SAPIENZA
UNIVERSITÀ DI ROMA

Master Degree in Data Science

Graduation Journal
October 2017

This journal is to celebrate the first graduation of the master in Data Science at Sapienza University. Graduates describe their experience and the theses they have developed at the end of this course. In particular, it is a first effort to spread the importance and the impact that data science could have across all disciplines and businesses. The reported theses will deal with topics like neurosciences, education, graph theory, signal theory, document classification, machine learning, visual analytics and more.

MASTER DEGREE IN DATA SCIENCE

GRADUATION JOURNAL

SAPIENZA
UNIVERSITÀ DI ROMA

October 2017

INTRODUCTION

Lorenzo Lancia

It is a capital mistake to theorize before one has data.

— The Adventures of Sherlock Holmes (1892)

Data Science is by definition a multidisciplinary discipline. Taking students across different academic paths and guiding them towards new topics and techniques in computer science, engineering and statistics was the aim of this master course.

For us, students coming from different bachelor degrees with different skill sets, it is a real challenge to collaborate and combine those skills towards the completion of projects and assignments.

During these two years students do not just learn the modern techniques of data mining, machine learning and statistics, but also to work together as a group exploiting our differences and skills.

We are facing both theoretical and practical lessons, learning to write clean ad efficient code and enhancing our speeching and presenting skills to effectively deliver the insight we build from raw data. Within our internships we get a glimpse of the real world problems in both working and academic environments, in Italy and abroad.

Today we present to you our first master theses concluding the journey to graduation. This journal is meant as a little memento of their works and experiences.

1 | BRAIN NETWORK ANALYSIS

Brain network analysis through multilayer core decomposition

Michele Gentili

Welcome to the first graduation in Italy of Data Science. My name is Michele Gentili, future PhD Students in Computer Science at Sapienza and now ex-student of the first Data Science Master in Italy. My background, management engineering, taught me some computer science, however I was surprised by having to code so much during this master. Despite this I really enjoyed it and I finally had a tool to make real my ideas. By becoming a data scientist I have seen how flexible I became coding for many different domains. For example I was able to code for designing artificial intelligence models but also for handling medical data describing diseases.

When I got in the Data Science course I was so glad to have met such an international and challenging course that could boost my skills. It also gave me the chance to get in touch with the energetic world of start-up and new technologies. My colleagues and I, participated and got the first prize at the national competition in the conference of the Società Italiana Statistica (2016). I also won the Digital Hackathon organized by Accenture and EnLabs (2016). Finally, this course gave me the chance to spend the last year in Barcelona, studying and working at the same time, giving me the chance to present and publish my work at the ACM Digital-Health Conference (London 2017).

My thesis: the brain. One of the most powerful organs of the human body, yet one of the less understood by scientists. Many conjectures have been made about the brain, does it define our consciousness? How does it store and manage all the information it collects during our lifetime? Many other questions and dreams are behind this part of the body that seems to influence the vast majority of our life.

"Beauty is in the eye of beholder". This well known truth motivated our work. Indeed, even though there are and there have been thousands of millions of humans in the whole history, none of them has shared a completely equal pair of brains. The way information are processed is different, and thus, there's a different outcome in two equal situations, even from the very same person: the brain is a continuously changing structure.

Connections, not shapes or locations, are the best tool to understand similarities and functionalities of cortical areas; they constitute a dynamic structure, *"As the water's stream calmly shapes the riverbed, so does the neural activity to the connectome"*[1]. A new approach has been introduced with the connectome. It is no longer important to study the physiognomy of neurons, but the connections through which functional areas interact. The totality of this connections is called the connectome.

Our work goes through different analysis starting from general statistics summaries to common networks analysis ending with more advanced algorithms that can cope with multilayer graphs. Data are collected during a study on neural correlates of the LSD experience[2].

In this analysis it's possible to find some specific areas that are mostly affected by the drug and classify the effects of the drug on different individuals. In particular focusing on hidden connected cores of the networks, that might be remained unseen until now, give to doctors recommendations to start new researches.

1 Sebastian Seung. *Connettoma*. 2014.

2 Carhart-Harris, Robin L., et al. "Neural correlates of the LSD experience revealed by multimodal neuroimaging." *PNAS 113.17 (2016): 4853-4858.*

2 | A DATA-DRIVEN SOLUTION FOR EDUCATION

Build Structured Quizzes for Knowledge Assessment

Cristina Menghini

Who am I? Before digging into my project, let me introduce myself! I am Cristina who two years ago decided to change her studies, moving from Statistics to Data Science. Since the first moment, I got that Data Science is a dynamic field which everyday puts you in front of challenges and asks you to be dynamic as well! The first year I won a *Machine Learning* competition and I traveled to London for an IBM event. Then, I studied at EPFL, where I had the opportunity to collaborate, with start-ups and big organizations like Wikipedia, and get in touch with smart and prepared people. The whole thing fascinete me such that I chose to pursue the PhD in Computer Science at Sapienza. One of the thing I appreciate most about this field is the wide community it has around.

Quizzettone: In the educational framework, knowledge assessment is a critical component. Quizzes are the most widespread means to test a student's knowledge, among all levels of education. Due to their important role, it is worth to study in-depth how they should be built in order to be as reliable as possible.

Concretely, building a quiz means: generate a set of possible questions and choose a subset, which forms the quiz, such that we can have a trustworthy idea of the student's knowledge about a specific domain of interest. The process related to the question generation has been widely studied by a large number of researchers, however, the problem of picking a subset of *right* questions to build a quiz, has not been investigated yet.

Specifically, the latter with respect to some constraints, from the *quiz creator* viewpoint, requires lots of efforts and turns out to be a time consuming and combinatorics problem.

In this thesis, we define a set of requirements and characteristics a quiz should hold, and we provide an algorithm which helps teachers to quickly produce a well-structured quiz, which meets those requirements.

At first, we define our framework supposing that we have a set of articles we want the students to read, and a set of possible questions and the set of entities extracted from the articles. The first property we want is to be sure that the students can potentially get the maximum score reading the provided documents. It means that all the questions the quiz is made of, should be answerable by reading the texts. Moreover, we want to be sure that a student who scores the maximum has gone through all the documents. In other words, there should be at least one *dedicated* question per article, only answerable reading a specific article. The latter condition implies the former and that the minimum number of the questions should be at least equal to the number of articles. So, we establish as main requirement the presence of a dedicated question per article.

In addition,in order to ensure the students to have a high level idea of the domain of interest, we want to choose questions which are about entities that are globally, in the entire corpus, and locally, in the single article, relevant. Furthermore, we desire to pick questions in such a way that they are evenly distributed both across the entities and the articles, so that we avoid to have a quiz which focuses only on some topics of the domain. We want the questions we pick to cover all the articles and as many entities as possible. Basically, it is a combinatorics problem which can not be solved in polynomial time. Thus, an approximated solution is provided using a greedy algorithm whose outcome is a set of questions which maximizes an objective function that summarizes the characteristics we want the quiz to have fulfilling the aforementioned requirements.

So far, the evaluation of the outcomes has been done by comparing the structures of a random and a not-so-random quiz. In details, we compare the two varying the number of picked questions. Results show that when the number of questions per quiz is low, the random quiz tends to not cover all the articles, not testing the broad student's knowledge. Currently, a user study is running to see what can be the impact of these well-structured quizzes on real scenarios.

3 | VISUAL ANALYTICS FOR MACHINE LEARNING

Interpreting Black-Box Classifiers through Instance-Level Explanations

Paolo Tamagnini

How do computer make decisions? Can we trust artificial intelligence? How can we understand the reasoning behind machine learning algorithms?

In today's world those are important questions we need to answer in order to control the tide of technology. Artificial intelligence is used more every day to take important decisions affecting human life. In particular the growing volumes of big data are being used for *machine learning*: a branch of computer science which is able to make machines autonomously learn the patterns hidden within the data.

Given this framework, computers will often be able to succeed more efficiently than humans in predicting the future, for example when foreseeing election outcomes, or in classifying items, like when recognizing handwritten digits in images. Despite this, computers fail in explaining how and why they performed such tasks. In order to trust and improve the machine learning model, the human would like to have a greater understanding over the automatic process of performing those tasks.

To solve this problem I developed a software working as a research intern at New York University. The work was published at the Workshop on Human-In-the-Loop Data Analytics co-located with SIGMOD 2017 in Chicago, where I presented my work as a Sapienza master student.

My tool, called *Rivelo*, uses the techniques of *visual analytics* to explain decisions of a broad variety of machine learning binary classifiers through interactive visual representations. *Rivelo* leverages *instance-level explanations*: techniques that compute feature importance locally, for a single data item at a time. For instance, in a text classifier such techniques produce the set of words the classifier uses to make decisions for a specific

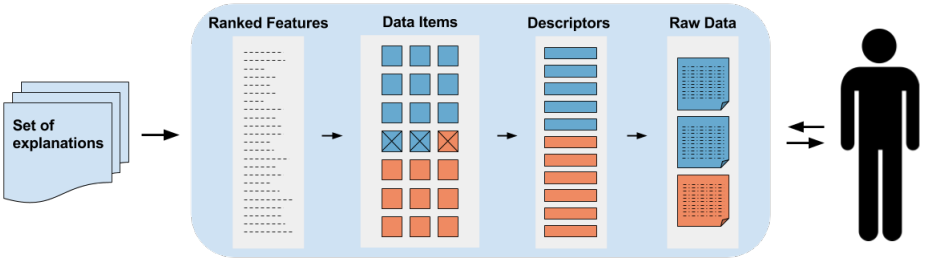

Figure 1: The explanation-driven workflow can extract global behaviours of the model from patterns of local anomalies through human interaction.

document. Those explanations are computed by treating the model as a black-box and observing how the output predictions are changing when modifying the input instances. The explanations are then processed and aggregated to generate an interactive workflow that enables the inspection and understanding of the model both *locally* and *globally*.

The workflow of *Rivelo* (Figure 1) consists of the following steps. The system generates one explanation for each data item contained in the data set and creates a list of features ranked according to how frequently they appear in the explanations. Once the explanations and the ranked list are generated, the user can interact with the results as follows: (1) the user selects one or more features to focus on specific decisions made with them; (2) the system displays the data items explained by the selected features together with information about their labels and correctness; (3) the user inspects them and selects specific instances to compare in more detail; (4) the system provides a visual representation of the descriptors / vectors that represent the selected data items (e.g., words used as descriptors in a document collection) and permits to visually compare them; (5) the user can select one or more of the descriptors / vectors to get access to the raw data if necessary (e.g., the actual text of a document).

The software and the relative paper is publicly available online.[1] [2]

1 *Rivelo* webpage: http://nyuvis-web.poly.edu/projects/rivelo.
2 Paolo Tamagnini website: https://paolotamag.github.io/.

4 | LEARNING GRAPH TOPOLOGIES FROM DATA

Application to brain functionality mapping

Elena Troccoli

Dear reader,
first of all, thank you for staying with us until the end of this graduation journal!

In this small article I will briefly talk you about my experience in the Data Science MS and about my final thesis, but first let me introduce myself.

I am Elena, one of the Data Science students at Sapienza University, and after a Bachelor's degree in Mathematics, I decided to start working in an IT consulting company. Two years ago, when I heard about the new MS in Data Science, I thought that it could have been a good chance to content my two souls: the one that is in love with beautiful and elegant mathematical theories, and the one understanding the distance between academic world and employment reality. So I decided to dive in this adventure...and I am glad I did. Not only subjects and environments were stimulating, I also had the chance to do an internship in a unique reality in Italy, such as the R&D department at UniCredit spa. Here, me and three colleagues of mine, under the guide of our supervisors, were able develop a work that brought us to WIMS2017 conference, where we presented our paper *HERMEVENT: A New Collection for Emerging-Event Detection.*

I have written my final thesis, *Learning graph topologies from data: application to brain functionality mapping*, working in a research group in the department of Information Engineering. The aim of the project was first to analyze and compare different solving approaches to task of associating a graph structure to a set of measurements; I then focused on the problem of inferring the functional network underlying a set of Electro-

corticographic (ECoG) recordings, in order to draw a brain functionality mapping, in relation to epileptic seizures.

Modern data processing tasks often deal with data whose intrinsic structure can be mathematically modelled by graphs. When the graph topology is not known a priori, it is handful to design strategies able to capture the hidden structure of observed signals.

Epilepsy is the world's most common serious brain disorder, defined by recurrent unprovoked seizures that result from complex interactions between distributed neural populations. Macroscopic characteristics of emergent ictal networks are explored by considering ECoG recordings from a human subject with intractable epilepsy. Graph learning strategies were employed to reveal functional dependencies among different brain regions, before and during seizure onsets. More specifically, we were interested in deriing time-varying graphs, in order to appreciate the evolution of brain functional network through time. Network analysis techniques have then been applied to analyze the topological organization of these results.

The study suggests that graphs derived from brain activity possess in general a small-world topology in which most connections are local and few are distant, and that small-world characteristics increase at seizure onsets. Furthermore, there is an evident pattern in the retrieved brain networks: pre-ictal networks are substantially more sparse (i.e. they have less connections) than ictal ones.

Unfortunately, when dealing with real-world data, we do not have a ground-truth to compare our results with. As someone said: "*all models are wrong, but some are useful*". Brain functionality mapping is an interesting application for graph learning models that deserves further investigations, and requires feedback from the neuroscience community as a fundamental step to proceed.

Master Degree in Data Science
Graduating session of October 2017, candidates:
Michele Gentili, Cristina Menghini, Paolo Tamagnini, Elena Troccoli

Contacts

Representative of Students
☞ lancia.lor@gmail.com

Master Degree Chairman
☞ datascience@diag.uniroma1.it

Candidates
☞ michele.gentili93@gmail.com
☞ cricri.menghini@gmail.com
☞ paolotamag@gmail.com
☞ elena.troccoli@gmail.com