WS 2019/2020 Syllabus of the MSc Course

INF52

Introduction to Data Science

Prof. Dr. Ingo Scholtes
Data Analytics Group
University of Wuppertal, DE

4L + 2E / 9 LP

Lectures: Tuesday 12¿15 – 13:45, HS6 & Wednesday 08:15 – 09:45, HS6
Exercises: Evaluation of group project reports

Executive Summary

Digital technologies provide us with an ever-increasing amount of data, e.g. in the context of online services, electronic commerce, financial services, digital humanities, computational social science, or life sciences. But how can we turn massive volumes of potentially noisy, time-stamped, and high-dimensional data into knowledge? How can we reason about relationships and patterns? Can we use these patterns to make predictions that can inform decision-making or help to design recommender systems? And how can we explore large volumes of noisy and high-dimensional data?

This course equips students both with theoretical and practical skills in data analytics, data science and statistical learning that can be used to address these questions. It combines a series of theory lectures (**Lxx**) on key concepts and algorithms with interactive practice lectures (**Pxx**), which demonstrate how they can be applied using state-of-the-art python packages. The course material consists of annotated slides for theory lectures and `jupyter` notebooks for the practice lectures. Students can further test and deepen their knowledge through three group projects that accompany the course.

Chapter I: Motivation and Introduction

The first chapter of our course motivates the need for data science and machine learning, and shows why those techniques are increasingly important in science, industry, and society. We further introduce the mathematical, statistical, and computational foundations of data science and show how we can develop professional data science projects in `python`.

L01 What is Data Science?

We give motivating examples for applications of data analytics and statistical learning in science, industry, and society and provide an overview of the course.

- Introduction and motivation
- History of data science and machine learning
- Exemplary applications of data analytics
- Outline of the course

L02 Probabilistic Foundations

We introduce statistical and philosophical foundations of statistical learning and introduce key terminology and notations. We introduce the frequentist and Bayesian view on inference and link them to questions in epistemology.

- Probabilistic foundations
- Statistical modelling
- Frequentist vs. Bayesian learning
- Philosophy of (data) science

P01 Introduction to `python`

We introduce key technologies needed to develop (collaborative) data science projects in `python` in a professional manner.

- Setting up a `python` data science environment
- A `python 3` crash course
- `jupyter` and Visual Studio Code
- Collaborative projects with `git`

P02 Data Analytics in `python`

Connecting the theoretical foundations from L02 to practice, we introduce basic techniques for data management and cleaning, data visualisation, scientific computing, and computational statistics in `python`.

- Matrix and vector arithmetics with `numpy/scipy`
- Computational statistics in `scipy.stats`
- Managing and cleaning data with `pandas`
- Visualising data with `seaborn` and `matplotlib`

L03 Learning from Data

We introduce statistical and philosophical foundations of statistical learning and introduce key terminology and notations. We introduce the frequentist and Bayesian view on inference and link them to questions in epistemology.

- Frequentist vs. Bayesian learning
- Philosophy of (data) science

P02 Data Analytics in python

Connecting the theoretical foundations from L02 to practice, we introduce basic techniques for data management and cleaning, data visualisation, scientific computing, and computational statistics in python.

- Managing and cleaning data with pandas
- Matrix and vector arithmetics with numpy/scipy
- Computational statistics in scipy.stats
- Visualising data with seaborn and matplotlib

Chapter II: Supervised Techniques for Inference and Prediction

In the second chapter of our course we will cover supervised statistical learning techniques that help us to (i) reason about (linear) relationships in data and (ii) to make predictions about data with categorical labels. Through a detailed discussion of one a particularly simple statistical learning technique, in the first three lectures we introduce key challenges in data science and machine learning. We further discuss statistical hypothesis testing and model selection techniques that are of central importance for real-world data science applications.

L04 Modelling Linear Relationships

We introduce a simple yet powerful supervised statistical learning technique that can be used to reason about linear relationships in noisy data. We specifically show how we can efficiently estimate the parameters and assess the accuracy of linear models.

- Supervised Learning and Inference
- Ordinary Linear Regression
- Model Accuracy and Confidence
- Multiple Regression and Categorical Data

L05 Testing Hypotheses in Data

We study how we can use statistical methods to formulate and test hypotheses in large data sets and how those techniques can be applied to argue about linear relationships. We further highlight fallacies in the interpretation of p-values and introduce methods to avoid them.

- Statistical Hypothesis Testing
- Regression Diagnostics
- Parametric and Non-Parametric Tests
- Misusing Data

L06 Selecting Optimal Models

We introduce the kernel trick, which can be used to apply linear regression to data that exhibits non-linear relationships. We further introduce overfitting, one of the key problems in data science and machine learning, and show how it can be mitigated by means of cross-validation and statistical model selection techniques.

- Non-linear models: Kernel trick
- Cross-Validation
- Statistical Model Selection
- Information-based model selection

P04 Robust Regression

We show how linear regression analyses can be performed in python. We first implement an ordinary least squares regression ourselves and generalise the model to multi-variate predictors. We then show how we can use sklearn to perform robust uni- and multi-variate linear regression.

- Fitting a Linear Model in python
- Evaluating the Accuracy of Linear Models
- Mutiple and Multivariate Linear Regression
- Outliers and Robust Linear Regression

P05 Validity of Linear Models

We apply the statistical hypothesis testing techniques from L04 to assess the validity of linear models using the packages scipy, sklearn, and statsmodels. We further exemplify common data science fallacies and show how to mitigate them.

- Testing normality of regression residuals
- Testing for linear relationships with statsmodels
- Comparing distributions
- Multiple hypotheses testing

P06 Overfitting and Model Selection

We show how the kernel trick can be applied in the data science package sklearn and we use a polynomial regression to illustrate the overfitting problem. We then implement the cross-validation and model selection techniques introduced in the theory lecture to determine the optimal maximum degree of the polynomial model.

- Non-linear models: Kernel trick
- Cross-Validation
- Statistical Model Selection
- Information-based model selection

L07 Statistical Classification

We introduce the statistical classification problem and motivate its importance based on real-world applications in recommender systems, medicine, financial systems, and autonomous driving. We show how the classification problem can be addressed using regression with a logistic function, as well as using Bayes formula.

- Classification vs. Regression
- Binary Logistic Regression
- Multinomial Logistic Regression
- Naive Bayes Classification

L08 Non-Linear Classification

Building on last week's introduction to the classification problem, we introduce advanced state-of-the-art classification techniques. We specifically explain support vector machines with linear and non-linear kernels, as well as classification techniques that build on decision trees.

- Support Vector Machines
- Non-linear kernel SVMs
- Decision Trees
- Random Forests

P07 Logistic Regression & Naive Bayes

We show how the binary classification problem can be solved by means of a logistic regression with `sklearn`. We generalise binary logistic regression to data with more than two classes and implement basic metrics to assess classifier performance. We finally show how to implement a Naive Bayes Classification with `sklearn`.

- Logistic Regression
- Gradient Ascent Optimisation
- Assessing classifier performance
- Bayesian Classification

P08 Advanced Classification in `sklearn`

We demonstrate the use of advanced classification techniques in `python` and `sklearn`. We illustrate the limitations of SVMs with linear decision boundaries and show how we can address them using the kernel trick and tree-based classification techniques.

- SVM-based classification
- SVM with non-linear decision boundaries
- Decision Tree Classification
- Training Random Forests

Chapter III: Unsupervised Learning and Explorative Data Analysis

In the third chapter we turn our attention to unsupervised learning techniques, that help us to do explorative analyses of data sets that lack a response variable. We introduce techniques to find cluster structures in such data and show how we can detect cluster or group structures in graph models of relational data. We further discuss key techniques to deal with high-dimensional data and to discover latent features in complex multi-variate and relational data.

L09 Finding Clusters in Data

We introduce the problem of finding groups or clusters of similar objects in noisy data and provide exemplary applications in practice. We show how we can treat the problem as a simple optimisation problem and highlight the issue of over- and underfitting the cluster number.

- The clustering problem
- K-means clustering
- Expectation maximisation clustering
- DBSCAN algorithm

L10 Graph Analytics

We motivate the representation of relational data in terms of graphs and complex networks and study the link prediction problem. We finally introduce statistical graph ensembles and show how they can be used to detect cluster structures in graphs.

- Mining Data with Graphs
- Random Graph Models
- Graph Model Inference
- Stochastic Block Model

L11 Graph Clustering

We revisit the problem of overfitting in the context of graph clustering. We introduce a regularised version of the stochastic block model that can be used to avoid overfitting and present the flow compression algorithm `InfoMap`.

- Stochastic Block Model Regularization
- Entropy of Graph Ensembles
- Random Walks and Flow Compression
- Clustering with InfoMap

P09 Validating Cluster Analyses

We implement the k-means clustering algorithm in `python` and apply it to a synthetic data set. We show cases where this simple method fails and we introduce advanced techniques that avoid this issue. We finally discuss the problem of over- and underfitting clusters.

- k-means clustering
- Expectation maximisation
- DBSCAN algorithm
- Over- and underfitting cluster structures

P10 Graph Analytics with `pathpy`

We show how we can perform basic graph analysis and visualisation tasks using the package `pathpy`. We implement statistical graph ensembles and perform a simple maximum likelihood estimation of a generative random graph model.

- Graph analysis and visualisation with `pathpy`
- Random graph models
- Model inference
- Implementing the stochastic block model

P11 Walk-based Graph Clustering

We practically explore under- and overfitting of group structures in social network analysis. We then show how we can solve this issue through a simple variant of the flow compression algorithm `InfoMap` for undirected networks.

- Random walks in graphs
- Spectral clustering in graphs
- Description length and flow compression
- Graph clustering with `InfoMap`

L12 Analysing high-dimensional data

We study methods to analyse multivariate data with high dimensionality, i.e. data where the number of features is large compared to the sample size. We introduce methods to reduce the dimensionality in a way that preserves relevant patterns in high-dimensional data.

- Multivariate Data Analysis
- Dimensionality Reduction
- Singular Value Decomposition (SVD)
- Principal Component Analysis

L13 Discovering Latent Features

We give an overview of unsupervised techniques to discover latent features in complex data sets. We motivate the problem in recommender systems and show how it can be addressed using algebraic and statistical techniques.

- Motivation: Recommender Systems
- Matrix Factorisation Techniques
- Matrix Factorisation in Graphs
- Latent Dirichlet Allocation

P12 Feature Extraction

We show how dimensionality reduction techniques like Principal Component Analysis can be used for feature selection and extraction. We then demonstrate how unsupervised dimensionality reduction can be combined with supervised classification methods.

- Linear PCA in `sklearn`
- Kernel PCA
- Feature Extraction
- PCA and Statistical Classification

P13 Topic Modelling

We practically demonstrate the discovery of latent features. We specifically show how matrix factorisation can be used to implement a simple recommender system and how latent dirichlet allocation can be apllied in topic modelling.

- Matrix factorisation
- Recommender Systems
- Latent Dirichlet Allocation
- Topic Modelling

Chapter IV: Outlook

L14 Data Science & Society

We motivate the use of data science in social computing and discuss resulting societal and ethical challenges. We study recent cases raising concerns about algorithmic accountability, fairness, and transparency and highlight the moral responsibility of data scientists.

- Data science and social computing
- Case study: Social scoring
- Case study: Human vs. machine predictions
- Case study: Predicting personality

L15 Data Science & Industry

In a final guest lecture, we take a practitioner's perspective on common issues that arise in the implementation of data science projects and strategies in industry.
This guest lecture will be given by René Pfitzner, founder of moonshot zurich GmbH, CEO of Experify.io and former lead data scientist at Neue Züricher Zeitung.

P14 Biased Data & Fairness

We explore ethical issues that can arise when we use biased data to train statistical classification algorithms. We further explore the ethical dimension of assessing classifier performance in terms of precision and recall and highlight issues arising in highly imbalanced data.

- Biased data and protected attributes
- Fairness measures for classifiers
- Dealing with imbalanced data
- Minimising type I vs. type II errors

P15 Exam Q&A

As a preparation for the exam, we give participants the opportunity to ask questions about past lectures as well as general issues in data science. Students are further invited to request a repetition of topics from any of the theory or hands-on practice lectures of this semester.

Group Projects

Through three group projects that accompany the course, participants can gain practical data analytics experience in real data. Each group consists of a minimum of three and a maximum of five students. Successful completion of a project involves the development of a `jupyter` notebook and a report that summarises the results. Solutions are evaluated and presented in a regular exercise session. Students can gain bonus points for the final exam. Provisionary topics of the group projects are given below:

Project I: Linear Regression and Hypothesis Testing

Data	Commit log data of six major Open Source Software projects
Tasks	Data cleaning and Data management Identify factors that affect the productivity of software developers
Issue date	22.10.2019 (Week 03)
Submission	12.11.2019 (Week 06)
Evaluation	19.11.2019 (Week 07)

Project II: Classification

Data	TBA
Tasks	TBA
Issue date	19.11.2019 (Week 07)
Deadline	10.12.2019 (Week 10)
Evaluation	17.12.2019 (Week 11)

Project III: Cluster Detection

Data	TBA
Tasks	TBA
Issue date	17.12.2019 (Week 11)
Deadline	21.01.2020 (Week 14)
Evaluation	28.01.2020 (Week 15)

Note: The information on contents and sequence of lectures in this syllabus is merely indicative and subject to change.