Università degli Studi di Napoli

Dipartimento di Scienze Economiche e Statistiche

NPDE

Data and Coding Lab

Luca Coraggio and Armando Miano

Course syllabus

Course description

The course is designed to equip students with a broad understanding of computer software and statistical techniques pertinent to researchers in Economics and Statistics. The curriculum encompasses the following topics: Git version control system; R and Python programming languages; STATA software; cluster analysis; supervised learning; text analysis; and web scraping. The course is tailored for graduate students in economics who lack prior knowledge or experience in these specific areas. These topics are introduced in a generalized manner rather than focusing on technical details and are examined through practical applications using programming languages.

The course covers several topics, and students are required to deepen their knowledge independently, both through studying the references provided in class and through completing the assigned programming exercises. During the course, students will be asked to complete and submit three ungraded assignments. At the end of the course, students will take a final exam in two parts. The first part involves creating an original project, applying a selection of the topics covered in the course. In the second phase of the exam, students will have to present a randomly assigned colleague's exam project. The outcome of the exam will depend on both the quality of the personal project and the presentation of the assigned colleague's project.

The course counts 15 slots of 1.5 hours each.

Slots and Topics

Slot 1 (LC) Introduction and Git.

- Introduction: computer fundamentals; kernel, shell, terminal and command line operations; file system; file type and mimetype.
- Git: git basics; creating and managing local repositories; git remotes; branching system.

Slot 2 (LC) Review of R syntax

- basic syntax; data type and data structures; control flow (conditionals and loops); function; installing and loading packages.

Slot 3 (LC) Unsupervised learning

- Cluster analysis: introduction to cluster analysis; dichotomies in clustering (hierarchical vs. flat; hard vs. fuzzy; probabilistic vs. non-probabilistic); clustering approaches (cost-based; model-based; density-based; spectral clustering; hierarchical clustering).

Slot 4 (LC) Practical R session (Clustering)

- Tutorial implementing a clustering algorithm.
- **Assignment I**

Slot 5 (LC) Review of Python syntax

- basic syntax; data type and data structures; control flow (conditionals and loops); function; classes; installing and loading packages; virtual environments.

Slot 6 (LC) Supervised Learning

- Introduction to supervised learning and objectives; risk-minimization framework.
- Linear models (linear and logistic regression); Nearest Neighbor; Tree-based methods; Artificial Neural Network.

Slot 7 (LC) Practical Python session (Regression/Classification)

- Tutorial implementing a regression/classification algorithm.
- **Assignment II**

Slot 8 (AM) STATA software

- Introduction to STATA syntax; basic commands; cleaning operations; regression analysis; MATA (introduction).

Slot 9 (AM) Practical STATA session

- Tutorial STATA session.
- **Assignment III**

Slot 10 (LC) Web Scraping I

- Introduction to web scraping; web scraping with Python (urllib and requests library; beautifulsoup library); scraping static websites; common strategies to address blocking.

Slot 11 (LC) Web Scraping II

- Scraping dynamic websites (selenium library); semi-automated scrapers.

Slot 12 (AM) Text Analysis

- Introduction to text analysis; modeling corpus of documents; vocabulary methods; visualizations (bag-of-words); type of applications and modern methodologies (sentiment analysis, topic modeling, text tagging, word2vec).

Slot 13 (LC) Model Selection I

- Model selection in supervised learning; hyperparameter tuning; resampling methods; overfitting bias; bias-variance trade-off; generalization error.

Slot 14 (LC) Model Selection II

- Model selection in cluster analysis: selecting an optimal clustering solution; internal, relative, and external validation techniques; elbow-criterion; quadratic-score validation; resampling approaches.

Slot 15 (LC and AM) Final practical session

- Tutorial session merging multiple techniques shown throughout the course.

References

- Lecture notes and material distributed by the teachers.
- Chacon, S., Straub, B. (2014). *Pro Git* (2nd ed.). Apress.
- Hennig, C., Meila, M., Murtagh, F., Rocci, R. (2015). *Handbook of Cluster Analysis*. Chapman Hall/CRC.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer New York.
- Mitchell, R. (2018). *Web Scraping with Python* (2nd ed.). O'Reilly Media, Inc.