COURSE MANUAL
SUPERVISED MACHINE LEARNING

Research Master Business Data Science
Erasmus University Rotterdam, University of Amsterdam, Vrije Universiteit Amsterdam

Course Code	EBDS19102
Academic year	2019-2020
Period	1.2
Credits	4
Recommended knowledge	Business Foundations, Programming Basics (specifically programming in R, use of R Markdown or knitr), Mathematics, Statistics
Required knowledge	Linear Algebra

Table of Contents

business data science
GRADUATE PROGRAM | Erasmus | VU

1. COURSE COORDINATOR AND LECTURERS

Coordinator/Lecturer: prof. dr. Patrick Groenen (EUR)
Email: **groenen@ese.eur.nl**
Short bio: Patrick Groenen is a professor of Statistics at the Erasmus School of Economics. His work focuses on development of data science methods and their numerical algorithms. He is the co-author of a textbook on multidimensional scaling published by Springer and has published articles in the top peer-reviewed journals including, among others, the Journal of Machine Learning Research, the Journal of Marketing Research, Computational Statistics and Data Analysis, Psychological Methods, Psychometrika, the Journal of Classification, the British Journal of Mathematical and Statistical Psychology, and the Journal of Empirical Finance.

Lecturer: dr. Pieter Schoonees (EUR)
Email: schoonees@rsm.nl
Short bio: Pieter Schoonees is an assistant professor in the Department of Marketing Management at RSM, Erasmus University. His expertise lie in the fields of computational statistics, machine learning and psychometrics. Pieter's research focuses on developing statistical and machine learning algorithms and applying these to secondary data. A special interest is the use of such techniques for the analysis of data gathered from neuroscientific studies.

Teaching Assistant: TBA
Email: TBA

2. COURSE CONTENT

Statistical learning methods arising from statistics, machine learning, and data science have become more widely available. These methods can be split into supervised learning, with the aim of predicting a response variable, and unsupervised learning, which aims to describe the relations between all variables simultaneously. This course focusses on supervised learning and has as its goal that the student obtains a thorough technical understanding of a selection of supervised machine learning techniques, can implement the technique in the high level language R and can write a report about an application of the technique.

The book of Hastie, Tibshirani, and Friedman (2001, 1st edition) has been a milestone in connecting statistical ideas into machine learning techniques. Parts of the second edition of this book (2009) form the basis of this course. An overview of techniques and ideas to be treated are:

- linear methods for regression,
- linear methods for classification,
- basis expansion and regularization,
- model assessment and selection,
- classification and regression trees,
- ensemble learning (random forests, bagging, and boosting),
- support vector machines.

3. LEARNING OBJECTIVES

By the end of the course students will be able to:

KNOWLEDGE AND UNDERSTANDING	Understand the fundamental building blocks of several supervised machine learning methods, with specific attention to: o linear methods for regression and classification, o basis expansion and regularization, o model assessment and selection, o classification and regression trees, o ensemble learning (random forests, bagging, and boosting), o support vector machines.
APPLICATION OF KNOWLEDGE	Program these methods by translating technical knowledge of a method into their own code in R.
MAKING JUDGEMENT	Apply and interpret these methods.
COMMUNICATION	Write a small report in the form of a short scientific article.

4. STUDY MATERIAL

The following book is considered essential for your learning experience and it is part of the examined material. Changes in the reading list will be communicated via Canvas.

Book:
Hastie, T., Tibshirani, R. and J. Friedman. (2009). "The elements of statistical learning (2nd edition)." *Springer*. Available at https://web.stanford.edu/~hastie/Papers/ESLII.pdf.

5. FORM OF TUITION

The lectures aim at stimulating your academic skills, and providing you with new knowledge. In this course, lectures are accompanied by tutorials. The information provided in the lecture is essential for the exercises, assignments and discussions during those sessions. We expect students to come to the lectures well prepared and to participate in the interaction. Each week, one group will give a presentation on an exercise or assignment submitted before class.

The tutorials aim at practicing the theory using exercises and allowing students to ask for additional explanation for those parts of the material perceived as more difficult.

Class participation is a part of the course that is important for learning. Students are expected to contribute to asking questions, answers and to the general learning atmosphere.

6. ASSESSMENT

Your overall course grade is composed based on different components that are presented in the assessment overview. You need a minimum overall grade of 5.50 to pass the course. The final grade is rounded to the nearest multiple of .0 or .5, with the following exceptions: any grade between 5.0 and 5.5 is rounded to a 5; a 5.5 is rounded to a 6; a 0.5 does not exist (in other words, besides 5.5, only integer grades are awarded). Grades for homework or midterm examinations do not need to be rounded.

To prepare for the individual assignment at the end of the course, students practise in groups of three with similar group exercises and one group assignment. These assignments are aimed at being able to program a method in their own code and being able to write a report in article form. Feedback to exercises is given. The group assignment is graded. In each week, one of the groups also give a short presentation on the exercise or group assignment which is followed by a group discussion.

Format	% grade	Knowledge and Understanding	Application of knowledge	Making Judgements	Communication
Exam (Individual assignment)	85%	X	X	X	X
Group Assignment	15%	X	X	X	X

7. DETAILED COURSE SCHEDULE

Please check Canvas for an up-to-date schedule, reading material and assignments.

Week	Date	Time	Format	Theme/Topics	Preparation
1			Lecture 1 (2 times 2hr)	**Linear methods for regression, model selection, and assessment**	Mandatory readings: • Hastie et al. (2009), Chapter 3.1, 3.2, 3.3

Week	Date	Time	Format	Theme/Topics	Preparation
			Tutorial 1 (1hr)	**Setting the Scene**	Managing expectations ● Team formation ● Explain Exercise Week 1 ● Explain what we expect
			Submission Deadline	**Exercise Week 1**	Submit your exercise before Lecture 2 via Canvas (see Section 9.1)
2			Lecture 2 (2 times 2hr)	**Regularized regression and k-fold cross validation**	*Mandatory readings:* ● Hastie et al. (2009), Chapter 3.4.1-3.4.3, 3.8.4, 7.10 Presentation on Week 1 Exercise by one group followed by discussion during second lecture
			Tutorial 2 (1hr)		Assistance for Exercise Week 2
			Submission Deadline	**Exercise Week 2**	Submit your exercise before Lecture 3 via Canvas (see Section 9.2)
3			Lecture 3 (2 times 2hr)	**Basis function expansions, kernels, bias-variance trade-off**	*Mandatory readings:* ● Hastie et al. (2009), Chapter 5.1-5.2.1, 5.8, 7.3, Presentation on Week 2 Exercise by one group followed by discussion during second lecture
			Tutorial 3 (1hr)		Assistance for Exercise Week 3
			Submission Deadline	**Exercise Week 3**	Submit your exercise before Lecture 4 via Canvas (see Section 9.3)
4			Lecture 4 (2 times 2hr)	**Support vector machines**	*Mandatory readings:* ● Groenen, Nalbantov, Bioch (2009) *Background readings:* ● Hastie et al. (2009), Chapter 12.1-12.3
			Tutorial 4 (1 hr)		Assistance for the Group Assignment
			Submission Deadline	**Group Assignment**	Submit your group assignment before Lecture 5 (see Section 9.4)
5			Lecture 5 (2 times 2hr)	**Classification and regression trees, random forests, bootstrap**	*Mandatory readings:* ● Hastie et al. (2009), Chapter 7.11, 9.2, 15
			Tutorial 5		Assistance for Exercise
			Submission Deadline	Exercise Week 5	Submit your exercise before Lecture 6 via Canvas (see Section 9.5)
6			Lecture 6 (2 times 2hr)	**Boosting**	*Mandatory readings:* ● Hastie et al. (2009), Chapter 10 Handing out individual assignment (see section 9.6)
			Tutorial 6		Assistance for individual assignment
8			Exam		

Week	Date	Time	Format	Theme/Topics	Preparation
	TBA		Submission Deadline		Submit your individual assignment within two weeks after it has been handed via Canvas (see Section 9.6)
	TBA		Exam inspection	**Feedback**	Check Canvas for updates

8. ASSIGNMENT INFORMATION

9.1 Exercise Week 1

Write the methods section for an article on multiple regression (approximately 1 to 1.5 pages). Give an intuitive explanation in words what multiple regression can be used for, provide technical details and diagnostics. Provide your own code for computing a regression solution and compare its results with a standard package (for example, in R, compare your code to the lm() function).

9.2 Exercise Week 2

Write an article (maximum 4 pages) with introduction, data, methods, result, and conclusion sections. The central technique is the elastic net. Provide your own code that computes an elastic net solution using the MM-algorithm as provided in the slides. Compare the results of your code with that of the glmnet package in R.

9.3 Exercise Week 3

Write the methods section for an article on the use of two methods for basis expansions in multiple regression (approximately 1 to 1.5 pages). Discuss how overfitting can become important when using basis expansions and why ridge regression could be used to solve it. Provide your own code for K-fold cross-validation and apply it sensibly to ridge regression.

9.4 Group Assignment (Week 4)

Write an article (maximum 4 pages) with introduction, data, methods, result, and conclusion sections. The central technique is the binary support vector machine (SVM). Provide your own code that computes a solution for the using the majorization algorithm from the article of Groenen, Nalbantov, and Bioch (2009). Compare the results of your code with that of the SVMMaj package in R.

9.5 Exercise Week 5

Write the methods section for an article on regression and classification trees and random forests (approximately 1.5 to 2 pages). It should contain a discussion how random forests implement the idea of regularization. Write your own code for an approximation of random forests by repeatedly fitting a tree (using a standard package) on randomly chosen predictors.

9.6 Written exam: individual assignment

The final exam is in the form of an individual take home assignment. It is handed out at the end of the final lecture in Week 6 of the course. The deadline is two weeks after handing out the individual assignment. The form of the individual assignment is the same as practiced in the group assignment and exercises: you will be assigned one or more methods and you will have to write a 4-page report in the form of an article (consisting of introduction, data description, methods, results, and conclusion). In addition, you will have to program your own code for one of the methods in the course.

APPENDIX A – INDIVIDUAL/GROUP ASSIGNMENT

Criterion	5 or lower	6	7	8	9 or 10
1. Research question **5%**	Question is unclear or illogical. Question is not functional. Question is too simple or too limited for the program or the study load.	Question refers to the technique, not to a substantive concept. Adequate and functional research question, but set at a minimum level of ambition.	Adequate and functional research question set at a level of ambition broadly appropriate for program and study load.	Well-formulated and clearly functional research question that can be answered by the available data and methods.	Original research question, displaying unusual insight and skill to translate relevant issues into well-formulated and researchable questions.
2. Method **40%**	Many of the elements under 7 are incorrect or missing.	As with 7 but some elements are incorrect or missing.	Methods section provides an intuitive explanation of the technique; the main goals and features are briefly described in simple terms. Notation is consistent and correct. The technical descriptions need to be correct. Key concepts of the method are correctly explained. Diagnostics of the method is properly described.	As with 7, explanation of techniques shows insight in the method.	As with 8, but addresses methodological and technical issues that go beyond what is covered in this program. Very extensive efforts in data / study collection.
3. Description and analysis of results **25%**	Many violations of the requirements outlined under 7. Poorly organized. Contains important errors of interpretation or logic; reveals lack of understanding of own research approach.	As with 7 but with several omissions. For example, standardized and/or mechanical presentation of results. Broadly effective, but inefficient or somewhat clumsy presentation. Contains minor errors of interpretation. Considerable unused potential for further analysis.	The order of discussion the results must be correct from more general modelling decisions towards interpretation. The results should be interpreted in terms of the meaning of the variables. The interpretation of the results should be sufficiently deep. Tables and figures support the decisions and interpretation, are readable, and have an informative caption.	As with 7 but is well-organized and thoughtful presentation of results, showing a good understanding of the nature of the data and many of the issues in interpretation.	As with 8, but very thorough analysis, showing a deep understanding of the research question, the research design, and the data. Presentation is highly effective.
4. Conclusion and discussion **5%**	No clear answer to research question, or an answer that does not follow from the research findings. No or just trivial suggestions for further research (e.g. 'collect more data').	Research question is answered by simple summary of findings. Perfunctory discussion of limitations and suggestions for further research.	Functional summary of findings, leading to discussion of extent to which research question is or is not answered. Meaningful reflection on limitations of own research. Some suggestions for useful further research.	A well-considered review of the findings in the light of the research question. Shows a clear understanding of limitations of own research. Several suggestions for further research that are properly explained.	Succeeds in putting the findings and the research question in the widest possible context, drawing out significant implications for theory development, research methodology and practice.

5. Written communication **10%**	Unstructured text. Fails to convey the key message of the report and/or to address questions. The text does not meet the academic editorial standards.	The text is somewhat unstructured and unclear. The text barely passes the academic editorial standards, as more polishing work is needed.	Overall well written, with occasional typos, or inaccuracies. The text passes the academic editorial standards, although the writing style is mechanical.	Structured text. The text is clear and concise, but here and there more (or less) details could improve the readability. Tables and Figures are self-explicatory and timely introduced in the text. The text meets the academic editorial standards, although the writing style is a bit mechanical at times.	Structured, coherent and polished text. Excellent writing style. The text is accurate, clear and concise, with the right level of detail. Tables and Figures are self-explicatory and timely introduced in the text. The text meets the academic editorial standards.
6. Code **15%**	Code is not working, difficult to read, or does not reproduce the results. Many of the requirements of 7 are missing.	As with 7, but some of the elements are missing.	Code implements the method correctly, is written for general data, structured, allows to reproduce the results. Naming of variables and comments are helpful to understand the code.	As with 7, but code is efficient and to the point.	Excellent code. Code goes contains an element of surprise that enhances its usage.
TOTAL					

APPENDIX B – EXAMPLE OF TAKE-HOME EXAM (INDIVIDUAL ASSIGNMENT)

Write a report of at most 4 pages (12pt font, no double columns) in which you solve a substantive research question using the techniques assigned to you. Use no more than 5 pages for introduction, data, methods, results, and conclusions, including tables and figures.

1. The report should be in the form of a small article (introduction, description of the data, substantive research question, methods, results, discussion and conclusions).
2. The description (including the technical properties, diagnostics, and usage) of the particular method is important.
3. Write a function for the technique assigned to you in R.
4. Use additional pages for code and the comparison of your program with the output of a standard function. Present your code as well as the script that runs the analysis. Make sure that your code can handle any data matrix of unknown size.
5. Search for your own data set from the sources indicated on the slides of week 6 or from another place. Describe the data briefly. Do not use a data set that have been used in this course before on the same technique by any of the groups or during the lecture.
6. Give a reference of the source of the data. If these data are your own, just say so.
7. Indicate whether the data propriety or can be freely used by others.
8. Do NOT copy and paste parts (e.g. method section) from your earlier assignments or exercises. Write them in your own words.
9. Substantive conclusions are important.
10. Show that you understand how to use the techniques sensibly.
11. Justify your conclusions by reporting appropriate results (possibly in tables or figures).

Deadline: two weeks after handing out the assignment. Check Canvas for definitive information.
Make a title page with: (in a large font) the number assigned to you, your name, your student number and your group number. (This page is not counted for the number of pages).