Hertie School **Master of International Affairs/Master of Public Policy**
Spring Semester 2021

Course Syllabus, Version 09.12.2020

GRAD-E1291: Machine Learning

Concentration: Policy Analysis

Slava Jankin and Hannah Bechara

1. General information

Class time	Mon, 16-18h
Course Format	This course is taught online only via the platform Clickmeeting/Teams. Clickmeeting/Teams allows for interactive, participatory, seminar style teaching.
Instructor	Slava Jankin and Hannah Béchara
Instructor's office	3.15 and 3.14
Instructor's e-mail	jankin@hertie-school.org, bechara@hertie-school.org
Instructor's phone number	Slava Jankin: +49 30 259 219 167 Hannah Béchara: +49 30 259 219 252
Assistant	Name: Alex Karras Email: karras@hertie-school.org Phone: +49 30 259 219 156 Room: 3.45
Instructor's Office Hours	Upon request

Link to Module Handbook **MIA** and **MPP**
Link to **Study, Examination and Admission Rules**

Instructor Information:
Slava Jankin is Professor of Data Science and Public Policy at the Hertie School. He is the Director of the Hertie School Data Science Lab. His research and teaching is primarily in the field of natural language processing and machine learning. Before joining the Hertie School faculty, he was a Professor of Public Policy and Data Science at University of Essex, holding a joint appointment in the Institute for Analytics and Data Science and Department of Government. At Essex, Slava served as a Chief Scientific Adviser to Essex County Council, focusing on artificial intelligence and data science in public services. He previously worked at University College London and London School of Economics. Slava holds a PhD in Political Science from Trinity College Dublin.

Hannah Béchara is an NLP post-doc who inadvertently found herself hired by Hertie's Data Science Lab. In between training neural networks and support vector machines, Hannah occasionally teaches programming classes in Python, the programming language for winners. She has previously been spotted teaching classes on NLP methods and Maths for Machine Learning. Hannah's current research interests include semantic relationships between words and phrases, and encompasses entailment, contradictions, and causal relations. Most importantly, Hannah plans to use NLP to

solve all of the world's problems.For reasons yet unclear, the University of Wolverhampton decided to award Hannah a PhD in Computer Science.

2. Course Contents and Learning Objectives

Course contents:
Machine learning is a core technology of artificial intelligence and data science that enables computers to act without being explicitly programmed. Recent advances in machine learning have given us, inter alia, self-driving cars, AlphaGo, Amazon, and Netflix. This technology has also allowed us to predict armed conflict and post-electoral violence, detect fake news, develop targeted provision of care and public services, and implement early policy interventions. This is a technical course that introduces core mathematical concepts and theory of machine learning. It also covers the implementation of these concepts using open-source software frameworks. The course provides a solid foundation for more advanced or more specialised study.

Main learning objectives:
By the end of this course students will have a sound understanding of the key theoretical concepts of machine learning and develop the ability to analyse data using some of its main methods.

Target group:
Students interested in developing strong methodological foundations for machine learning research and practice.

Teaching style:
Lectures covering theoretical concepts followed by practical lab sessions. This is an intensive course with a significant research component undertaken by the students.

Prerequisites:
Python Programming (E1326). You should also have familiarity with NumPy, Pandas, and Matplotlib.

Software:
We will be using production-ready Python frameworks like Scikit-Learn. In addition, for practical work we will make heavy use of Jupyter notebooks, Google Colab, and GitHub.

Diversity Statement:
As you may know, the Hertie School is committed to implementing a new Diversity and Inclusion Strategy. We strive to have an inclusive classroom but ask your informal feedback on inclusivity throughout the course.

3. Grading and Assignments

Composition of Final Grade:

Assignment 1: Project Proposal and Literature Review	Deadline: Session 4	Submit via Moodle	20%
Assignment 2: Midterm Report	Deadline: Session 7	Submit via Moodle	20%

Assignment 3: Final Report	Deadline: Session 11	Submit via Moodle	40%
Assignment 4: Presentation	Project Presentations: Session 12	Submit via Moodle	10%
Participation grade			10%

The assessment for the course consists of a research project, presentation and participation. The research project must be done in teams of 2-4 (individual submissions will not be accepted for the project). The aim is to develop research projects as close as possible to an academic publication in the area of applied machine learning and communicate your research to the broader public.

The aim of the assessments is three-fold:

- First, it will provide you with the opportunity to apply the concepts learned in this class creatively, which helps you with understanding material more deeply.
- Second, designing and working on a unique project in a team which is something that you will encounter, if you haven't already, in the workplace, and the project helps you prepare for that.
- Third, along with the opportunity to practice and the satisfaction of working creatively, students can use this project to enhance their portfolio or resume. We will discuss with individual project groups whether they can be turned into academic publications

Note about grading. There is no "perfect project." While you are encouraged to be ambitious, the most important aspect of this research project is your learning experience. Hence, you don't want to pick something that is too easy for you, but similarly, you don't want to choose a project where you are not certain that is out of the scope of this class. The project proposal is not graded by how exciting your project is but based on whether you follow the objectives of the project proposal, project presentation, and project report. For instance, if your project ends up being unsuccessful – for example, if you choose to design a classifier and it doesn't achieve the desired accuracy – it will not negatively affect your grade as long as you are honest, describe the potential issues well, and suggest improvements or further experiments. Again, the objective of this project is to provide you with hands-on practice and an opportunity to learn.

Assignment Details

Assignment 1: Project proposal and literature review (20%) – 3 pages and 5 references

- The main purpose of the project proposal is to receive feedback from the instructor regarding whether your project is feasible and whether it is within the scope of this class. Also, the project proposal offers a chance to receive useful feedback and suggestions on your project. The goal is for you to propose the research question to be examined, motivate its rationale as an interesting question worth asking, and assess its potential to contribute new knowledge by situating it within related literature in the scientific community.
- For the project, you will be working in a team consisting of 2-4 students. The members of each team will be randomly assigned by the instructor. If you have any concerns about working with someone in your group, please discuss it with the instructor.
- You must include a link to a GitHub repository containing the code of your project. Your repository must be viewable to the instructor by the submission deadline. If your repository is private, make it accessible to us (GitHub IDs *sjankin* and *hbechara*). If your repository is not visible to us, your assignment will not be considered complete, so if you are worried please submit well in advance of the deadline so we can confirm the repository is visible.

Furthermore, we will assess individual contribution to the team, should such an issue arise, based on the frequency and quality of GitHub commits in your project repository, so make sure you start the repository as the very first stage of your project.

- After you have received feedback from the instructor and your project proposal has been graded, you are advised to stick to the project outline in the proposal as closely as possible. However, if there is a concept introduced in a later lecture, you have the option to modify your proposal, but you are not penalized if you don't. If you wish to update your project outline, talk to the instructor first.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 2: Midterm report (20%) – 4 pages and 10 references

- By the middle of the course, students should present initial experimental results and establish a validation strategy to be performed at the end of experimentation. This serves as a project milestone. The milestone should help you make progress on your project, practice your technical writing skills, and receive feedback on both.
- Ultimately, your final report will be written in the same style as an ML research paper. For the midterm, we ask you to write a preliminary version of some sections of your final report. Producing a high-quality milestone is time well-spent, because it will make it easier for you to write your final report. You might find that you can reuse parts of your project proposal in your milestone. This is fine, though make sure to act on any feedback you received on your proposal.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 3: Final report (40%) – 8 pages and unlimited references

- The final report will include a complete description of work undertaken for the project, including data collection, development of methods, experimental details (complete enough for replication), comparison with past work, and a thorough analysis. Projects will be evaluated according to standards for conference publication—including clarity, originality, soundness, substance, evaluation, meaningful comparison, and impact (of ideas, software, and/or datasets).
- You ***must*** include a link to a GitHub repository containing full replication code of your project.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 4: Presentation (10%)

- At the end of the semester, teams will produce a blogpost (use this template: https://github.com/hertie-data-science-lab/distill-template) and pre-recorded video presenting the results of their work to the class and broader community. These will be posted on the Data Science Lab website.
- Detailed description of the presentation task will be made available on Moodle.

Participation grade (10%)

- We appreciate everyone being actively involved in the class. For full participation credit, we expect you to contribute relevant questions and ideas to the online class forum on Piazza, and answer questions from your peers. The top ~5 contributors will get full participation grade; others will get credit in proportion to the contribution of the ~5th person. Use your real name and your Hertie email address for participation credit. We will regularly show the "leaderboard" of contributors in class.

Any other act that improves the class for this year or subsequent years, which the instructor notices and deems worthy will receive 1% credit.

Late submission of assignments: For each day the assignment is turned in late, the grade will be reduced by 10% (e.g. submission two days after the deadline would result in 20% grade deduction).

Attendance: Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity: The Hertie School is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, deceitful actions as well as free-riding in group work are not tolerated. See Examination Rules §16.

Compensation for Disadvantages: If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating circumstances: An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructors and the Examination Office *in advance* of the deadline.

4. General Readings

- Aurélien Géron (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. 2nd edition. O'Reilly Media, Inc. [we'll designate it as **AG** throughout]
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. (2009). *The Elements of Statistical Learning: Data mining, inference, and prediction*, 2nd edition, available at https://web.stanford.edu/~hastie/ElemStatLearn/. [we'll designate it as **HTF** throughout]

5. Session Overview

Session	Session Date	Session Title
1	08.02.2021	Logistics, Software, Linear Algebra
2	15.02.2021	Probability and Statistics
3	22.02.2021	The Machine Learning Landscape
4	01.03.2021	End-to-End Machine Learning Project
5	TBD	Classification
6	15.03.2021	Training Models
Mid-term Exam Week: 22 – 26.03.2021 – no class		
7	29.03.2021	Support Vector Machines
8	12.04.2021	Decision Trees

9	**19.04.2021**	**Ensemble Learning and Random Forests**
10	**26.04.2021**	**Dimensionality Reduction**
11	**03.05.2021**	**Unsupervised Learning Techniques**
12	**10.05.2021**	**Project Presentations**
Final Exam Week: 17 – 21.05.2021 – no class		

6. Course Sessions and Readings

In the case that there is a change in readings, students will be notified by email.

Required readings are to be read and analysed thoroughly. Optional readings are intended to broaden your knowledge in the respective area, and it is highly recommended to at least skim them.

Lab sessions will use textbook notebooks: https://github.com/ageron/handson-ml2

Session 1: Logistics, Software, Linear Algebra	
Learning Objective	We discuss the importance of an original research project in machine learning in a team as the learning opportunity. We cover core components of the research project and related expectations. We also revise core concepts of linear algebra and calculus. Finally, we look at the computational tools that we'll be using in the course.
Required Readings	Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola. 2020. *Dive into Deep Learning*. URL: https://d2l.ai. Chapter 18.1 – 18.5, 19.1, 19.4.
Optional Readings	
Lab content	Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola. 2020. *Dive into Deep Learning*. URL: https://d2l.ai. Chapter 19.1, 19.4.

Session 2: Probability and Statistics	
Learning Objective	We continue revising the core mathematical foundations of machine learning and cover random variables, maximum likelihood, distributions, Naïve Bayes, statistical and information theory.
Required Readings	Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola. 2020. *Dive into Deep Learning*. URL: https://d2l.ai. Chapter 18.6 – 18.11.
Optional Readings	
Lab content	Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola. 2020. *Dive into Deep Learning*. URL: https://d2l.ai. Chapter 18.6 – 18.11.

Session 3: The Machine Learning Landscape	
Learning Objective	What is ML? Why use ML? We discuss the types of machine learning systems and their main challenges. We look at data quality is a key ingredient for success, and testing and validation strategies.
Required Readings	AG: Chapter 1.
Optional Readings	HTF: Chapter 2
Lab content	https://github.com/ageron/handson-ml2/blob/master/01_the_machine_learning_landscape.ipynb

Session 4: End-to-End Machine Learning Project	
Learning Objective	We discuss how to frame a research problem in machine learning, select performance measures, and check assumptions. We cover how to work with real-world data and create visualizations. We discuss how to train, fine-tune, and evaluate your models.
Required Readings	AG: Chapter 2.
Optional Readings	
Lab content	https://github.com/ageron/handson-ml2/blob/master/02_end_to_end_machine_learning_project.ipynb

Session 5: Classification	
Learning Objective	We discuss binary, multiclass and multilabel classifiers. We also cover relevant performance measures.
Required Readings	AG: Chapter 3.
Optional Readings	HTF: Chapter 4
Lab content	https://github.com/ageron/handson-ml2/blob/master/03_classification.ipynb

Session 6: Training Models	
Learning Objective	We are looking under the hood at the training algorithms of core machine learning models. We cover gradient descent and look at the model fit of several regression functions.
Required Readings	AG: Chapter 4.
Optional Readings	HTF: Chapter 4

Lab content	https://github.com/ageron/handson-ml2/blob/master/04_training_linear_models.ipynb

Mid-term Exam Week: 22 – 26.03.2021 – no class

Session 7: Support Vector Machines	
Learning Objective	We cover the core concepts of SVMs.
Required Readings	AG: Chapter 5.
Optional Readings	HTF: Chapter 12
Lab content	https://github.com/ageron/handson-ml2/blob/master/05_support_vector_machines.ipynb

Session 8: Decision Trees	
Learning Objective	We introduce decision trees and discuss how to train and visualize them. We cover CART and issues of stability.
Required Readings	AG: Chapter 6.
Optional Readings	HTF: Chapter 9.2
Lab content	https://github.com/ageron/handson-ml2/blob/master/06_decision_trees.ipynb

Session 9: Ensemble Learning and Random Forests	
Learning Objective	We discuss the most popular ensemble methods like bagging, boosting, and stacking. We also introduce random forests.
Required Readings	AG: Chapter 7.
Optional Readings	HTF: Chapters 8.7, 10, 15-16
Lab content	https://github.com/ageron/handson-ml2/blob/master/07_ensemble_learning_and_random_forests.ipynb

Session 10: Dimensionality Reduction	
Learning Objective	We discuss the curse of dimensionality and applications in high-dimensional space. We also cover core dimensionality reduction techniques.
Required Readings	AG: Chapter 8.
Optional Readings	HTF: Chapters 14.5, 14.7-14.9, 18

Lab content	https://github.com/ageron/handson-ml2/blob/master/08_dimensionality_reduction.ipynb

Session 11: Unsupervised Learning Techniques	
Learning Objective	We discuss approaches to deal with unlabeled data. We focus on algorithms for clustering and anomaly detection.
Required Readings	AG: Chapter 9.
Optional Readings	HTF: Chapter 14.3
Lab content	https://github.com/ageron/handson-ml2/blob/master/09_unsupervised_learning.ipynb

Session 12: Project Presentations

Final Exam Week: 17 – 21.05.2021 – no class