Exercises, STK4040, week 46

November 9, 2007

Exercise 1

Show that the leave-one-out ('full') cross-validated MSE for a model with only the intercept (i.e. $y_i = \beta_0 + \varepsilon_i$) can be written as $\text{Var}(y)n/(n-1)$, where $\text{Var}(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$ is the unbiased sample variance of the response vector.

Exercise 2: Principal component regression

The file `dataset_46.txt` contains a data set with measurements on 79 wheat grain samples. The first coloumn contains a measurement of water content (in percent), and the remaining coloumns (2 – 101) are Near Infrared (NIR) reflectance spectra. The data can be imported an set up in R like this:

```
tmp <- read.table("dataset_46.txt")
water <- tmp[,1]
NIR <- as.matrix(tmp[,2:101])
```

a)

Plot and examine the water measurements and the NIR data. (Tip: The NIR data is probably best viewed by plotting each spectrum (row) as a line.) Are there any special structures, or extreme observations that should perhaps be removed?

b)

Do a principal component regression (PCR) with ten components, regressing water on NIR: Use `prcomp` and calculate regression coefficients, predicted (fitted) values, and residuals.

Plot and judge the residuals, and the predicted against measured response values. Plot the first scores and loadings and try to interpret them. Also look for extreme observations and judge what to do with them.

c)

The plot of predicted against measured response values in b) suggests that 10 components are enough to get a good prediction of water from NIR. But do we really need 10 components, or could we do with fewer?

Use full (leave-one-out) cross-validation to estimate the MSE for models with $0, 1, \ldots, 10$ components.

Plot the estimated MSE values against the number of components, and decide how many components there should be in the model. How large is the estimated MSE for the chosen model size? Which components seem to be the most important for predicting the response?

d) 'Bonus exercise'

Use the PCR implementation in the R package 'pls' (or any ready-made PCR procedure in your favourite software) to replicate the calculations in b) and c). (Tip: If the 'pls' package is not installed already, it can be installed with `install.packages("pls")`. To actually use the package, it has to be loaded with `library(pls)`.)