Machine Learning for Language Technology
(Term: Autumn 2016)
Course website: http://stp.lingfil.uu.se/~santinim/ml/2016/ml4lt_2016.htm

ML4LT Assignments - Debriefing

23 Jan 2017

In this document, some final reflections on the experiments included in the assignments are presented. These reflections are meant to complement interpretations and comments that students have put forward in their submissions. This document is also an opportunity to suggest additional reading. ☺

Assignments' Goals

The general goal of the three assignments is to assess what students have learnt after:

- **the online lectures and in-class labs**
- **having read the required course literature**

An additional goal of the three assignments is to assess to what extent students have developed *independent thinking* on the topics covered during the course.

Expected Learning Outcomes (2016):

* apply basic principles of machine learning to **natural language data**;
* **evaluate the performance** of machine learning schemes;
* use standard **off-the-shelf software** for machine learning;
* apply **supervised** and **unsupervised** models for classification.

Remarks on the results of the experiments

All the experiments in the three assignments are designed to produce problematic results. Students are supposed to orientate themselves using what they learned in the course, and provide plausible interpretations.

The results of the experiments in **Ass1** were the most misleading and showed high accuracies but weak classifiers;
The results of the experiments in **Ass2** were disappointing to the students, no matter how they tweaked the features;
The results of the experiments in **Ass3** were hard to interpret.

Often finding the best supervised and unsupervised ML methods for a set of data is based on a trial-and-error approach. In order to interpret the results, we must have a full grip on how to interpret evaluation measures. Mathematical algorithms can be fascinating in themselves, but they only represent a part of the whole story. When we apply supervised and unsupervised ML to data, we should always keep in mind that there are *at least* three elements interacting: a mathematical algorithm, a dataset and set of features.

Mathematical algorithms, although they all have different inductive biases, can be tamed and tweaked using parameters. This was experienced when students forced J48 and k-NN to output the same results, as asked in Question 6, Ass1.

A dataset represents a statistical population. Statistics is sometimes tricky, and Daume', Section 5.1, presents a clear example of how learning is affected by an unbalanced dataset. The experiments of **Ass1** return very high Acc, P, R, F-score, but the classifier is basically blunt since it

systematically misclassifies IRR-classes. Experiments in Ass1 show empirically how these measures are all, somehow, biased when the dataset in unbalanced (see additional reading: Power, 2011). K statistic and/or ROC curves and/or visualization tools can, however, help us unveil deceptive results.

There is not one-fits-all remedy when dealing with unbalanced datasets such as the past tense dataset. Since this dataset contains an overwhelming majority of REG verbs in one single class, and many differying IRR classes, maybe it would be worth trying collecting all the IRR classes in a single IRR class and perform a binary classification, ie REG vs IRR. If this is ineffective, more sophisticated approaches (that we have not dealt with empirically in the course) can be envisaged (eg. see Daume', Sect 5.1 , and also the links in the Additional Reading section). Presumably, the unbalanced distribution of the past tense dataset is realistic and well represents the real distribution of verbs in natural languages (since the number of regular verbs is normally higher than the number of irregular verbs).

The experiments in **Ass2**, on the other hand, are based on a perfectly balanced dataset. But for some reasons, performance is just above 50% (which is disappointingly low for a 2-class classification problem) and, consistently, k statistic is also quite low. Tweaking feature thresholds provides marginal improvements. Here there are several hypotheses we can make, eg. features are not representative of the classification problem (are these words good enough to detect sentiments?); the ML algorithm is not ideal for this type of data, the number of features is too low or too high; the sample (i.e. the dataset) is too small, etc. Students have identified some of these problems in their submissions.

The experiments in **Ass3** are easy to run, but the interpretation of results is challenging. We know that we cannot expect a high performance from unsupervised clustering, because clustering algorithms learn without information about the class label of data. The primary advantage of unsupervised methods is that they do not require annotated data (always very expensive in terms of time and financial resources) to learn a model. However, it is often difficult to interpret the results and to evaluate them against a manually labeled gold standards such as the SUC (see additional reading about evaluation of unsupervised methods against gold standard, Vlachos (2011)).

The SUC is the Swedish National corpus and as such it is supposed to be representative of the Swedish language (of the 90s) as a whole. The language varieties are represented in terms of different textual categories. The corpus was built for linguistic studies and not for ML purposes. This leads to a *first reflection*: we do not know whether the fine-grained class distinction made by linguists makes sense to mathematical algorithms.

SUC textual categories are a mixture of domains and genres. In the assignments we interpret domains as topic-related categories (such as "Skills") and genres as categories characterized by textual conventions (such as "Academic writing") . In the assignment, we make the hypothesis that readability features "represent" the different textual categories of the SUC well. In practical terms, we are basically saying that the different textual categories in the SUC can be recognized based on how difficult or how easy it is to read the different texts in the corpus. If you have a look at the readability features in Falkenjack et al. (2013), you will see that the feature set is mostly based on stylometric cues, morphology and syntax, but few lexical features. This might trigger a *second reflection*: since domains are topic-based we might think that *content words* would do a better job on certain SUC categories and that grammatical features are better suited to account for the grammatical conventions of certain genres. Remember what Daume' says about the importance of good features (Sect. 4.1). Finding good and representative features is not trivial.

The automatic classification of the SUC is a difficult problem, also for supervised algorithms, as shown in Falkenjack et al. (2016). The SUC makes sense for humans, but less sense for ML algorithms, because it contains classes of different nature.

Both k-means and hierarchical clustering give some hints, however, about what to do next, since they perform consistently better on certain classes and consistently worse on other classes. But the readability features seem to be too weak to show clear-cut patterns. One possible next step would be to try out a separate feature set for genres and a separate feature set for domains. For examples, domains could be represented using the bag-of-word approach, while genres could be represented using POS n-grams… (just guessing).

At present, unsupervised methods (including more sophisticated algorithms such as topic modes) are knowing increasing popularity because the internet is full of unlabeled data to make sense of.

There are many other interesting observations that could be made on these assignments, but it is now time to close the course. Feel free to ask any additional questions you might have.

Congratulations on your achievements!

Additional Reading

Powers, David Martin (2011). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011). Journal of Machine Learning Technologies, Volume 2, Issue 1, 2011, pp-37-63 <http://www.bioinfopublication.org/files/articles/2_1_1_JMLT.pdf >

On the Classification of Imbalanced Datasets
<http://www.ijcst.com/icaccbie11/sp1/krishnaveni.pdf>
Crossvalidated: Class imbalance in Supervised Machine Learning <http://stats.stackexchange.com/questions/131255/class-imbalance-in-supervised-machine-learning >

Vlachos, A. (2011, July). Evaluating unsupervised learning for natural language processing tasks. In Proceedings of the First workshop on Unsupervised Learning in NLP (pp. 35-42). Association for Computational Linguistics.

--the-end--