STAT3064: Statistical Learning for Data Science
STAT5061: Statistical Data Science
taught jointly in 2022

Content by week

Week 1: Introduction to Data Science

Visualisation, some multivariate statistics background, good simulations, reproducibility and performance evaluation

Week 2: Principal Component Analysis part I

Key concepts for the population and sample, examples, properties of principal components

Week 3: Principal Component Analysis part II

Raw, scaled and sphered data, high-dimensional low sample size data, PC regression

Week 4: Factor Analysis

Key ideas for population and sample FA, varimax criterion, sample FA, which FA? Gaussian FA and ML solutions, testing for the number of factors

Week 5: Canonical Correlation Analysis part I

Key concepts for the population and sample, examples, properties of canonical correlations

Week 6: Canonical Correlation Analysis part II

CCA for transformed data, Test statistic and testing for correlation, Canonical Correlation Regression

Week 7: Agglomerative Hierarchical Clustering

Examples, dendrograms, distances and linkages, cluster algorithm, properties of clusters: variability, patterns, how many clusters?

Week 8: k-Means Clustering

k-means algorithm, optimality, assessing cluster performance, visualising cluster arrangements

Week 9: A Case Study and More Clustering

The Dow Jones case study: finding pattern with different methods and their interpretation, PC clustering, clustering binary data

Week 10: Linear Discriminant Analysis

Fisher's key ideas for population and sample, between and within class concepts, LDA for Gaussian populations

Week 11: Cross-Validation and Logistic Regression

Naive Bayes and Bayes rule, performance assessment of a rule, misclassification, prediction, testing and training, CV, Leave-one-out, case study in logistic regression: heart failure data

Week 12: Logistic Regression part II

Logistic regression estimator, case study: heart failure data, decision threshold, comparison with LDA, QDA, regularisation in logistic regression, intro to trees and random forests with examples

Each 2-hour lecture is accompanied by a 2-hour lab. We will use 'R' and RStudio in the computing labs.

There will be 3 assignments, weekly quizzes, a practice test and a final exam on the computer.

Students will be provided with lecture slides, recordings of lectures, and chapters from I. Koch and A. Pope: Statistical Learning for Data Science.

Lecturer: Inge Koch
Professor of Statistics and Data Science
The University of Western Australia