Ljubljana Doctoral Summer School

15 – 19 July 2019 (WEEK 1)

COURSE TITLE: Practical Introduction to Machine Learning and Data Analytics

ECTS credits: 4

Course schedule: from 9:00 to 13:00

Lecturers:
Zupan Blaž & **Pretnar Ajda** & **Toplak Marko**
University of Ljubljana, Faculty of Computer and Information Science, Slovenia

Aims of the course:

Machine learning algorithms are essential components of today's artificial intelligence and modern data analysis. Machine learning technologies provide cutting edge advantage to modern, data-driven companies, and help them to manage their relations with customers, discover new markets, or plan and optimize their products. In the course, we will introduce various machine learning techniques and present them in the setup of data analysis and modelling. While requiring no prior knowledge of statistics or computer science, we aim to cover a wide range of most useful and known machine learning techniques. The course will be entirely hands-on: we will use a modern visual-programming tool for data mining and solve some interesting practical problems on real-life data.

Course syllabus:

- Data preparation for machine learning
- Data visualisation, univariate vs. multivariate data visualisations
- Classification methods: classification trees, random forests, logistic regression
- Model evaluation: cross-validation and scoring, classification accuracy vs. AUC scoring
- Overfitting and cheating with testing on train data
- Feature scoring and selection
- Regression techniques: linear regression, polynomial expansion, overfitting, regression trees and forests

- Clustering: hierarchical clustering, k-means
- Data projection: principal component analysis, multi-dimensional scaling, t-SNE
- Deep learning-based embedding of text and images

Practical projects:

- Can we predict which project will be founded on the Kickstarter?
- Can we predict the market value of the house?
- Who's the author of a book? Or of a tweet? (text mining)
- Can we cluster or classify the products based on their images alone? (image analytics)

Tentative schedule:

Monday, 9 July:

Data preparation, data loading, classification trees (+ play with your own data)

Tuesday. 10 July:

Cheating with classification trees, model evaluation, forests, feature ranking (+Kickstarter project success prediction)

Wednesday, 11 July:

Regression, overfitting, regularization (+ prediction of the market value of a house)

Thursday, 12 July:

Clustering, data projection (+ customer segmentation)

Friday, 13 July:

Introduction to text mining and image analytics (+ author classification, + image maps)

List of readings:

Lecture notes (~80 pages) will be provided for course participants at the start of the course. No reading or extra preparation prior to the course is required, though participants are welcome to check first few videos from Orange Data Mining YouTube channel (http://youtube.com/orangedatamining).

Teaching methods:

This is a practical, hands-on course. Participants are required to bring their laptops and before the course install Orange Data Mining software from https://orange.biolab.si. From the first hour of the course on we will work on real-life data and introduce machine learning and data visualization techniques on practical problems.

Lecturers' Biographical Note:

Dr. Blaž Zupan has worked on machine learning seemingly forever. He heads the bioinformatics lab at the University of Ljubljana and is a Visiting Professor at the Baylor College of Medicine in Houston. He believes that crafting simple tools that anybody can use to understand data is essential to advancements in humanity and democracy. His lab developed Orange (http://orange.biolab.si), an open-source, ever-evolving data mining suite with a visual programming environment. He also enjoys writing scripts for YouTube videos to explain data science, and preparing courses that introduce data science to students of engineering and humanities.

Ajda Pretnar, M.Sc., is an anthropologist and a Ph.D. student who has lately been organizing courses with Orange worldwide and specializes in digital humanities and text mining.

Dr. Marko Toplak is a senior researcher in machine learning and a long-time developer of Orange.