Exercises, STK4040, week 47

November 16, 2007

Oppgave 1

The files `dataset_47_train.txt` and `dataset_47_test.txt` contain data from recognition of handwritten digits. The first five coloumns contain different features extractet from the digits (0, 3 and 8). The last coloumn is the digit. `dataset_47_train.txt` contains 200 observations of each digit, and is to be used for training. `dataset_47_test.txt` contains about 200 observations of each digit, and is to be used as a test set.

a) Explorative data analysis

Inspect the training data with for instance PCA and different plots. Are there any special properties/structures in the data?

b) LDA

Construct an LDA on the training data. Test it on the test data set, and calculate the error rate and the confusion matrix.

Extra exercise: Plot the test data in the coordinate system determined by the discriminant functions. Compare with a score plot from a PCA of the test data. Are the plots different, and if so: why?

c) QDA

Construct a QDA on the training data and test it on the test data set. Does the QDA perform better or different in any way? Why (not)? Was that (un)expected?