Preparation course Statistical Learning (and Prediction) 2021-2022

Instructors

Marjolein Fokkema m.fokkema@fsw.leidenuniv.nl

Tom F. Wilderjans t.f.wilderjans@fsw.leidenuniv.nl

Course format and schedule

This course is set up as a combination of self-study (to be done **before the start** of the 2-day course, see further for guidelines on preparations per session) and a set of five interactive sessions: Q&A combined with working on exercises (mostly in R) and discussing solutions. Each session also has a lecture, but we prefer interactivity, thus the theoretical lecture part will be kept short.

- Quality of the sessions mainly depends on the questions raised by the students. Send your questions about the material covered in each session to the respective instructor.
- It is important that students prepare themselves well for each session, following the preparation instructions presented further. Some theory will be discussed during the lectures but this will be limited (the theory behind the techniques is assumed more or less known).
- During the sessions, students mainly will work on exercises regarding the covered material. Therefore, **bring your laptop to each session** (with R and RStudio installed and the required R packages, see further).
- Tuesday session (February 1, 2022) will take place at FSW room 0A28.
- Wednesday session (February 2, 2022) will take place at FSW room 1A01.

Session	Date*	Time	Instructor	Chapters from ISLR
1	Tue Feb 1, 2022	11h-13h	Marjolein Fokkema	2**
2	Tue Feb 1, 2022	14h-16h45	Tom Wilderjans	3.5, 4.1-4.5, 5
3	Tue Feb 1, 2022	17h00-18h15	Tom Wilderjans	6.1, 6.2
4	Wed Feb 2, 2022	10h-12h	Tom Wilderjans	6.3, 6.4, 12
5	Wed Feb 2, 2022	13h-18h	Marjolein Fokkema	7, 8, 9

* *Location*: Leiden University, FSW: Pieter de la Court building, Wassenaarseweg 52, 2333 AK Leiden.

** also including chapters 1 & 3, but this material is considered known.

Course material

For the course we will use the book of James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R (Second Edition)*. New York, USA: Springer (denoted by ISLR). A pdf of this book can be downloaded for free from https://www.statlearning.com/.

Preparation for each session

As a preparation, for each session, you are required to watch a number of videos and/or to read the related chapters in the ISLR book (there is a large correspondence between the videos and the book):

- You can find the videos at https://www.dataschool.io/15-hours-of-expert-machine-learning-videos/.
- Send your questions regarding the material to be prepared for each session to the instructor in question.

An overview of the specific material to be prepared for each session can be found below.

Materials

All materials (slides, R-code, etc.) can be found online at

https://github.com/marjoleinF/winter-course-SLP.

Exam

There is no exam for this course. Active participation is required for obtaining the course credits.

Material to be prepared for each session

We assume Chapters 1 and 3 of ISLR to be known.

Session 1

- Read chapter 2 of ISLR.

- Download the R Package *ISLR*.

- Watch videos of Chapter 2: Statistical Learning

- **Statistical Learning and Regression** (11:41)
- **Curse of Dimensionality and Parametric Models** (11:40)
- **Assessing Model Accuracy and Bias-Variance Trade-off** (10:04)
- **Classification Problems and K-Nearest Neighbors** (15:37)

- Check if you are familiar with the material in the lab sessions of Chapter 2 and 3 (sections 2.3 and 3.6). Replicate the analyses you are unfamiliar with.

Session 2

- Read chapters 3.5, 4.1-4.5 and 5 of ISLR.

- Download the R Packages *class*, *e1071*, *boot* and *MASS*. Based on the lab videos (ISLR chapters 4.7 -but not 4.7.7- and 5.3), get to know the *glm*, *cv.glm*, *knn* and *boot* function.

- Watch videos of Chapter 2: Statistical Learning (also included in preparation of session 1)

- **Classification Problems and K-Nearest Neighbors** (15:37)

- Watch videos of Chapter 4: Classification

- **Introduction to Classification** (10:25)
- **Linear Discriminant Analysis and Bayes Theorem** (7:12)
- **Multivariate Linear Discriminant Analysis and ROC Curves** (17:42)
- **Lab: Logistic Regression** (10:14)
- **Lab: Linear Discriminant Analysis** (8:22)
- **Lab: K-Nearest Neighbors** (5:01)

- Watch videos of Chapter 5: Resampling methods

- **Estimating Prediction Error and Validation Set Approach** (14:01)
- **K-fold Cross-Validation** (13:33)
- **Cross-Validation: The Right and Wrong Ways** (10:07)
- **The Bootstrap** (11:29)
- **More on the Bootstrap** (14:35)
- **Lab: Cross-Validation** (11:21)
- **Lab: The Bootstrap** (7:40)

- Work on the following exercise. The file *student-mat.csv* contains data for 395 students on several variables. For this session, we will only use the variables *studytime* (amount of time studied), *schoolsup* (whether or not the school provides extra support), *romantic* (whether or not being involved in a romantic relationship) and *G1* (the score on a mathematics test at measurement moment 1). The goal is to predict the mathematics achievement (*G1*) based on three predictors mentioned before.

- Use the validation approach to determine the test error for the following two models:

    - A model with the main effects of the predictors (there are 3 predictors + intercept)
    - A model with the main effects of and the two-way interactions between the three predictors (6 terms + intercept)
- Determine the test error for the same two models but now by using 10-fold cross-validation (CV). Which model is the best model?
- Determine the standard error for the terms included in the second model by means of a bootstrap analysis. Compare the bootstrap estimate for the standard error with the standard error provided by the regression analysis.

- Work also on this exercise. Take the data from *student-mat.csv*, but now make the dependent variable *G1* binary (a new variable *G1bin*) by coding a score of 8 or lower as 0 ("failed") and a score above 8 as 1 ("passed"). Perform K-Nearest Neighbors (KNN) to predict *G1bin* based on the three predictors from the previous question.

- Determine the optimal value of *K* (try values for *K* going from 1 up to 50) for the KNN classifier (by means of 10-fold CV). Determine the test error (by means of the validation approach) of the optimal KNN model. Use a model with only main effects for the KNN classifier.
- For the previous question it is important that you think about how to set up the training, validation and test set(s). For more information on this topic, consult (the end of) section 7.2 of Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.)*. New York, USA: Springer (download a pdf at http://statweb.stanford.edu/~tibs/ElemStatLearn/).

Session 3

- Read chapters 6.1 and 6.2 of ISLR

- Download R packages *glmnet* and *leaps*. Based on the lab videos (ISLR Chapters 6.5.1 and 6.5.2), make yourself acquainted with the *glmnet* and *cv.glmnet* function in the *glmnet* package.

- Watch videos of Chapter 6: Linear Model Selection and Regularization

- **Linear Model Selection and Best Subset Selection** (13:44)
- **Forward Stepwise Selection** (12:26)
- **Backward Stepwise Selection** (5:26)
- **Estimating Test Error Using Mallow's Cp, AIC, BIC, Adjusted Rsquared** (14:06)
- **Estimating Test Error Using Cross-Validation** (8:43)
- **Shrinkage Methods and Ridge Regression** (12:37)
- **The Lasso** (15:21)
- **Tuning Parameter Selection for Ridge Regression and Lasso** (5:27)
- **Lab: Best Subset Selection** (10:36)
- **Lab: Forward Stepwise Selection and Model Selection Using Validation Set** (10:32)
- **Lab: Model Selection Using Cross-Validation** (5:32)
- **Lab: Ridge Regression and Lasso** (16:34)

- Work on the following exercise. Take the data from *student-mat.csv*, but now use *G3bin* (a binarized version – make this variable yourself using 8 again as cut off point – of *G3*, the mathematics achievement at time point 3) as dependent variable. Use all other variables as predictors (also *G1*, *G2*, *G3* but NOT *G1bin*)

- Fit a ridge logistic regression and a lasso penalised logistic regression (predict *G3bin* and use all other variables as predictors, also *G3*). Determine the optimal λ for both regularization methods by means of 10-fold cross-validation. Determine (also with 10-fold CV) the test error for both methods. Which method performs best for this data set? Carefully think about what is the training, validation and test set(s).

Session 4

- Read chapters 6.3, 6.4 and 12 of ISLR

- Download R packages *pls* (ISLR Chapter 6.5.3).

Part A. Principal Component Regression (PCR) and Partial Least Squares (PLS)

- Watch videos of Chapter 6: Linear Model Selection and Regularization

- **Dimension Reduction** (4:45)
- **Principal Components Regression and Partial Least Squares** (15:48)

- Work on the following exercise. Use the *College* dataset from the *ISLR* package (use the variable *Apps* as response variable)

- Determine for this data set the test error obtained with PCR and PLS. Use 10-fold cross-validation to determine the optimal number of components for both methods.
- Which method performs better (in terms of predicting independent data)?
- Do the first two PLS components differ much from the first two PCR/PCA components?
- Compare both methods with respect to the amount of variance in the predictors and the response that is explained by the components.

Part B. Unsupervised methods

- Watch videos of Chapter 10: Unsupervised Learning (note that this is Chapter 12 in the second edition of the book)

- **Unsupervised Learning and Principal Components Analysis** (12:37)
- **Exploring Principal Components Analysis and Proportion of Variance Explained** (17:39)
- **K-means Clustering** (17:17)
- **Hierarchical Clustering** (14:45)
- **Lab: Principal Components Analysis** (6:28)
- **Lab: K-means Clustering** (6:31)
- **Lab: Hierarchical Clustering** (6:33)

- Work on the following exercise. Use the *College* dataset from the *ISLR* package (use only the variables *Top10perc*, *Personal* and *PhD*, which are the 5th, 12th and 13th variable)

- Perform a K-means cluster analysis by using different values for the number of clusters *K*.
    - Which value of *K* is optimal?
    - Interpret the obtained clustering
- Perform a hierarchical clustering (take only 20 cases, sampled at random)
    - Decide on the dissimilarity measure and linkage method to use.
    - How many cluster do you retain?
    - Compare the clustering obtained with hierarchical clustering with the clustering obtained from a K-means analysis (on the same reduced data set consisting of 20 cases only).

Session 5

- Download the R Packages *tree*, *randomForest*, *partykit*, *splines*, *gam* and *MASS*.

- Read chapter 7 (Moving Beyond Linearity) and watch the following videos:

- **Polynomial Regression and Step Functions** (14:59)
- **Piecewise Polynomials and Splines** (13:13)
- **Smoothing Splines** (10:10)
- And replicate the analyses from the lab session (Section 7.8)

- Read chapter 9 (Support Vector Machines) and watch the following videos:

- **Maximal Margin Classifier** (11:35)
- **Support Vector Classifier** (8:04)
- **Kernels and Support Vector Machines** (15:04)
- And replicate the analyses from the lab session (Section 9.6; you can skip 9.6.3, 9.6.4 and 9.6.5)

- Read chapter 8 (Tree-Based Methods) and watch the following videos:

- **Decision Trees** (14:37)
- **Pruning a Decision Tree** (11:45)
- **Classification Trees and Comparison with Linear Models** (11:00)
- **Bootstrap Aggregation (Bagging) and Random Forests** (13:45)
- **Boosting and Variable Importance** (12:03)
- And replicate the analyses from the lab session (Section 8.3)