Theses



Daniele Apiletti, Elena Baralis, Luca Cagliero, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, Danilo Giordano, Alessandro Fiori

General information

- Duration: 4-6 months full time
 - equivalent overall duration if part time

Internal thesis

- cooperation on active research topic or research project
- good programming and analytical skills required
- supervised by a group member
- can work at home or in our lab (LAB5)
- *External* thesis (stage)
 - supervised by external tutor

To get more info on specific topics please contact the reference person of the thematic area of interest by email (name.surname@polito.it)

Main topics

- Data mining and machine learning algorithms
 - Design and implementation of novel ML algorithms
- Data science pipeline
 - Design, personalization and implementation of KDD processes in diverse application areas
 - Industry, health, finance, ...
- Big data analytics
 - Design of scalable data mining and machine learning algorithms
 - Design of scalable KDD processes
- Database management
 - Data warehouse and NoSQL data modeling



Data mining and machine learning algorithms

- Clustering and semi-supervised clustering
- Scalable data mining algorithms for big data
 - E.g., (approximate) clustering
- Time series analysis
 - Forecasting models, trend detection
- Textual data analysis
 - Summarization, clustering, classification
- Predictive maintenance for
 - Industrial processes, robots, automotive components, ...



Data science pipeline

- Design and implementation of *personalized KDD* processes
- Black-box *prediction interpretation*
- KDD process automation by means of self-tuning and concept drift detection techniques
- Semantic enrichment by means of entity recognition and latent-based models
- Theory-guided data science, blending data-driven modeling with domain-provided knowledge



Big data mining

- Study of innovative, parallel, and distributed data mining approaches for
 - Pattern mining algorithms
 - Clustering techniques
 - Classification algorithms
- Design and development of novel parallel algorithms based on the Spark framework
- Exploitation of the novel algorithms for big data analytics applications (e.g., network traffic data, fraud detection, social networks)
- Analysis modules based on HADOOP and Spark Ecosystems





REFERENCE PERSON: PROF. PAOLO GARZA



Automation in the data analytics process

- Tailor the analytic steps to the different key aspects of the data under analysis
- Automate the analytic workflow to reduce manual user interventions
- Translate the domain-expert knowledge into automated procedures
- Automatically configure input parameters by means of self-tuning strategies
- Design informative dashboards and explanation techniques to support the translation of the extracted knowledge into effective actions

REFERENCE PERSON: PROF. TANIA CERQUITELLI

Conversational data science

Objective: making data analysis accessible to all and **democratizing the data science**, through:

- self-tuning algorithms
- friendly interaction between user and a data analysis engine through a conversational approach
- complexity of the analysis hidden to the user



Opens issue:

- define ad hoc self tuning strategies for each considered algorithm
- offer a friendly and dialogue-based interaction to the user, properly handling its requests
- choose the most suitable visualization to show the results to the user

REFERENCE PERSON: PROF. TANIA CERQUITELLI



- Predictive model performance usually degrades over time
 - New incoming data can widely differ from the data distribution on which the model was trained
 - Not all possible classes (labels) are known at training time
 - Real time predictions performed on new unseen data may be misleading or totally wrong

Explainable AI

- Understanding the behavior of machine learning models
 - From the individual and subgroup perspective



• Exploring the relationship with the desiderata of XAI research





REFERENCE PERSON: PROF. ELENA BARALIS

Explaining End-to-End SLU models

 Characterizing the behavior of End-to-End (E2E) Spoken Language Understanding (SLU) systems



- Understanding system behavior from the subgroup perspective
- Explaining model errors and anomalous behaviors



REFERENCE PERSON: PROF. ELENA BARALIS

Explaining Transformers

- Analyzing and explaining transformer models
- Focus on Natural Language Models



 Characterize shortcomings (e.g., gender bias) and propose regularization techniques

XAI in NLP

Goal: Designing innovative XAI solutions that offer a level of transparency greater than existing methods tailored to NLP

 We currently implemented an XAI solution, named T-EBAnO, producing local and global explanations for deep neural networks in text classification tasks



<u>Open Issues</u>

- A quantitative XAI evaluation methodology
- Comparison with other XAI techniques
- Extending the approach to tasks other than classification (e.g., Q&A)
- Exploit the XAI explanations for re-training the models in case of wrong predictions or correct predictions but for the wrong reason, with the human-in-the-loop feedbacks

REFERENCE PERSON: PROF. TANIA CERQUITELLI

Deep Natural Language Processing

- Vector representations of text data
 - Trained using Deep Learning architectures
 - Commonly used to address NLP tasks
 - Examples: Word2Vec, BERT, BART, PEGASUS, GPT-3
- Current issues
 - Summarization
 - Extract a concise descriptions of specific multimodal/multilingual data sources
 Patents, social posts, etc.
 - Semantic evaluation in NLP
 - Political and ethical hate speech detection, emotion detection in social media
 - Text style transfer
 - Reformulate text in natural language, e.g., from informal to formal style



REFERENCE PERSON: PROF. LUCA CAGLIERO

Transformers



- Multimodal training for representation learning
 - Learn unimodal representations embracing multimodal data



Multimodal fake news detection and verification
 Learn unimodal representations embracing multimodal data

Example of MULTIMODAL FAKE NEWS

The Ukrainian army has been destroyed almost entirely. There are very few professional soldiers left in the ranks of the Armed Forces of Ukraine. Ukraine also suffers from the shortage of military equipment and hardware





REFERENCE PERSON: PROF. LUCA CAGLIERO

Cross-lingual sentiment embeddings

- Extraction of embeddings based on the sentiment of words
 - e.g. for sentiment analysis tasks
- Not trivial for languages other than English
- Focus on the adoption of *propagation techniques* from English to other languages

x	x's nearest neighbors	Cosine similarity	10 9.5 9.9 9.8 9.9
excellent	eccellente	0.575	excellent
	ottimo	0.513	• 0.52
	apprezzabile	0.369	0.513
	buon	0.367	• accollanta
	adatto	0.322	ottimo



Time series forecasting

- Design multivariate supervised models for predicting time series directions/values
 - Application of regression/classification models in different domains
 - Seq2Seq, Reinforcement Learning, and Transformer models



Spatio-temporal data mining

- Problem
 - We are overloaded by heterogeneous spatio-temporal data
 - Satellite images and measurements (e.g., Copernicus data)
 - Ground-based sensor measurements
 - Etc.
- Thesis goals
 - Design and implement data mining algorithms for
 - Describing spatio-temporal phenomena
 - By means of sequential and/or graph-based patterns
 - Predicting spatio-temporal events
 - By means of graph-based patterns and ST-GNN (Spatio-Temporal Graph Neural Networks)

Theory-Guided Data Science

- Data science models, although successful in several domains, have limited applicability in scientific problems involving complex phenomena
- Theory-guided data science blends data-driven models with domain-provided knowledge
- Exploiting Physics-Informed Neural Networks (PINNs) to rewrite the way of learning of traditional models
- Examples: modeling physical processes, building energy management and sustainability, sustainable manufacturing, etc.







REFERENCE PERSON: DANIELE APILETTI



Digital Cultural Heritage

Development of a web-based digital atlas for historical buildings

- Store heterogeneous information (building structure, textual documentation, pictures, video, ...) on geo-located historical buildings
- Support geo-referenced queries to retrieve information on buildings located in a geographic area
- Profile user interaction with the system and guide end-user in identifying the information of interest



 Adopted technologies: data repository: MongoDB; backend: Django; frontend: JQuery, Bootrstrap, Leaflet, TinyMCE.

REFERENCE PERSON: PROF. SILVIA CHIUSANO

Apps for e-Education

designer.polito.it)

DesignER

- Web-Based Application for Computer-Aided Design of Relational Databases
- Guide users in the conceptual and logical design of a relational database through a tutoring-based approach
- Profile user design patterns and user mistakes in design process

REFERENCE PERSON: PROF. SILVIA CHIUSANO

Development of a platform for unplugged coding

- Coding courses are currently offered at pre-school and primary schools
- this thesis aims at developing a platform to help teachers to realize unplugged coding exercises and to be able to share them with the school community in a simple and intuitive way

REFERENCE PERSON: PROF. ALESSANDRO FIORI



Data analytics for healthcare

- Physiological data analysis
 - analyze physiological data collected to detect/predict discomfort conditions in automotive environnement
 - characterize discomfort conditions
 - reduce the effect of discomfort conditions









Summarization of Italian legal judgements

- Creation of judgement maxims
 - Brief description of the specific case
 - Principle of law affirmed in the legal judgment
- Examples: Word2Vec, BERT, BART, PEGASUS, GPT-3
- Entity recognition and disambiguation from Italian legal documents
 - Identification of the most relevant aspects covered by the text



REFERENCE PERSON: PROF. LUCA CAGLIERO

Manufacturing optimization

- Deep-learning techniques to predict KPIs of future products from highdimensional historical data of the manufacturing process
- Investigating context-aware machine learning solutions
 - E.g., reduce the overall duration, maximize throughput, shape energy consumption, etc.





Creativity-injection into AI-powered multimedia storyboards

Video Sequence



Machine Learning on HPC

Objective: Modeling and predicting the job/activity time on HPC by exploiting machine learning algorithms





Data Science Pipeline for Defect Prediction

Thanks to the Industry 4.0 the manufacturing **processes** expose many data about production

- conditions: environment conditions, wear and tear of the machines, etc.
- parameters: oven temperature, production speed, components characteristics, etc.



Conditions and parameters affect the outcome of the production line and can lead to **defects** in the final product

defects that are detected only at <u>the end</u> of the production by specific test procedures

The thesis aims to

- identify the most important conditions/parameters describing and healthy/defect production
- create a Machine Learning pipeline to predict the outcome (healthy/defect) based on the production conditions/parameters

The thesis leverages real production line data



REFERENCE PERSON: PROF. DANILO GIORDANO

Big Data Analytics for Financial Crime Detection

Each year, banking systems execute **hundreds** of **millions** of **transactions** to exchange **billions** of **dollars** between parties. In this scenario, **transaction monitoring** aims at analysing transactions to enforce Anti Money Laundering (AML) and Combating the Financing of Terrorism (CFT).

... current transaction monitoring solutions are mostly based on manual and rulebased investigations...

The thesis aims to explore new techniques for transaction monitoring based on Big Data and Machine Learning by

- Characterising real transaction data
- Reducing hundreds of millions of transactions to a subset of suspicious transactions
- **Proposing representation solutions** for transaction data
- Investigating techniques for **detecting anomalies** in transaction monitoring
- Develop a Machine Learning algorithm to flag suspicious transactions to increase the detection rate of the transaction monitoring task force

REFERENCE PERSONS: PROF. LUCA CAGLIERO, DANILO GIORDANO, ELENA BARALIS



29



Thesis internship in Bosch

Description

You will be inserted as intern in the Application team of Bosch, in the Turin offices, having the chance to experiment and acquire new competences in technical automotive field, always supported by your tutor inside the Bosch team. As part the team you will be involved in all the technical and innovation topic. You will be an active part in the definition and standardization of a process that has the scope to analyze and elaborate vehicle data using machine learning methodology.

Master Thesis Objects:

- application of machine learning in an automotive use case
- new AI function model development for Diesel Engine
- Bosch pipeline architecture improvement

Related Activities :

- Supervising the data acquisition process
- · Verify data quality, and/or ensure it via data cleaning
- Mange and process the data using Bosch standard architecture
- Explore and visualize data to gain an understanding of it
- Define\create test cases and data pipeline
- Understand today's test cases and create/develop new models

Qualifications:

- <u>Education</u>: Course of study in engineering (Computer Engineering, Automotive, Mechatronics or similar)
- Languages: English (minimum B2), Italian (minimum B2)
- <u>Know-how</u>:
 - \checkmark Proficiency with a deep learning framework
 - ✓ Proficiency with Python/Matlab and basic libraries for machine learning
 - ✓ Proficiency with Big Data
- Work style: meticulous, flexible, goal-oriented and team-worker.
- <u>Personality</u>: good communication skill, interest for innovation and technology in automotive.

Other information: There is a monthly repayment of 700€.



REFERENCE PERSON: PROF. ELENA BARALIS



30



Optimization of nuclear fusion reactor design

As the world seeks cleaner and more *sustainable energy solutions*, nuclear fusion has emerged as a promising contender. This thesis proposal aims to adopt machine learning techniques to make fusion reactors even better.

Objectives:

- Improve the design optimization, *exploring* currently adopted *design patterns* and *generating* new ones accordingly
- Develop a framework that integrates various types of simulations, to provide a more accurate and complete picture of the reactor's behavior, hopefully at a fraction of the computational cost!

Details:

Collaboration with the Department of Applied Science and Technology (DISAT), with Prof. F. Laviano, Dr. D. Torsello

Possibility of a period abroad (MIT)





REFERENCE PERSON: FLAVIO GIOBERGIA