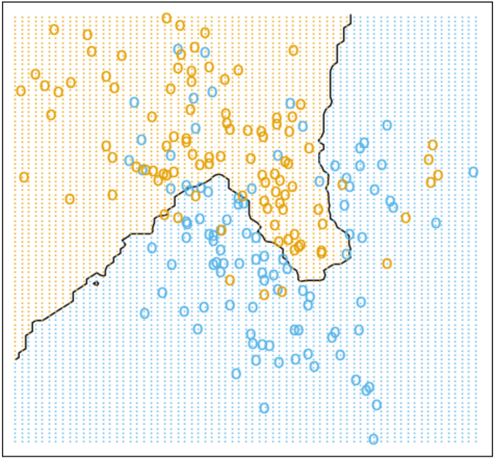
True model



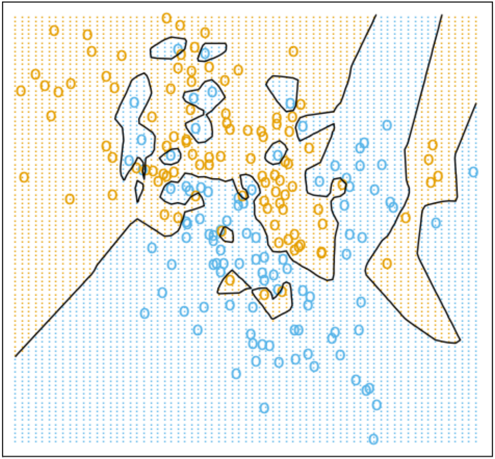
Underfitting (linear model)



Reasonable model (15-nearest neighbour)



Overfitting (1-nearest neighbour)



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
- ...

Test your understanding: Supervised learning

- What is supervised learning?
- What is the difference between classification and regression?
- What kinds of problems could you solve with these methods?
- What is overfitting and underfitting?
- Why training set error / performance is unreliable?

Test your understanding: Supervised learning

- How does k-NN classification / regression work?
- How complicated functions can it learn?
- What parts of the method can you modify?
- How does choice of k affect how the method works, how to select this value?
- How do you control model complexity with k-NN
- Computational and memory costs of k-NN?
- When can the method fail?

Test your understanding: Supervised learning

- How does ridge regression work?
- How complicated functions can it learn?
- What does the objective function of ridge regression measure?
- How do you control model complexity for ridge regression?
- How does a linear model work, what does it mean if a coefficient in the learned **w**-vector is large/small/zero?
- What kind of algorithm do you need for training ridge regression?
- Can you name any other linear regression / classification methods?

Test your understanding: Supervised learning

- What is model selection?
- What is model evaluation / assesment?
- What is generalization error? (or more generally, expected squared error / F-score / AUC /... on new data)
- Why do we need separate test data?
- Bias and variance in error / performance estimation?
- How to combine model selection and final evaluation?

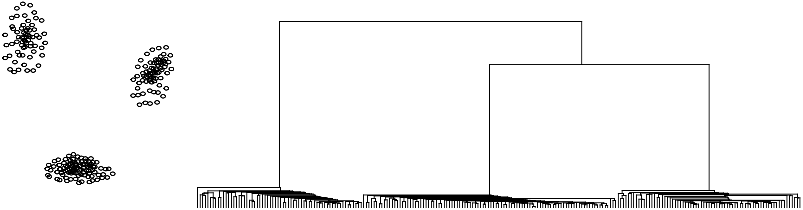
Test your understanding: Supervised learning

- What is cross-validation
- Why and when do we need it?
- Leave-one-out and K-fold CV?
- How to combine model selection and final evaluation when using cross-validation?

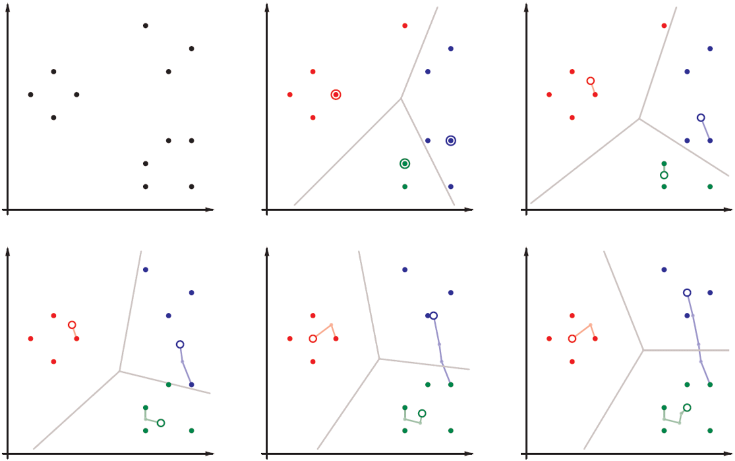
Test your understanding: Supervised learning

- What performance measures are out there for regression and classification?
- What limitations does misclassification rate / classification accuracy have as a performance measure?
- What are cost and confusion matrices?
- What are true positives, false positives etc., how can one calibrate classifier to achieve different tradeoffs between them?
- What is a ROC curve, how to interpret area under ROC curve?

Hierarchical clustering



k-Means clustering



Test your understanding: Unsupervised learning

- What is unsupervised learning?
- What is clustering?
- What is association analysis?
- What kinds of problems could you solve with these methods?
- Why is evaluation of unsupervised learning results more difficult, than for supervised learning?

Test your understanding: Unsupervised learning

- Main idea behind hierarchical clustering?
- How does the agglomerative hierarchical clustering algorithm work?
- How can you measure dissimilarity between clusters, what kind of cluster structure do different choices lead to?
- What is a dendogram?
- How can you choose the number of clusters for hierarchical clustering?

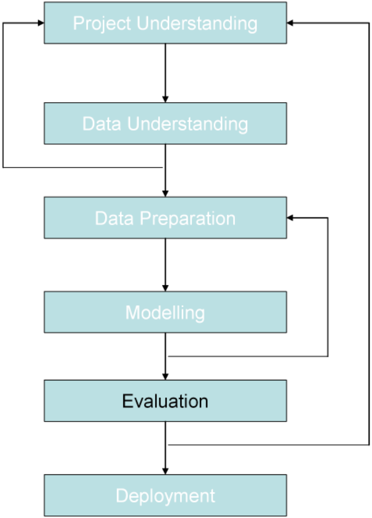
Test your understanding: Unsupervised learning

- Main idea behind k-means clustering?
- How does the k-means clustering algorithm work?
- How can you select k?
- Can you invent an example where k-means would fail?

Test your understanding: Unsupervised learning

- Main idea behind association analysis?
- Item sets, support, confidence of a rule?
- How do the algorithms for finding association rules work?

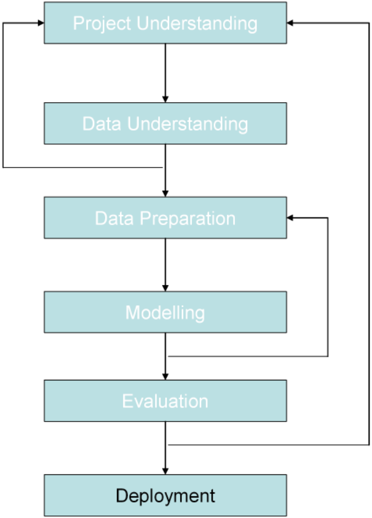
Evaluation


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

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.
- Several data science related courses during the spring
  - Evaluation of Machine Learning Methods (III)
  - Machine Learning and Pattern Recognition (III)
  - Introduction to Deep Learning (III / IV)
  - Introduction to Human Language Technology (IV)
  - Computer Vision and Sensor Fusion (IV)
  - Biosignal analytics (III)
- Internet is full of free MOOCs (Massive Open Online Course), tutorials etc.

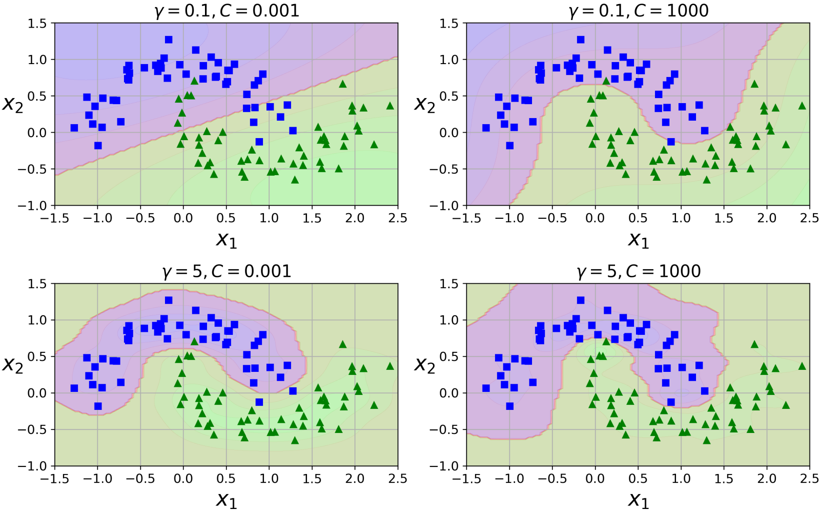
Where next?

- Several research groups at UTU pursuing data science and machine learning research
- Every summer trainee positions open
- Doctoral Programme in Technology (DPT)

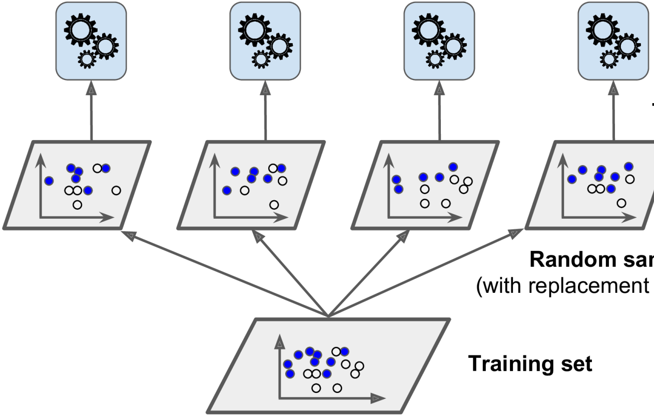
Recommendations for self-study

- Aurélie Géron: Hands-on Machine Learning with Sciki-Learn, Keras & TensorFlow, 2nd Ed. (2019)
- Many things we did not cover in this course

Chapter 5: support vector machines



Chapter 7: ensemble learning and random forests



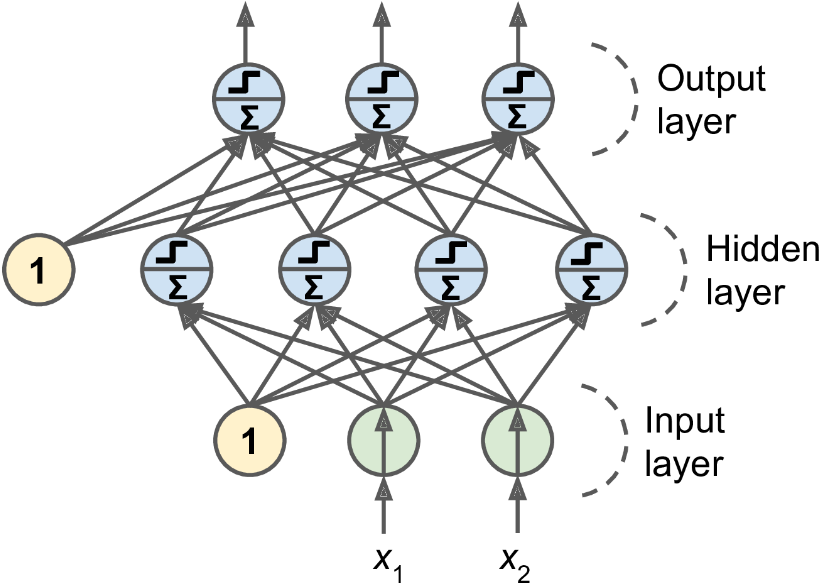
Predictors
(e.g., classifiers)

Training

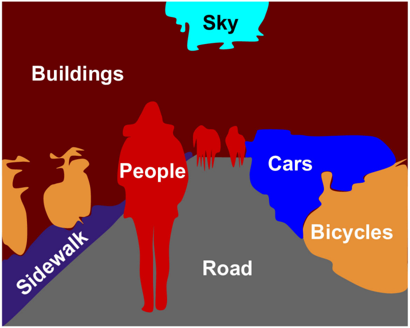
Random sampling
(with replacement = bootstrap)

Training set

Part II Neural Networks and Deep Learning



Chapter 14: Computer Vision



The end

Thank you and good luck for the exam!