Series 8

1. We analyse the dataset on babies' survival, that we know from the lecture notes. We apply linear logistic regression and investigate some performance measures.

R-Hints:
Use the R-skeleton available on the course's webpage.

a) Fit a linear logistic model to the data. Plot the resulting ROC. Further, plot the misclassification rate as a function of the threshold on the probability. You can use the R -package `ROCR` for this.

b) In order to estimate the generalization error, we perform 10-fold cross-validation. Again, plot the ROC and the misclassification rate. For this, average the curves over all ten folds. What do you observe? Is there an indication that logistic regression is overfitting to the data?

c) We now look at a set of different cost functions than just the misclassification rate. Each of these is given by

$$\frac{c_1 * FP + c_2 * FN}{n},$$

where n denotes the total number of data points, FP the number of false positives and FN the number of false negatives. This allows for weighting the cost of a false positive and a false negative differently. Plot the loss for a range of different values of c_1 and c_2. To make things comparable, fix $c_1 + c_2 = 2$. What do you observe? How can you explain this?

d) In the ROC, we plot the TPR as a function of the FPR, say, $TPR = f(FPR)$. For a given cost function as above, find a formula for $f'(FPR^*)$, where FPR^* is the point minimizing the cost. Express your result in terms of c_1, c_2 and p, where p denotes the overall fraction of positives. Interpret the formula in the different limits, i.e., $c_1 \to 0$, $c_2 \to 0$, $p \to 1$ and $p \to 0$.
Hint: Assume that $f(FPR)$ is a strictly concave function, i.e., we do better than random guessing, and that $f''(FPR)$ exists.

e) Assume that in reality falsely claiming a baby survives is much more severe than falsely claiming it will die. How would you (qualitatively) chose c_1 compared to c_2 in this situation?

2. In this exercise we are investigating the ozone dataset which you have already seen in the lecture. The dataset `ozone` is available in numerous R-packages, e.g. in the package `gss`. You can load it with `data(ozone, package = "gss")`. If you do not have access to the package, you can get the data from `http://stat.ethz.ch/Teaching/Datasets/ozone.dat`. A short description of the variables is available at `help(ozone, package = "gss")`.

R-Hints:
Use the R-skeleton available on the course's webpage.

a) Get an overview of the data with `pairs()`. You should take the log of the response (`upo3`) and remove the outlier in the predictor `wdsp`. Explain why you should do this.
R-Hints:
The transformation can be done using:

```
ozone$logupo3 <- log(ozone$upo3)
d.ozone <- subset(ozone, select = -upo3)
```

b) We want to compare linear regression with an additive model (7.2 in the lecture notes).
In order to better fit the model, we allow polynomial fits of the data in the predictive variables up to degree d for linear regression, i.e.
$y_i = \beta_0 + \beta_{1,1}x_{i1} + \beta_{1,2}x_{i1}^2 + ...\beta_{1,d}x_{i1}^d + ... + \beta_{p,1}x_{ip} + \beta_{p,2}x_{ip}^2 + ...\beta_{p,d}x_{ip}^d + \epsilon_i$.
This means instead of having $p + 1$ predictive variables we have $pd + 1$.
Here we fit the 5 models with linear regression for the degrees up to 5. We also fit the data with an additive model as programmed in R.
R-Hints:
Use the function `poly()` in the formula when working with `lm()`. The function `gam()` (to fit an

additive model) can be found in the package `mgcv`. Usage: `gam(formula, data)`. `formula` must be of the form
`logupo3 ~ s(vdht) + s(wdsp) + ...`
The function `wrapFormula()` from package `sfsmisc` can help to quickly create the desired formulas. Make use of `summary()` to get an overview of your `gam()` output.

c) Plot the respective fits. Use function `termplot()` for the linear models. What do you observe? Which polynomial degree would you choose? Use `mult.fig()` and the default `plot()` function for the additive model. Compare the additive model plot to the linear model ones. What do you observe?

d) Now we want to perform model selection on our models. We fix σ as the estimated scale from the linear fit of degree 5. Then we calculate Mallows' C_p statistic for all 6 models. Which of the linear models do we prefer? And which of all the models?

e) Instead of GAM, we can also use MARS as implemented in R-package `earth`. Fit a MARS model with degree 1, i.e., no interaction terms to the data. According to the fitted model, how does the expected log ozone concentration change if the temperature on the Sandburg Air Base (variable `sbtp`) increases from 41 to 60 degrees celsius keeping the other variables fixed?

f) Finally, we want to compare the performance of MARS models with degree 1 to 3 with GAM. We use a 10-fold cross validation and take the squared error loss as performance metric. Which model would you prefer based on this analysis?

Preliminary discussion: Friday, May 06.

Deadline: Friday, May 13.