

Data Science Challenge

Prof. Michalis Vazirgiannis, Professor, AUEB, mvazirg@aueb.gr, 210 8203519

Dr. Ioannis Nikolentzos, post doctoral fellow.

Overview

The field of data science has emerged in response to the significant increase in the availability of data that took place in the last decade. This course will introduce students to this rapidly growing field and will equip them with tools for working with data. The course will give students the chance to use these tools on real-world data. The students will participate in a data challenge competition. The challenge will let students go through the complete data science process. More specifically, the course will help the students to:

- develop an understanding of the key technologies in data science.
- practice problem analysis and decision-making.
- gain practical, hands-on experience through the challenge.

Key Outcomes

By the end of the course, students will be able to:

- understand the whole process of extracting knowledge from data.
- apply data analysis techniques to real-world datasets.
- solve real-world data science problems using the principles and methods they have been taught.
- design and implement machine learning pipelines.
- work with textual data and with graph-structured data.

Requirements and Prerequisites

This is mainly a hands-on course. Students will spend a large amount of time on writing scripts and experimenting with different algorithms. Students are expected to be familiar with at least one programming language such as Python, Java and C++. We will use Python as the main programming language throughout the course, however, the students are allowed to develop their methods in any programming language they like. The course also assumes knowledge of core computer science concepts and also of basic machine learning concepts.

Required Course Materials

There is no required textbook. All course materials will be provided in class and available for downloading. Students will need to bring their laptops in class in order to try out interactively the material being presented in the lab sessions.

Books

There are several books on the subject and a lot of resources available on the Internet. We next list some books and research articles for those students who wish to delve deeper on the topics discussed in class:

- Hand, D.J., Mannila, H. and Smyth, P., *Principles of Data Mining*. MIT Press, 2001.
- Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., *Deep Learning*. MIT press, 2016.
- Zhang, A., Lipton, Z.C., Li, M. and Smola, A.J., *Dive into Deep Learning*. 2020.
- Goldberg, Y., A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research* 57, pp. 345-420, 2016.
- Hamilton, W.L., Graph Representation Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 14(3), pp.1-159, 2020.

Software/Computing requirements

For the purposes of the course, the students are advised to install Python 3 beforehand. Specifically, the students could install the Anaconda distribution of Python (which is available at: <https://www.anaconda.com/products/individual>) since it is oriented toward machine learning applications and contains several libraries that are necessary for the lab exercises (e.g., scikit-learn, NetworkX). The students should also install the TensorFlow 2 library which contains Keras, a high-level neural networks API designed to enable fast, user-friendly experimentation with deep neural networks. The students will download all the necessary datasets from the challenge's website.

Grading

This is a practical course. Hence, there will be no final exam. Instead, as mentioned above, the students will only participate in data challenge competition (i.e. a data science problem organized as a competition hosted on Kaggle). The challenge will be announced at the first lecture of the course. The challenge will run for a number of weeks and the students will have the opportunity to submit their solutions during that period. The final grade will be determined based on students' performance on the challenge, the approach they will follow, the report they will submit and the presentation that they will give. Late submissions (after the deadline has passed) are not accepted.

Attendance Requirements

It is very important that students attend classes from the beginning of the course. Class attendance is an important part of a student's educational experience. Students are required to participate in the learning process and interact with the instructor. Students' presence in the classes is essential to the liveliness of the course and can help students better understand the teaching material. Therefore, regular attendance is expected and considered mandatory.

Academic Integrity

Students are expected to demonstrate academic integrity in the context of the course. In particular, students are expected to acknowledge the work of other people and to give credit where they have used

other people's ideas, results, software, etc. Gaining unfair advantage, usually violating regulations, is strictly forbidden. For instance, students are not allowed to gain access to and copy the work of other students. Unauthorized collaboration is also not allowed. For instance, students may not work together on the data challenge unless they are members of the same team.

Course Syllabus

The following topics will be covered by the course:

1: Data Preprocessing, Feature Extraction/Engineering

One of the most important steps in building machine learning models is data preprocessing. In this session, the students will be introduced to basic data preprocessing techniques such as scaling, feature encoding, feature normalization and missing value imputation. The session will also cover feature selection techniques whose objective is to retain a subset of informative features as well as dimensionality reduction techniques which produce lower-dimensional representations of the input samples. Furthermore, the students will be taught basic feature engineering approaches such as how to generate new features by combining existing features.

2: Supervised Learning

This session will cover traditional supervised learning algorithms for both regression and classification tasks. Such algorithms include linear regression, the logistic regression classifier, decision trees, etc. The session will provide an in-depth overview of these machine learning models. Besides the algorithms, the session will also cover cross validation methodologies which can be employed to reliably evaluate predictive models, strategies for avoiding underfitting and overfitting and strategies for optimizing the hyperparameters of the different algorithms.

3: Deep Learning

This session will be devoted to deep learning which has produced extremely promising results for various tasks in recent years. The students will be first introduced to multi-layer perceptrons and to backpropagation, the central mechanism by which neural networks learn. The session will cover other more advanced architectures such as convolutional neural networks and long short-term memory networks. The students will also be introduced to regularization techniques such as dropout which can help to avoid overfitting and to layers that aim to address the vanishing/exploding gradient problem such as batch normalization.

4: Text Mining

The first part of this session will cover traditional approaches to text mining such as the bag of words representation and topic models. In the second part of the session, the students will be introduced to more advanced recently proposed representation learning algorithms which embed words and/or documents to vector spaces such as Skipgram and BERT. Furthermore, the session will cover recent deep learning architectures such as convolutional neural networks and transformers that can deal with different text mining tasks including text categorization and question answering.

5: Graph Mining

The students will be introduced to standard concepts from graph mining such as centrality measures, graph generators, degree distributions, etc. The students will also get a detailed picture of community detection algorithms whose objective is to partition the graph into clusters of nodes. The next part of the session will focus on graph kernels, a very popular family of methods for performing machine learning tasks on graphs. Finally, the session will cover modern graph representation learning techniques such as node embedding algorithms and graph neural networks which have attracted a lot of attention recently and have been applied with great success to many problems.