ML4LT (2016)
Marina Santini
Department of Linguistics and Philology
Uppsala University

ML4LT (2016): Assignments - *Undergraduate Students*

Refresh the page: Changelog 12 Dec 2016

The Machine Learning for Language Technology course is examined by means of 3 home assignments (check the assignments' deadline on the course website).

TIP: Before starting working on the assignments, review all the lab tasks and build upon the practical experience you gained during the lab sessions. We used several datasets of different nature and composition that highlighted different views and interpretation of data and ML algorithms. The lab documents contain step-by-step instructions, tips and useful references.

Assignment 1: Decision Trees and k-Nearest Neighbours

This assignment is about learning **morphological information**, namely the different classes of English verbs.

- Remember that accuracy can be a misleading metric to assess the goodness of a classifier.
- Use visualization to understand the behaviour of the classifiers.
- Tweak parameters and see what happens.
- Did any classifier(s) perform statistically worse than the reference classifier?
- Did any classifier(s) perform statistically better than the reference classifier?
- Which classifier(s) classified most correct instances?
- Which classifier(s) classified least correct instances?
- How many correct instances did the reference classifier classify (in percent)?
- How many incorrect instances did the reference classifier classify (in percent)?
- Did any classifier had unclassified instances?
- Draw ROC curves, cost/benefit analysis and apply any other evaluation metrics that you think is useful to make the right choice.
- Think of the effect of the inductive bias on classification results.
- etc! *be creative, but always ground your interpretentions on data and evidence.*
- Conclusion: Which classifier, in your opinion (**based on some evidences**), has highest performance on the dataset?

Assignment 2: Naive Bayes

This is a sentiment classification task on a tweet dataset. We used a metaclassifier and several filter in class. Finding sentiment orientation or mining opinions in the social media is very trendy branch of LT. You will study more on this topic in the SAIS (semantics for LT) course.

Marina Santini
Department of Linguistics and Philology
Uppsala University

Assignment 3: k-Means and Hierarchical Clustering

This is an assignment on text classification. You will apply unsupervised machine learning to classify the text categories included in the Swedish national corpus (SUC).

Among other things, the assignment asks you to investigate whether you can observe different behaviours depending on the **type** of text categories, namely *genre classes* and *domain classes*.

Keeping these two types of textual categories apart has proved to be useful in several LT/NLP fields, such as Machine Translation. If you are interested in this topic, here are a couple of reading suggestions:

- van der Wees, M., Bisazza, A., & Monz, C. (2015). Translation model adaptation using genre-revealing text features. *DISCOURSE IN MACHINE TRANSLATION*, 132.
- van der Wees, M., Bisazza, A., Weerkamp, W., & Monz, C. (2015). What's in a domain? Analyzing genre and topic differences in statistical machine translation. In Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP (Short Papers) (pp. 560-566).

Good Luck!
Lycka till!
Viel Glück!
In bocca al lupo* !

* Wikipedia: "In bocca al lupo" (Italian pronunciation: [im 'bokka al 'lu:po], "in the mouth of the wolf") is an Italian idiom used in opera and theatre to wish a performer good luck prior to a performance. The standard response is "crepi il lupo!" or, more commonly, simply "crepi!" (Italian: ['kre:pi il 'lu:po], "may (the wolf) die"). Equivalent to the English actor's idiom "Break a leg", the expression reflects a theatrical superstition in which wishing a person "good luck" is considered bad luck.

--- the end---