# CSCI4155/CSCI6505: Assignment 7

This is an individual assignment that must be submitted by email in PDF format to Paul Hollensen (paulhollensen@gmail.com) by Friday December 4 at noon. Late submissions are not accepted.

1. **Gaussians data:** In assignment 2 and 3 you were working with the data set provided in file `data1.mat`. In this assignment you should test two methods with the same data. The first method is an example of a generative supervised model, that of linear discriminant analysis. The second method is similar but unsupervised, that of a Gaussian mixture model. Please briefly discuss the results in comparison with each other and the results from previous methods.

2. **20newsgroups:** The data in file 20news-bydate-matlab.tgz contain data derived from examples text of 20 news groups. The directory includes data files with extension `.data` for training and testing. These data have the format ”docIdx wordIdx count”. The files with extension label .label contain the class labels as unique number, while the files with extension `.map` provide the mapping between this number and the newsgroup name. Hint: Use sparse matrices with function `spconvert(x)` for the feature matrix.

   Use the Naive Bayes method on the binary feature vector that represents if words are present or absent to predict the test labels. Provide the results in form of a confusion matrix.

3. **Denoising autoencoder and RBM:** (Bonus question) This assignment is about demonstrating the abilities of a denoising autoencoder and an a RBM on the example similar to the example shown in the last slide of the dimensionality reduction slides. To generate the data, produce a long feature vector with feature values following a bell curve with a separate maximum for each class. You should also add noise in any form that you deem interesting. You should then evaluate the unsupervised filters for different noise levels and describe your method and results.