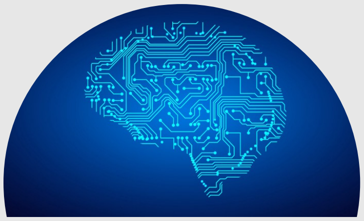

Machine Learning I:

Unsupervised Learning

Instructor Info

Prof. Dr. Marco Steenbergen

Office Hrs: Mon & Wed 13-14

AFL H-335

http://steenbergen.ch

steenbergen@ipz.uzh.ch

Course Info

Prereq: Knowledge of R, probability theory, and statistical inference

Tue

10.15-12.00

Lecture Room

Overview

Broadly speaking, unsupervised machine learning concerns itself with detecting communalities between observations. Although the term is of a relatively recent origin, some of the techniques go back over a century and have been used extensively in the social sciences. In this course, we consider three areas of unsupervised machine learning and one that is on the fence between supervised and unsupervised learning. The unsupervised methods include: (1) component analysis; (2) cluster analysis; and (3) mixture models. The hybrid area pertains to recommender systems and their uses. The goal is to provide you with a thorough grounding in the statistical theory behind those methods and to show how they can be applied to social science research questions.

Learning Objectives

Upon completing the course, you should be able to

- understand the purposes and applications of unsupervised machine learning.
- grasp key methods of unsupervised learning, their uses and limitations, and their meaning.
- program methods of unsupervised learning in R, interpret the results, and communicate them to a non-technical audience.

Material

Required Texts

We use a variety of texts that are indicated in the course schedule. Those texts will be made available on OLAT.

Other

You should install R on your own computer, as well as RStudio. The following packages will be required throughout the course:

matrixcalc, etc.

Requirements

The course requirements consist of 3 homework assignments and a term paper. The homework assignments require that you apply R to instructor-defined problems and interpret the results. The deadlines for the homeworks are indicated in the course schedule.

The term paper gives you an opportunity to apply the methods to a problem of your own choice. The paper should consist of the following parts: (1) research question; (2) detailed description of the data you will be using; (3) justification and description of a particular methodology; (4) results; and (5) discussion and interpretation. The paper is due on January 4, 2020. It should be no longer than 2500 words (exclusive of tables, figures, and references). It should be submitted as a single pdf with R code in the appendix.

Grading Scheme

20%	Homework 1
20%	Homework 2
20%	Homework 3
40%	Term Paper

Grades for each assignment are on the UZH half note scale. Curving is at the discretion of the professor.

FAQs

Machine Learning—What Is It?

The rubric of machine learning includes computer algorithms that learn from data in an autonomous manner.

What is Unsupervised Learning?

In supervised learning we know the values of the outcome. In unsupervised learning, we do not. The focus then is not on prediction as much as on finding patterns in data.

How Much Math is Involved?

In this course, we'll rely heavily on linear algebra. No worries, however: everything you need to know will be taught in the course.

How Much R Do I Need to Know?

For the applications, we rely heavily on R. That means that you need to feel comfortable using this computational platform. If this is not the case, please take advantage of one of the many MOOCs and online courses that are available prior to starting the class.

Course Policies

- Plagiarism will not be tolerated under any circumstance.
- In this connection, note that collaboration with fellow students is only allowed on the weekly homework assignments.
- Extensions require that you put in a formal request with the exam coordinator, Naomi Czisch. Neither the instructor nor the TAs are allowed to grant an extension on their own.

Class Schedule

MODULE 1: Mathematical Foundations

Sep 17	Course Overview and Introduction to Matrices	Steenbergen, Marco R. 2019. *Matrices and Their Statistical Applications.* Zurich: IPZ. Chapters 2-4.
Sep 24	Vector Geometry	Steenbergen, Marco R. 2019. *Matrices and Their Statistical Applications.* Zurich: IPZ. Chapters 5-6.

MODULE 2: Dimensionality Reduction

Oct 1 & 8	Component Analysis	Dillon, William R. and Matthew Goldstein. 1984. *Multivariate Analysis: Methods and Applications.* New York: Wiley. Chapter 2. Lagona, Francesco and Fabio Padovano. 2007. A Nonlinear Principal Component Analysis of the Relationship between Budget Rules and Fiscal Performance in the European Union. *Public Choice* 130(3-4): 401-436. Steenbergen, Marco R. 2019. *Matrices and Their Statistical Applications.* Zurich: IPZ. Chapter 7. Tharwat, Alaa. 2018. Independent Component Analysis: An Introduction. *Applied Computing and Informatics*, forthcoming.
Oct 15	Factor Analysis and Q Sorts	Carlin, Ryan. 2018. Sorting Out Support for Democracy: A Q-Method Study. *Political Psychology* 39(2): 399-422. Dillon, William R. and Matthew Goldstein. 1984. *Multivariate Analysis: Methods and Applications.* New York: Wiley. Chapter 2. Vatter, Adrian. 2009. Lijphart Expanded: Three Dimensions of Democracy in Advanced OECD Countries. *European Political Science Review* 1(1): 125-154. → Homework 1 is due.
Oct 22	Correspondence Analysis FA for tables...	Hoffman, Donna L. and George R. Franke. 1986. Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Researcjh* 23(3): 213-227. Steenbergen, Marco R. 2019. *Matrices and Their Statistical Applications.* Zurich: IPZ. Chapter 8. Teney, Celine and Laurie Hanquinet. 2012. High Political Participation, High Social Capital? A Relational Analysis of Youth Social Capital and Political Participation. *Social Science Research* 41(5): 1213-1226.
Oct 29	Introduction to Text Analysis	Aggarwal, Charu C. 2018. *Machine Learning for Text.* New York: Springer. Chapter 1.2-1.3.

		Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gardarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58(4): 1064-1082.
Nov 5	Multidimensional Scaling	Dillon, William R. and Matthew Goldstein. 1984. *Multivariate Analysis: Methods and Applications.* New York: Wiley. Chapter 4.
		Jacoby, William G. and David Armstrong. 2014. Bootstrap Confidence Regions for Multidimensional Scaling Solutions. *American Journal of Political Science* 58(1): 264-278.

MODULE 3: Cluster Analysis

Nov 12	Hierarchical Cluster Analysis	Dillon, William R. and Matthew Goldstein. 1984. *Multivariate Analysis: Methods and Applications.* New York: Wiley. Chapter 5.
		Wolfson, Murray, Zagros Madjd-Sadjadi, and Patrick James. 2004. Identifying National Types: A Cluster Analysis of Politics, Economics, and Conflict. *Journal of Peace Research* 41(5): 607-623.
		→ Homework 2 is due.
Nov 19	K-Means Clustering and Other Algorithms	Gurusamy, Vairaprakash, S. Kannan, and J. Regan Prabhu. 2017. Mining the Attitude of Social Network Users Using k-Means Clustering. *International Journal of Advanced Research in Computer Science and Software Engineering* 7(5): 226-230.
		Kasambara, Alboukadel. 2017. *Practical Guide to Clustering in R: Unsupervised Machine Learning.* Marseille: STHDA. Chapters 4-6.

MODULE 4: Mixture Models

Nov 26	Latent Class Analysis	Alvarez, R. Michael, Ines Levin, and Lucas Nuñez. 2017. The Four Faces of Political Participation in Argentina: Using Latent Class Analysis to Study Political Behavior. *Journal of Politics* 79(4): 1386-1402.
		Porcu, Mariano and Francesca Giambona. 2016. Introduction to Latent Class Analysis with Applications. *Journal of Early Adolescence* 37(1): 129-158.
Dec 3	Latent Profile Analysis	Ahlberg, Mikael G., Petra Svedberg, Maria Nyholm, Anthony Morgan, and Jens M. Nygren. 2019. Into the Realm of Social Capital of Social Capital for Adolescents: A Latent Profile Analysis. *PLOS One* 14(2): e0212564.
		Stanley, Laura, Franz W. Kellermanns, and Thomas M. Zellweger. 2017. Latent Profile Analysis: Understanding Family Firm Profiles. *Family Business Review* 30(1): 84-102.

MODULE 5: Association Rules

Dec 10 & 17	Rule-Based Systems	Bagui, Sikha, Dustin Mink, and Patrick Cash. 2007. Data Mining Techniques to Study Voting Patterns in the U.S. *Data Science Journal* 6: 46-63. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2014. *Introduction to Data Mining.* Harlow: Pearson. Chapter 6. → Homework 3 is due on Dec 10.