Machine Learning II
SURV 753
2 credits/4 ECTS
Summer Term 2020

Instructors
Trent D. Buskirk, PhD, buskirk@bgsu.edu
Christoph Kern, PhD, c.kern@uni-mannheim.de

Video lectures by Christoph Kern, PhD and Trent D. Buskirk, PhD

Short Course Description
Social scientists and survey researchers are confronted with an increasing number of new data sources such as apps and sensors that often result in (para)data structures that are difficult to handle with traditional modeling methods. At the same time, advances in the field of machine learning (ML) have created an array of flexible methods and tools that can be used to tackle a variety of modeling problems. Against this background, this course discusses advanced ML concepts such as cross validation, class imbalance, Boosting and Stacking as well as key approaches for facilitating model tuning and performing feature selection. In this course we also introduce additional machine learning methods including Support Vector Machines, Extra-Trees and LASSO among others. The course aims to illustrate these concepts, methods and approaches from a social science perspective. Furthermore, the course covers techniques for extracting patterns from unstructured data as well as interpreting and presenting results from machine learning algorithms. Code examples will be provided using the statistical programming language R.

Course and Learning Objectives
By the end of the course, students will...
- will have a profound understanding of advanced (ensemble) prediction methods
- have built up a comprehensive ML toolkit to tackle various learning problems
- know how to (critically) evaluate and interpret results from "black-box" models

Prerequisites
Topics covered in SURV751: Introduction to Machine Learning and Big Data (ML I), i.e.:
- Conceptual basics of machine learning (training vs. test data, model evaluation basics)
• Decision trees with CART
• Random forests

Familiarity with the statistical programming language R is strongly recommended.

Participants are encouraged to work through one or more R tutorials prior to the first-class meeting. Some resources can be found here:

• https://rstudio.cloud/learn/primers
• http://www.statmethods.net/
• https://swirlstats.com/
• https://www.rcommander.com

Class Structure and Course Concept
This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video-recorded lectures and reading the required literature for each unit prior to participating in mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching prerecorded videos, attending the live online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week.

This is a 2-credit/4-ECTS course that runs for 8 weeks. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings:
Fridays, 12:00 PM EDT/6:00 PM CEST, starting on June 5

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://www.elms.umd.edu/. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In
addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Thursday, 12:00 PM EDT/6:00 PM CEST).

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can of course use other online meeting platforms such as Google Hangout or Skype.

**Grading**

Grading will be based on

- 4 homework assignments (10% each)
- 8 online quizzes (5% each)
- Participation in discussion during the weekly online meetings (20% of grade)

\[
\begin{align*}
A+ & : 100 - 97 \\
A & : 96 - 93 \\
A- & : 92 - 90 \\
B+ & : 89 - 87 \\
B & : 86 - 83 \\
B- & : 82 - 80 \\
Etc. & : 
\end{align*}
\]

The grading scale is a base scale recommended by the IPSDS. Variations for grading on a scale are at the discretion of the instructor.

Dates of when assignment will be due are indicated on Canvas. Late assignments will not be accepted without prior arrangement with the instructors.

**Technical Equipment Needs**

The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore, we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.
Long Course Description
Social scientists and survey researchers are confronted with an increasing number of new data sources such as apps and sensors that often result in (para)data structures that are difficult to handle with traditional modeling methods. At the same time, advances in the field of machine learning (ML) have created an array of flexible methods and tools that can be used to tackle a variety of modeling problems. Against this background, this course discusses advanced ML concepts such as cross validation, class imbalance, Boosting and Stacking as well as key approaches for facilitating model tuning and performing feature selection. In this course we also introduce additional machine learning methods including Support Vector Machines, Extra-Trees and LASSO among others. The course aims to illustrate these concepts, methods and approaches from a social science perspective. Furthermore, the course covers techniques for extracting patterns from unstructured data as well as interpreting and presenting results from machine learning algorithms. Code examples will be provided using the statistical programming language R.

The course is structured such that each session focuses on specific prediction tasks and presents tools that can be used to tackle modeling problems in this setting. Topics include, e.g., accounting for informative data structures in the context of model training and tuning, dealing with class imbalance in categorical outcomes, building effective prediction models by applying cutting edge ML methods, and performing feature selection in high-dimensional data settings. The presented methods will be motivated from a social and survey science perspective and critically discussed with respect to their advantages and limitations.

Code examples will be provided using the statistical programming language R.

Readings


Lists of required and recommended readings for each class are provided below for each specific unit.

**Academic Conduct**
Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland) and


Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

**Accommodations for Students with Disabilities**
In order to receive services, students at the University of Maryland must contact the Accessibility & Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; https://www.counseling.umd.edu/ads/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/.

**Course Evaluation**
In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.
Class Schedule

Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.

Unit 1: Intro: Bias-variance trade-off, cross-validation (stratified splits, temporal cv) and model tuning (grid and random search)

Video lecture: available Friday, May 29, 2020

Online meeting: Friday, June 5, 2020, 12 PM EDT/6 PM CEST

Online quiz 1: due Friday, June 12, 2020, 11 AM EDT/5 PM CEST

Required readings:


Recommended readings:


Unit 2: Classification: Performance metrics (ROC, PR curves, precision at K) and class imbalance (over- and undersampling, SMOTE)

Video lecture: available Friday, June 5, 2020

Online meeting: Friday, June 12, 2020, 12 PM EDT/6 PM CEST

Online quiz 2: due Friday, June 19, 2020, 11 AM EDT/5 PM CEST

Homework 1: due Friday, June 26, 2020, 11 AM EDT/5 CEST

Required readings:


Recommended readings:


Unit 3: Ensemble methods I: Bagging and Extra-Trees

Video lecture: available Friday, June 12, 2020

Online meeting: Friday, June 19, 2020, 12 PM EDT/6 PM CEST

Online quiz 3: due Friday, June 26, 2020, 11 AM EDT/5 PM CEST

Required readings:


Recommended readings:


Unit 4: Ensemble methods II: Boosting (Adaboost, GBM, XGBoost) and Stacking

Video lecture: available Friday, June 19, 2020

Online meeting: Friday, June 26, 2020, 12 PM EDT/6 PM CEST

Online quiz 4: due Friday, July 3, 2020, 11 AM EDT/5 PM CEST

Homework 2: due Friday, July 10, 2020, 11 AM EDT/5 PM CEST

Required readings:


Recommended readings:

Unit 5: Variable selection: Lasso, elastic net and fuzzy/recursive random forests

Video lecture: available Friday, June 26, 2020

Online meeting: Friday, July 3, 2020, 12 PM EDT/6 PM CEST

Online quiz 5: due Friday, July 10, 2020, 11 AM EDT/5 PM CEST

Required readings:


Recommended readings:


Unit 6: Support Vector Machines

Video lecture: available Friday, July 3, 2020

Online meeting: Friday, July 10, 2020, 12 PM EDT/6 PM CEST

Online quiz 6: due Friday, July 17, 2020, 11 AM EDT/5 PM CEST

Homework 3: due Friday, July 24, 2020, 11 AM EDT/5 PM CEST

Required readings:


Recommended readings:

https://www.datacamp.com/community/tutorials/support-vector-machines-r

Unit 7: Advanced unsupervised learning: Hierarchical clustering and LDA

Video lecture: available Friday, July 10, 2020

Online meeting: Friday, July 17, 2020, 12 PM EDT/6 PM CEST

Online quiz 7: due Friday, July 24, 2020, 11 AM EDT/5 PM CEST

Required readings:


Recommended readings:


Unit 8: Interpreting (Variable Importance, PDP, ...) and reporting ML results

Video lecture: available Friday, July 17, 2020

Online meeting: Friday, July 24, 2020, 12 PM EDT/6 PM CEST

Online quiz 8: due Friday, July 31, 2020, 11 AM EDT/5 PM CEST

Homework 4: due Friday, August 7, 2020, 11 AM EDT/5 PM CEST

Required readings:

Recommended readings:
