# IN4320 Machine Learning Exercise

March 15, 2017

## Semi-Supervised Learning

When it comes to programming, this assignment can be put into one sentence: take two-class LDA[1] and implement two different ways of doing semi-supervised learning for this classifier. OK, we do want you to do a bit more: the second and most important part of the exercise is concerned with constructing/designing insightful experiments that illustrate the pros and cons of your methods.

When it comes to the implementation of your two semi-supervised approaches for LDA, you are certainly allowed to take any inspiration from other works, papers, web pages, etc., you are even allowed to implement existing methods. In any case, do provide proper references to where you got your inspiration from!

Now, let us make this challenging assignment a bit more concrete. Here are the more specific questions for you to answer and exercises for you to do.

### Prelude

**a** Define and describe your two [really different?] ways of semi-supervised learning for the LDA *on an algorithmic level.* Keep the descriptions for the two methods clearly separate. Before giving these descriptions, do note item **d**. The more different your two choices are, the easier it will be to solve those later exercises.

**b** Take the *Spambase Data Set* from the UCI repository[2] and *first normalize all 57 features*[3]. Based on this normalized data set, make learning curves against the number of *unlabeled* samples for a total of 75 *labeled* samples in the training set *per class.* Check, at least, adding 0, 10, 20, 40, 80, 160, 320, 640, 1280 unlabeled samples per class and see how the expected error rates change. Compare the two curve [and

[1] LDA is linear discriminant analysis: the classifier that assumes the class-conditional distributions to be Gaussian with the same covariance matrix.

[2] See `http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.data`. Note that the *last* column contains the class labels, which are encoded as 0 and 1. The first 57 columns are the features.

[3] That is, make all 57 feature standard deviations equal to 1.

their standard deviations!] to the supervised error rates. Make sure you repeat your experiments sufficiently often to get some nice, and possibly smooth, curves. Do you get significant changes in error rates?

**c** With the same preprocessed data set as in **b**, make the same type of plots, but now plot the log-likelihood[4] [and not the error rate] versus the number of unlabeled data.

## Toccata?

**d** Construct two data sets. On the one data set, your first semi-supervised LDA should work well and improve over the regular supervised learner, but the second should give deteriorated performance on this same set: its performance should be worse than the supervised classifier. On the other data set, it should be the other way around: the second semi-supervised LDA should work better than the supervised learner and the first learner should fail to do so. Consider the setting in which you take 75 labeled samples and a large number of unlabeled samples. Explain why the respective improvements and failures are expected.

**My assessment:** you should be able to keep your report within three pages.

[4] On a test set of course. Also, you probably have to make sure that you test on sets of the same sizes.