Explainable Artificial Intelligence

00 – Administrivia

MSc in Artificial Intelligence

MSc in Computer Engineering

Marco Lippi

marco.lippi@unifi.it

UNIVERSITÀ
DEGLI STUDI
FIRENZE

UNIVERSITÀ DEGLI STUDI FIRENZE

Table of Contents

▶ Course organization

Course content (1/5)

Introduction and general concepts

- Safe and trustworthy AI: motivations and challenges
- Explainable AI: taxonomy, metrics
- Interpretability by design
- Explainability of black-box models
- Applications: justice, medicine, satefy-critical systems

Course content (2/5)

Interpretable models

- Early approaches: decision trees, generalized linear models, rule lists
- Explainable Boosting Machines
- Prototype-based models, concept embeddings
- Optimal decision trees
- Inductive logic programming
- Argumentation

Course content (3/5)

Post-hoc explanations

- SHAP
- LIME
- Attention, saliency maps
- Sensitivity analysis

Course content (4/5)

Neuro-symbolic AI

- Combining symbolic and sub-symbolic AI
- Markov Logic, DeepProbLog
- Concept Bottleneck Models

Course content (5/5)

Applications and practical sessions

- Python libraries for specific methods
- Data manipulation and visualization
- Model inspection
- Presentation of results, computation of metrics
- Real-world problems and case studies

Timetable

Classes

- Tuesday, 14:00-16:00, room S12 @ Santa Marta (**14:00-15:30**)
- Thursday, 16:00-18:00, room 177 @ Santa Marta (**16:00-17:30**)

Office hours

- Tuesday, 11:30-13:30 (office @ DINFO, Santa Marta)
- Or by appointment via email
- In any case, **always** write me in advance!

Teaching material

Teaching material will be made available on Moodle

- Slides (sometimes containing only topic summaries)
- Sketchbooks of whiteboards
- References to papers and book chapters

Please **subscribe** to the Moodle page

- Course information
- Timetable variations, announcements
- Updates to teaching material
- Information about exam organization

Exam

The exam consists in an **oral test** on the course content

- Registration via SOL
- Three dates in summer: June 17th, July 1st, July 21st
- One date in autumn: September 1st
- Three dates in winter/spring: to be decided

Project works

1 Course organization

Students that have chosen the Lab (3 CFU) for this course in their study plan will be proposed a project regarding the application of XAI techniques to real-world case problems

- Law
- Medicine
- Safety-critical systems
- Astronomy
- Bibliometrics
- ...