Overview of the teaching sessions

[Salta al contenido principal](https://egela.ehu.eus/mod/page/view.php?id=4325205&inpopup=1#maincontent)

After Olatz Arbelaitz and Alicia Perez's lessons on initial machine learning topics in the "Introduction to Machine Learning", it is my turn for the following days. They have shared with me the topics covered in class, and I will try to go one step ahead: and locate this subject between "Introduction to Machine Learning" and "Deep Learning" subjects, trying to enlarge your theoretical-knowledge and software skills in the world of data science. Trying always to ilustrate them with NLP applications.

First of all, I will present myself. My name is Iñaki Inza, and I complete my research activities in the ["Intelligent Systems Group"](http://www.sc.ehu.es/ccwbayes/isg/) of the Computer Science Faculty, UPV-EHU: My main research areas and profile can be found here [[personal webpage](http://www.sc.ehu.es/ccwbayes/members/inaki.htm)] [[GoogleScholar profile](https://scholar.google.com/citations?user=ogYjUPAAAAAJ&hl=es)]

Research on novel techniques for machine learning and data analysis, together with applications in environmental sciences and bioinformatics, are my main reseach areas. A single journal application in NLP, "[Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers](http://www.sciencedirect.com/science/article/pii/S0925231212001282)".

During the next teaching days in the "Machine Learning" course in your EMLCT-HAP master, I will try to show you data analysis tools that can be useful in your everyday NLP practice. I will try a have a practical-application perspective. The presentation of the machine learning techniques will be coupled with applications over [NLP benchmark datasets](http://www.sc.ehu.es/ccwbayes/master/selected-dbs/nlp-naturallanguageprocessing/): another link in our egela describes each dataset. The exercise will be performed using popular machine learning software tools such as R, WEKA...

The following is a list of the machine learning tools that we will cover during the next sessions: 

1. General terms on the "data science" world: the "data science" term, relation among AI and data science, the big data term, kaggle repository, kdnuggets.com, data science for a better world...
2. Principal classification scenarios: supervised classification, unsupervised classification (clustering), weakly supervised classification (alternative scenarios).
3. One-class classification and outlier detection: usefulness in NLP tasks. Software, R packages.
4. Semi-supervised classification: usefulness in NLP tasks. Software, RSSL package in R.
5. Multi-label classification: a crucial scenario in NLP. Software: R.
6. Using statistical tests to compare the accuracy of different classifiers. Software: WEKA, R, web pages
7. Feature selection techniques. Techniques for selecting a "competitive" subset of original features.
8. General techniques and filters for data preprocessing. Preprocessing filters for any kind of data: missing data imputation, one-hot encoding, discretization, imbalanced class distributions...
9. "A short introduction to the tm (text mining) package in R: text processing". How to contruct by text mining operators a proper document-term matrix for further machine learning analysis. Starting from raw text such as files, html pages, twitter... A tutorial using R software.
10. "The machine learning approach: clustering words and classifying documents with R". A tutorial using R software.
11. "First steps on deep learning for NLP by R’s h2o package (+word2vec)". A tutorial using R software

While the first 5 items are more data-analysis-general oriented, I consider the last three topics tightly fitted to your NLP interests. However, during the first 4 items, we will try to use NLP datasets to illustrate the use of the teached data analysis "general purpose" techniques.

Última modificación: martes, 7 de diciembre de 2021, 11:48