Fulbright School of Public Policy and Management
Master of Public Policy in Policy Analysis
Academic Year 2021-2023
Summer 2022

SYLLABUS
Data Science for Public Policy - 3 credits

TEACHING TEAM

Huynh Nhat Nam nam.huynh@fulbright.edu.vn
Vladimir Yapit Mariano vladimir.mariano@fulbright.edu.vn
Vo Tuan Kiet mpp22.kietvo@student.fulbright.edu.vn

CLASS MEETING TIME

8.30 am - 11.45 am, Monday, Wednesday, and Friday 06/06/2022 to 29/06/2022, inclusive.

OFFICE HOURS

Huynh Nhat Nam [HNN]	Monday, Wednesday, Friday [16.00 – 17.30]
Vladimir Yapit Mariano [VYM]	Monday, Wednesday, Friday [13.00 – 14.30]
Vo Tuan Kiet [VTK]	Offline - Monday to Friday [17.30 – 19.00]
	Online – Saturday & Sunday [9.00 – 10.30]

LEARNING OBJECTIVES

- Understand the fundamentals of machine learning and its potential applications in public policy analysis and decision-making support.
- Be able to organise a data analysis project using popular supervised and unsupervised learning methods with applications to policy evaluation.
- Visualise and describe data with advanced graphical presentation.
- Be able to read, understand and criticise academic publications or industry reports using machine learning methods.

COURSE DESCRIPTION

The first part of the course will introduce the fundamental principles and concepts underlying common algorithms in machine learning and their applications in business and policy. More specifically, students will first be introduced to the concepts of bias-variance tradeoff and overfitting, which are arguably the bedrock of developing effective predictive analytic models. Techniques to help identify overfitting (e.g. cross-validation) and eliminate overfitting (e.g. regularisation) of predictive models will be presented. While students may have been familiar with parametric approaches to predictive analytics (e.g. regression models introduced in

'Quantitative Methods' courses), this course aims to introduce students to non-parametric approaches (with a focus on classification problems), such as K-nearest neighbours, tree-based methods, and support vector machines. They will also be introduced to the principles and application of popular unsupervised methods, such as hierarchical and k-means clustering algorithms.

The second part of the course will introduce students to the use of off-the-shelf artificial intelligence libraries for image processing, with a focus (and hands-on exercises) on their use for satellite image processing, facial recognition and vehicle counting.

The courses will consist of lectures and computing exercises which are designed to help students practice and thus understand better the theoretical concepts presented in the lectures. The lectures are not intended to be mathematical intensive. Mathematical details will be provided just enough to help students understand the data science concepts and associated techniques. The programming language Python will be used to demonstrate these concepts and techniques. Students will be required to write codes in Python for exercises, assignments and the final exam. While a brief introduction to Python for data manipulation will be provided, it is recommended that students have prior knowledge of programming, either with Python or another language.

The course will be taught in English (without interpretation).

REQUIRED READINGS

- McKinney, W. 2017. *Python for Data Analysis, Data Wrangling with Pandas, Numpy, and IPython, 2nd edition*. O'Reilly. [McKinney, 2017]
- James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. *An Introduction to Statistical Learning – with Applications in R*. Springer. [JWHT, 2013]
- Provost, F., Fawcett, T. 2013. *Data Science for Business*. O'Reilly [PF, 2013]
- Python Data Science Handbook (Google Colab)
- What Is The LandSat Program and Why Is It Important? (usgs.gov)
- How LandSat Benefits Us (nasa.gov)
- Additional short reading on specialised topics (to be advised closer to the lectures)

ASSESSMENT

1. In-class Participation: 20%
Students will be provided with a daily set of discussion questions that we will cover in the lecture. Participation will be heavily influenced by the quality and sophistication of your answers to those questions.
This evaluation component includes presentations by groups (as part of Lecture 4) on the review of the topics of cross-validation, nested cross-validation, variable selection, regularisation and principal component analysis. These topics were introduced in Quantitative Methods 2.
Specifically, each group (maximum of 3 students) will be required to review and present one of the above topics, including theoretical concepts and illustrative examples. While students

can reuse the materials (including Python codes) presented in Quantitative Methods 2, they are encouraged to use new examples to give the class a refresh on these topics.

2. Problem Sets (group, x2): 40%
Two problem sets using real-world data and public policy problems. Students will submit a script file analyzing the data and describing the results. To get full credit the file must run without error.

3. Final Project and Presentations (group): 40%
For the final project (about 10 pages in length), students will provide preliminary research that involves the identification of a policy question, then explain the analytical method to answer the research question, an explanation of the data source and the tools and theories to be used to evaluate the research question. Students will be expected to review the literature, explain the theory, detail hypotheses, explain the data and the analysis, and implications for public policy. The goal is to help students familiarise themselves with all phases of an analytical project and equip them with the fundamentals for future applications. Students will present their final project in groups. The presentation should describe the research question, the data and modelling approach, and the analytical results.

COURSE SCHEDULE

Monday, 06/06/2022 ***[HNN]***
Lecture 1. Introduction to Python for Data Science (1)
Lecture 2. Introduction to Python for Data Science (2)
Readings. Chapters 5, 6, 7, 9 [McKinney, 2017]

First problem set released

Wednesday, 08/06/2022 ***[HNN]***
Lecture 3. Bias-variance tradeoff and overfitting
Lecture 4. Detecting and avoiding overfitting
- Review of cross-validation and nested cross-validation (as group presentations)
- Review of variable selection, regularisation and principal component regression (as group presentations)

Readings. Chapters 2, 5, 6 [JWHT, 2013]

Group presentation - review of topics covered in Lecture 4.

Friday, 10/06/2022 ***[HNN]***
Computing exercises

Monday, 13/06/2022 ***[HNN]***
Lecture 5. Classification - Linear discriminant analysis and KNN
Lecture 6. Classification - Tree-based methods for classification
Readings. Chapters 4, 8 [JWHT, 2013]

Wednesday, 15/06/2022 ***[HNN]***

Lecture 7. Classification - Support vector machines
Lecture 8. Classification - Model selection
Readings. Chapter 9 [JWHT, 2013], Chapters 7, 8 [PF, 2013]

Friday, 17/06/2022 ***[HNN]***
Computing exercises

Monday, 20/06/2022 ***[VYM]***
Lecture 9. Clustering methods - Hierarchical clustering
Lecture 10. Clustering methods - K-means clustering
Readings and Code:

- Hierarchical Clustering from CognitiveClass.ai (Google Colab)
- K-Means Clustering on Python Data Science Handbook (Google Colab)

First problem set due; Second problem set released

Wednesday, 22/06/2022 ***[VYM]***
Lecture 11. Introduction to satellite image processing (1)
Lecture 12. Introduction to satellite image processing (2)
Readings:

- What Is The LandSat Program and Why Is It Important? (usgs.gov)
- How LandSat Benefits Us (nasa.gov)
- LandSat in Python (earthdatascience.org)

Friday, 24/06/2022 ***[VYM]***
Lecture 13. Introduction to image processing (1)
Lecture 14. Introduction to image processing (2)
Reading and Code:

- Introduction to OpenCV (four notebooks on Google Colab)
- Google Colab to Access Webcam for Images and Video

Monday, 27/06/2022 ***[VYM]***
Lecture 15. Application of image processing - facial detection
Lecture 16. Introduction to image processing - vehicle counting
Reading and Code:

- Google Colab to Access Webcam for Images and Video

Second problem set due

Wednesday, 29/06/2022 ***[HNN, VYM]***
Group presentation on final project