Lecture
"Foundations of Language Science and Technology: Semi-Supervised Learning"

Prof. Dr. D. Klakow

Exercise

Sample Solution

1. Why is semi-supervised learning useful?

ANSWER: *Semi-supervised learning only requires a small amount of labeled data instances. Labeled data instances are usually hard to obtain since they require manual annotation. The other type of data used in semi-supervised learning, unlabeled data, on the other hand, is usually fairly cheaply to obtain.*

2. Name a semi-supervised learning algorithm that has been explained in the lecture and explain why it can be regarded as a bootstrapping algorithm.

ANSWER: *Two algorithms could have been described in the answer for this question: the Yarowsky algorithm and the EM algorithm. In both algorithms a simple weak classifier exclusively trained on the labeled data represents the 'simple system'. The construction of the 'other more complicated system', i.e. the classifier that takes into account both labeled and unlabeled training data is different. In the Yarowsky algorithm, at each iteration only the most confident unlabeled data instances are given a hard label (i.e. the classification is definite). In EM, all unlabeled data instances are soft-labeled (i.e. $P(c_i|d_j)$ is computed for each class c_i and all of these probabilities become part of the model of the subsequent iteration).*

3. Explain why a semi-supervised classifier heavily relies on the correlation between features observed in the labeled data instances of a specific class c_i and the other features only observed in the unlabeled data instances of c_i.

ANSWER: *This correlation is vital for propagating correct labels from labeled data instances to unlabeled data instances. Little knowledge can be drawn from the labeled data which is compensated by taking the clustering of data instances into account. Labeled and unlabeled data instances of a particular class only cluster if their underlying features sufficiently co-occur.*

4. Explain how a semi-supervised classifier might be led astray.

ANSWER: *A bad feature x_i which coincidentally correlates with labeled data instances of class c_j causes the classifier to infer (erroneously) other features x_k in the unlabeled data instances correlating with x_i to be of class c_j. Likewise, x_k might also trigger further false correlations for c_j (error propagation).*

5. Why is feature selection more important in semi-supervised learning for text classification than in supervised learning (where a sufficient amount of labeled documents is available)?

ANSWER: *In supervised learning – where sufficient labeled data are available – noisy features are fairly reliably downweighted*[1]*. In semi-supervised learning there is less information contained in the labeled partition of the dataset. These algorithms are usually more susceptible to being led astray.*

6. Give a list of possible parameters that have to be taken into account in semi-supervised learning.

ANSWER: *Amount of unlabeled data, size of the feature set, number of learning iterations, regularization weight for the unlabeled data ...*

7. With regard to the amount of labeled training data, there are two (extreme) situations in which semi-supervised learning does not work. Name them.

ANSWER: *Case 1: There is too little information contained in the labeled dataset. Case 2: All useful information is already contained in the labeled dataset.*

8. Imagine, you are to implement an EM-classifier for spam classification. Your entire dataset comprises 1000 spam and ham mails each. You are to use 1 labeled document per class only. In your first version you only define all words you observe in the labeled documents as your overall vocabulary. You get very bad results. Explain why your current choice of the vocabulary is inappropriate? What would you suggest as an alternative vocabulary?

ANSWER: *The vocabulary will be extremely small, presumably about 100 words. (Remember there are only* 2 *labeled documents!) This choice of words is very document-specific, so only very few words will be significant for the classification task. Hardly anything can really be learnt from the unlabeled documents because the vocabulary will only have a very poor coverage on the entire dataset (including the unlabeled data). Alternative suggestion: use words which frequently occur in the entire dataset and also increase the size of the feature set.*

[1] Note that there are many other NLP tasks where this is not true!

9. One of your eager fellow students has done some experiments in semi-supervised learning on a standard dataset for binary text classification. As a feature set he manually compiled a list of words which he thinks are very discriminative for this classification task. He spent *three weeks* building this resource. With the new lexicon, he achieves a better classification accuracy than a supervised classifier trained on the labeled data (50 documents per class) only. He tells you that he is now convinced that semi-supervised learning is superior to supervised learning. You do not share his enthusiasm. Explain why!

ANSWER: *Building up the lexicon seems to be a very time-expensive matter. Manually labeling more documents (i.e. increasing the size of the labeled dataset) would have been easier (and faster). Moreover, we do not know how he actually constructed the lexicon. Perhaps it is just a feature set which is heavily tuned to the current dataset but does generalize at all. Apart from that, he should do some more comprehensive evaluation: only testing 50 documents on one dataset is insufficient in order to make such a general claim.*