ESSLLI 2013

August 5-16, 2013, Düsseldorf, Germany

Course proposal

Personal information

Proposers

Barbora Hladka, Martin Holub

Affiliation

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Malostranske namesti 25
118 00 Prague 1
Czech Republic

E-mail: {hladka, holub}@ufal.mff.cuni.cz
Tel. +420-21914223
Fax: +420-21924305

Home page: http://ufal.mff.cuni.cz/~hladka/ML.html

General proposal information

A Gentle Introduction to Machine Learning in NLP using R

An introductory course

Contents information

Abstract

The course provides a concise introduction to principles and algorithms of machine learning in natural language processing both theoretically and practically. We will focus on fundamental ideas in machine learning and the basic theory behind them. The principles of machine learning will be presented in a gentle way so that students do not have to be afraid of scary mathematical formulas.

We will give a brief introduction to the R system for statistical computing, which we use as a tool for practical demonstration. Students will gain practical know-how needed to apply the machine learning techniques to new problems.

The presented methods of machine learning will be practically demonstrated on selected tasks from the field of natural language processing. Understanding these tasks will not require any extra linguistic knowledge. We will show how to master NLP tasks using the R system and how to experiment with real data sets.

Basic description of the course

The course provides an introduction to machine learning (ML), both theoretically and practically. It is designed for students who:

- want to be familiar with ML and either are completely unfamiliar or have only a slight idea what ML is on;
- do not have an ambition to become an ML expert after a week course but want to get in the picture;
- have basic knowledge of math and computer science that will make the understanding of the ML fundamentals possible, including the practical aspects.

We will present selected concepts step by step and will not dive into deep details. That is why we advertise our course as a *gentle* introduction. We formulate three general goals to reach:

1. theoretical introduction to ML principles and algorithms,
2. practical introduction to ML methods using the R system,
3. demonstration of ML techniques on selected NLP tasks.

Each of the three goals is accompanied by our ambition what students should take home from the course.

Goal 1 – Introduction to the theory of ML

Machine learning is a field of computer science in which a computer program analyzes large collections of data and makes predictions. In other words, we teach computers (machines) to do various tasks using ML techniques based on analysis of data examples. ML exploits many fields of science, especially mathematical statistics, algebra, probability theory, and information theory.

We will concentrate on fundamental concepts in ML, not on proving theorems or detailed explanation of mathematical formulas. On the other hand, we will underline the fact that ML principles and algorithms are underlain by profound theoretical frameworks.

Our ambition: Students will learn what ML is on, what are the fundamental concepts and why it is useful to know it.

Goal 2 – Introduction to practical aspects of ML

Practical aspects of ML deal with the design and implementation of computer programs that read collections of data of different types and learn particular knowledge. Then the knowledge learned from the example data can be applied to new data instances. Such computer programs already exist and are available as libraries, such as the R system for statistical computing, which perfectly fits our needs.

Our ambition: Students will learn that they do not have to implement ML algorithms themselves because such systems are are available. Students will also learn how to use the R system.

Goal 3 – Demonstration of ML techniques on selected NLP tasks

ML principles and algorithms are application independent. We will address NLP tasks and experiment with English texts. The NLP tasks used for demonstration are easy to explain even to students who have not attended any computational linguistics course yet. Therefore knowledge of English is the only prerequisite. The examples have been carefully selected with respect to their gradual complexity.

Practical demonstrations will pursue the following 'what to do' scenario:

- input data analysis and/or interpretation;
- formal description of data examples;
- selection of a suitable ML algorithm;
- training process and parameter optimization;
- performance evaluation;
- result analysis: visualisation, interpretation, comparison of different methods, error analysis;
- making conclusions.

Our ambition: Students will learn how to design, run, evaluate an ML experiment from the very beginning to the end.

Tentative outline

- We think machine learning
    - Non-technical survey of ML,
    - first steps in R.
- Basic machine learning concepts more formally
    - Classification task, classifier as a function,
    - training and test data,
    - random variables and feature vectors,
    - evaluation and overfitting.
    - Machine learning algorithm #1: Decision trees
        - basic theory,
        - implementation in R.
- Machine learning algorithm #2: Naive Bayes
    - Basic theory,
    - implementation in R including parameter tuning.
- Machine learning algorithm #3: SVM
    - Basic theory,
    - implementation in R including parameter tuning.
- A closer look at selected key machine learning issues
    - Summary of machine learning algorithms #1, #2, #3 and their comparison from various aspects, e.g. bias, training data manipulation, computational complexity.
    - Presentation of tasks that demonstrate differences between the algorithms with respect to their (un) suitability to address a given task.
    - Overfitting and feature selection.

Expected level and prerequisites

- No required knowledge of machine learning or the R system.
- Required fundamental knowledge of algebra, statistics and probability theory (only level of introductory courses is expected)
- Required basic knowledge of general programming techniques (most common data structures, simple data manipulation etc.)

Appropriate references

1. Emms Martin and Saturnino Luz. Machine learning for Natural Language processing. *ESSLLI 2007* Course Reader. 2011. Available at http://ronaldo.cs.tcd.ie/esslli07/mlfornlp.pdf.
2. Kerns G. Jay. Introduction to Probability and Statistics Using R. 2011. Available at http://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf.
3. Short Tom. R reference card. 2004. Available at http://ufal.mff.cuni.cz/~hladka/info-on-lecture/R-Short-refcard.pdf.

Practical information

Relevant preceding meetings and events

Both lecturers have a lot of experience acquired while teaching one-semester introductory course on ML in NLP since 2003 - for more details see http://ufal.mff.cuni.cz/~hladka/ML.html. The course is given in English and is attended by both Czech and international students (under the programme European Masters Program in Language and Communication Technologies).

Potential external funding for participants

Charles University in Prague, Faculty of Mathematics and Physics will cover travel expenses, accommodation and per diems for one of the two lecturers.

Relevance to ESSLLI

Machine learning is an interdisciplinary field that hosts both linguistics and computer science (beside others). In the train of it, we offer a course for "interdisciplinary" students. A course is designed as *3-in-1*: introduction to machine learning, introduction to the R system and a practical guide to employing ML techniques in NLP.