Chalmers/GU, MSA220, Statistical Learning for Big Data
Final report due June 10, 2016

The exam consists of 4 sub-projects. The first 3, I provide some questions for. Use any statistical method you want to answer these questions as best you can. Please write a short report on each project component.
Important: Work independently! You are only penalizing yourself if you chose to work with others, plus this constitutes academic misconduct. You will submit all codes with your final report, and how you code is almost like a finger print...

Question 1

I have posted the data set `CATSnDogs.RData` on the class homepage. This is a small subset of the original data which comprises tens of thousands of images of cats and dogs. This smaller data set consists of 99 cats and 99 dogs (the order they appear in the data set), and images have been re-scaled to 64 by 64 pixels (which is pretty low resolution!).

(A) Classification: Can you tell cats from dogs based on low-resolution images?
Provide reliable estimates for the misclassification error rate on new data - discuss how you estimate this and why you think it's a reliable estimate for future data.
How many features do you need? How do you decide this?
Which classification technique do you prefer and why.
(In image classification there are more complex methods for extracting features than we have discussed - you don't need to use these but work with the methods we've discussed in class).

(B) Dimension reduction and data representation:
Is there a "typical cat" or a "typical dog"? Explain how you deduced this from methods you applied.
What do low-dimensional data representations extract from this data set? Apply and interpret results from SVD, NMF, MDS and SOMS.

(C) Clustering:
Is it easy to learn that there are two types of animals in the image data? Explain
Are there any particular cats or dogs that don't fit with the rest of their species?
Any particular features important for clustering?

Question 2

I have posted a contaminated spam data set on the class homepage. The data set comprises roughly 4000 emails and features extracted from them. The emails have been labeled "spam" and "nonspam". Note, I have added random noise to the data so it cannot be matched to the original data set. In addition, I have mislabeled about 5% of the observations (changing their labels).
Try to identify the mislabeled observations. You can use any methods you want. A mislabeled observation would probably be consistently wrongly predicted across methods and subsamples - can you exploit this?

Question 3

The TCGA data comprising 6 types of cancers with measured gene expression for 20530 genes is posted on the class home page.
Compare one-against-all, pairwise-classification and multi-class classification techniques. Discuss the results - are some cancers easy to tell from other cancers using one method over another? Which cancers are difficult to tell apart and why?
How do the methods compare with respect to selected features? Are the same features selected? Why/Why not? Roughly same number of features selected? Why/Why not?

Question 4

A project of your own choosing - and that is not a repeat of your project 1,2 or 3. You can do a simulation study or a data analysis project.
It should be Big p and/or Big n, where Big p means more than 100 features and Big n means more than 5000 samples since that means that by-hand model selection and/or graphical assessment of model fit is difficult (Big Data).