857-0002-00L

Methods III: Statistical Learning

Course Syllabus

Lecturer:
Dominik Hangartner

Tutorial Instructor:
Moritz Marbach

Graduate Teaching Assistant:
Eroll Kuhn

Spring 2019

"It is exceedingly difficult to make predictions, particularly about the future."

Danish proverb

Contact Information:
Dominik Hangartner, Public Policy Group, Department of Humanities, Social and Political Sciences, ETH Zürich.
Email: dominik.hangartner@gess.ethz.ch.

Moritz Marbach, Public Policy Group, Department of Humanities, Social and Political Sciences, ETH Zürich.
Email: moritz.marbach@gess.ethz.ch.

Eroll Kuhn
Email: eroll.kuhn@gess.ethz.ch

Course Information:

Meeting times: There will be twelve two-hour lectures, Monday 14.15-16.00 in IFW E 42 and eleven tutorials, offered on Tuesday, 14:15-16:00 in IFW E 42.

Course Description

This course provides an introduction to the key statistical methods for supervised learning in the social sciences. Upon completion of the course, students will have an understanding of modern computational methods for statistical modelling and prediction, the assumptions on which they are based, and be able to use them to address specific research questions in the social sciences. Topics include linear regression with interaction and fixed effects, binary logistic regression, classification, resampling methods, shrinkage approaches, tree-based methods, support vector machines, and clustering algorithms.

Organization

The first nine weeks of the course focus on statistical methods for supervised learning including linear regression, classification, resampling methods, model selection and regularization, non-linear models (polynomial regression, splines, local regression, generalized additive models, etc.), tree-based methods and support vector machines. The last few weeks focus on methods for unsupervised learning including clustering algorithms.

Prerequisites

Prerequisites and assessment: Knowledge of basic descriptive and inferential statistic to the level of Methods II: Quantitative Methods, or equivalent. If you need to review material on covered in Methods II: Quantitative Methods, please consult this textbook:

- Imai, Kosuke. 2017. *Quantitative Social Science: An Introduction.* Princeton University Press.

If you need to review some `R` basics, you may want to have a look at

- Fox, John. 2002. *An R and S-PLUS Companion to Applied Regression.* Sage Publications.

Software

`R` will be used in tutorials.

Materials

The main course texts is:

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2017 (7th ed.). *An Introduction to Statistical Learning.* Springer [ISL]. The PDF of the book is freely and legally available at: `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf`

Assessment

The final grade is based on the following assessments: a poster project (25%), a lab report (25%), and 11 problem sets (totalling 50%). For problem sets, the usual grading scale is used.

Paper copies of problem sets and the take-home exams should be submitted at the start of class. They can be Microsoft Word documents or PDFs created with LaTeX. All figures and tables should be included in-text. A print out of the `R` code used in analyses should also be submitted.

Schedule

Week 1 - February 18
Introduction to Statistical Learning
Topics: Supervised and unsupervised learning; statistical learning and regression; curse of dimensionality and parametric models, assessing model accuracy and bias-variance trade-off; classification problems and K-nearest neighbors.

Readings:

- ISL Chapters 1 and 2.

Tutorial 1:

- Introduction/Refresher to `R`

Week 2 - February 25
Linear Regression
Topics: Simple and multiple linear regression; hypothesis testing and confidence intervals; interpreting regression coefficients, qualitative predictors, interactions and polynomials.

Readings:

- ISL Chapter 3

Tutorial 2:

- Predicting Boston crime rates in the 1970s

Week 3 - March 4
Classification
Topics: Maximum likelihood estimation; simple and multivariate logistic regression; case-control sampling; multiclass logistic regression; univariate and multivariate linear discriminant analysis; Bayes theorem; ROC curves.

Readings:

- ISL Chapter 4.

Tutorial 3:

- Analyze Canadian judges' decision on asylum case appeals using logistic regression.

Week 4 - March 11
Resampling Methods
Topics: Estimating prediction error and validation set approach; K-fold cross-validation; the bootstrap.

Readings:

- ISL Chapter 5

Tutorial 5:

- Apply resampling techniques to the Boston crime rates predictions.

Week 5 - March 18
No Lecture This Week

Tutorial 5:

- Voluntary: Review session

Week 6 - March 25
Linear Model Selection and Regularization
Topics: Linear model selection; best subset selection; forward and backward stepwise selection; estimating test error: Cp, AIC, BIC, adjusted R-squared; estimating test error using cross-validation; shrinkage methods and ridge regression; the lasso; tuning parameter selection; dimension reduction.

Readings:

- ISL Chapter 6.

Tutorial 6:

- Selecting variables that predict wages using the LASSO.

Week 7 - April 1
Non-linear Modeling
Topics: Polynomial regression and step functions; piecewise polynomials; splines; local regression; generalized additive models.

Readings:

- ISL Chapter 7.

Tutorial 7:

- Improving wage-prediction using non-linear models.

Week 8 - April 8
No Class (Lecture and Lab) This Week ("Sechselaeuten")

Week 9 - April 15
Tree-Based Methods
Topics: Decision trees; pruning; classification tree; comparison with linear models; bagging; random forests; boosting; variable importance.

Readings:

- ISL Chapter 8.

Tutorial 8:

- Predicting civil wars with trees and forests.

Week 10 - April 22 / 23
Spring Break: No Class This Week

Week 11 - April 29
Support Vector Machines
Topics: Maximal margin classifier; support vector classifier; kernels; support vector machines; comparison with logistic regression.

Readings:

- ISL Chapter 9.

Tutorial 9:

- Extracting data from satellite imagery.

Week 12 - May 6
Overview and Review: Supervised Learning
Topics: Review of all supervised learning methods covered.

Readings:

- No new readings.

Tutorial 10:

- Preparation for poster presentation.

Week 12 - May 13
Poster presentations on supervised learning

- Preparation for poster presentation (May 13) and poster presentation (May 14).

Week 12 - May 20
Unsupervised Learning
Topics: Introduction to unsupervised learning; principal component analysis; clustering.

Readings:

- ISL Chapter 10.

Tutorial 11:

- Clustering.

***Week 13* - May 27**
Buffer week.
Week 15 - June 10
No class.
Submit lab report by 2pm.