Classification Methods

Exercise 1 (Conceptional: Training and Test Error)

Exercise 8 (p. 170): Compare logistic regression and KNN based on error rates.

Exercise 2 (Conceptional: Odds)

Exercise 9 (p. 170): Interpretation using odds.

Exercise 3 (Applied: Comparison of Classification Methods I)

Exercise 11 (p. 171): Perform a comparison of classification methods using the `Auto` data set.

Exercise 4 (Applied: Comparison of Classification Methods II)

Exercise 13 (p. 173): Perform a comparison of classification methods using the `Boston` data set.

Exercise 5 (To be graded in detail: Derive the formula of the discriminant function $\delta_k(x)$)

It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

Exercise 6 (To be graded in detail: Logistic Regression)

We want to analyze data on infections of mothers after abdominal delivery. The data is given in an aggregated fashion by the contingency table:

	Not planned			Planned	
	Infection			Infection	
	no	yes		no	yes
no antibiotics					
no risk	9	0	no risk	32	8
risk	3	23	risk	30	28
	no	yes		no	yes
antibiotics					
no risk	0	0	no risk	2	0
risk	87	11	risk	17	1

(a) Let be π the probability that a woman develops an infection after abdominal delivery. Specify the corresponding likelihood for π depending on the data whether antibiotics are given as prophylactic measure using the data in the table.

(b) We model the probability for infection using a logistic regression model

$$\log \frac{\Pr(\texttt{infection} = 1)}{1 - \Pr(\texttt{infection} = 0)} = \beta_0 + \beta_1 \cdot \texttt{anitbiotics} + \beta_2 \cdot \texttt{risk} + \beta_3 \cdot \texttt{plan},$$

where

$$\begin{aligned}
\texttt{infection} &= \begin{cases} 1 & \text{infection} \\ 0 & \text{no infection} \end{cases} \\
\texttt{anitbiotics} &= \begin{cases} 1 & \text{dosed with antibiotics as prophylactic measure} \\ 0 & \text{no antibiotics.} \end{cases} \\
\texttt{risk} &= \begin{cases} 1 & \text{a risk factor exists} \\ 0 & \text{otherwise.} \end{cases} \\
\texttt{plan} &= \begin{cases} 1 & \text{abdominal delivery planned} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Do the interpretation of the following output using odds ratios:

```
Call:
glm(formula = infection ~ antibiotics + plan + risk, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7149  -0.5298  -0.4933   0.7227   2.5141

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8207     0.4947  -1.659   0.0971 .
antibiotics  -3.2544     0.4813  -6.761 1.37e-11 ***
plan         -1.0720     0.4254  -2.520   0.0117 *
risk          2.0299     0.4553   4.459 8.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
%
    Null deviance: 299.01  on 250  degrees of freedom
Residual deviance: 226.52  on 247  degrees of freedom
AIC: 234.52

Number of Fisher Scoring iterations: 5
```

(c) Assume a woman with a risk factor, planned abdominal delivery and antibiotics as prophylactic measure. What is her probability of an infection?

(d) We also do a linear discriminant analysis and obtain the output below. Compute the probability of an infection for the women described in (c).

```
Call:
lda(infection ~ antibiotics + plan + risk, family = binomial)

Prior probabilities of groups:
        0         1
0.7171315 0.2828685

Group means:
  antibiotics      plan      risk
0   0.5888889 0.4500000 0.7611111
1   0.1690141 0.5211268 0.8873239

Coefficients of linear discriminants:
```

```
                  LD1
antibiotics -2.796391
plan        -0.733131
risk         1.968118
```

This homework is due at the beginning of the discussion section on Feb 23, 2016 at 3.30pm.

All references refer to the textbook: James, Witten, Hastie and Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R.* Springer. Available online: `http://www-bcf.usc.edu/~gareth/ISL/`.
