Question 1.6 from Lecture notes on ridge regression

Download the `multtest` R package from BioConductor.

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("multtest")
```

Activate the library and load leukemia data from the package:

```
library(multtest)
data(golub)
```

The objects `golub` and `golub.cl` are now available. The matrix-object golub contains the expression profiles of 38 leukemia patients. Each profile comprises expression levels of 3051 genes. The numeric-object `golub.cl` is an indicator variable for the leukemia type (AML or ALL) of the patient.

(a) Relate the leukemia subtype and the gene expression levels by a logistic regression model. Fit this model by means of penalized maximum likelihood, employing the ridge penalty with penalty parameter $\lambda = 1$. This is implemented in the `penalized` package available from CRAN. Note: center (gene-wise) the expression levels around zero.

(b) Obtain the fits from the regression model. The fit is almost perfect. Could this be due to overfitting the data? Alternatively, could it be that the biological information in the gene expression levels indeed determines the leukemia subtype almost perfectly?

```
##     1     2 3 4 5 6     7     8 9    10 11    12 13 14 15 16    17    18
## y   0 0.000 0 0 0 0 0.000 0.000 0 0.000  0 0.000  0  0  0  0 0.000 0.000
## f_x 0 0.001 0 0 0 0 0.001 0.001 0 0.001  0 0.002  0  0  0  0 0.001 0.001
##     19 20 21    22 23 24    25 26 27    28    29    30    31    32 33
## y    0  0  0 0.000  0  0 0.000  0  0 1.000 1.000 1.000 1.000 1.000  1
## f_x  0  0  0 0.001  0  0 0.001  0  0 0.998 0.999 0.999 0.998 0.998  1
##        34    35    36 37    38
## y   1.000 1.000 1.000  1 1.000
## f_x 0.999 0.998 0.999  1 0.999
```

(c) To discern between the two explanations for the almost perfect fit, randomly shuffle the subtypes. Refit the logistic regression model and obtain the fits. On the basis of this and the previous fit, which explanation is more plausible?

```
##         1     2     3     4     5     6     7     8     9    10    11
## y.p 0.000 1.000 0.000 1.000 1.000 0.000 0.000 1.000 1.000 0.000 0.000
## f_x 0.002 0.996 0.002 0.994 0.994 0.003 0.002 0.995 0.995 0.002 0.003
##        12    13    14    15    16    17    18    19    20    21    22
## y.p 0.000 0.000 1.000 0.000 0.000 0.000 0.000 0.000 0.000 1.000 0.000
## f_x 0.002 0.001 0.997 0.004 0.002 0.001 0.001 0.001 0.002 0.998 0.001
##        23    24    25    26    27    28    29    30    31    32    33
## y.p 1.000 0.000 0.000 1.000 0.000 0.000 0.000 1.000 0.000 0.000 0.000
## f_x 0.995 0.003 0.002 0.993 0.002 0.001 0.002 0.997 0.002 0.001 0.001
##        34    35    36    37    38
## y.p 0.000 0.000 0.000 1.000 0.000
## f_x 0.001 0.001 0.002 0.995 0.001
```

(d) Compare the fit of the logistic model with different penalty parameters, say $\lambda = 1$ and $\lambda = 1000$. How does λ influence the possibility of overfitting the data?

```
##         1     2    3     4    5     6     7     8     9    10    11    12
## y   0.000 0.000 0.00 0.000 0.00 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## f_x 0.086 0.165 0.08 0.088 0.06 0.102 0.146 0.154 0.077 0.126 0.091 0.234
##        13    14    15    16    17    18    19   20   21    22    23    24
## y   0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.00 0.00 0.000 0.000 0.000
## f_x 0.038 0.085 0.035 0.053 0.115 0.127 0.093 0.03 0.04 0.175 0.093 0.056
##        25   26    27    28    29    30    31    32    33    34    35    36
## y   0.000 0.00 0.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
## f_x 0.193 0.09 0.118 0.699 0.732 0.801 0.677 0.666 0.857 0.727 0.664 0.816
##        37    38
## y   1.000 1.000
## f_x 0.858 0.751
```

(e) Describe what you would do to prevent overfitting.