Line H. Clemmensen
Section of Statistics and Data Analysis
DTU Compute

Exercises 02582
Module 4
Spring 2022

February 23, 2022

Topics: Logistic regression, regularized logistic regression, Regularised discriminant analysis (RDA)

Exercises:

1 Logistic regression: Given a logistic model for lung cancer (yes/no) as a function of smoking (number of cigarettes per day) with $\beta = 0.02$. Show that one unit increase in smoking means an increase in lung cancer risk (odds-ratio) of $\exp(0.02) = 1.02 = 2\%$.

2 We have a data material (Golub et al 1999) with gene expression levels from 72 patients with two forms of leukemia, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Gene expression levels (how actively the cells are using the information in different genes) are measured for 7127 genes. We would like to build a biomarker for classification of the two cancer forms. Ideally, we would like to use only a few variables.

   (a) How can you use logistic regression here?

   (b) Build a classifier for training data in `GolubGXtrain.csv`. What regularization method do you prefer if you want to have few genes in the biomarker?

   (c) How many variables do you end up with?

   (d) Use the obtained model to calculate accuracy on the test data.

3 Implement and calculate a Regularized Discriminant Analysis (RDA) for the Silhouette data in `Silhouettes.mat` (You may use the file `plot_silhouettes.m` to visualize data).

   (a) What happens when we vary γ in RDA?

Line H. Clemmensen
Section of Statistics and Data Analysis
DTU Compute

Exercises 02582
Module 4
Spring 2022

February 23, 2022

Topics: Logistic regression, regularized logistic regression, Regularised discriminant analysis (RDA)

Resources for this exercise:

Listing 1: Resources in Matlab

```
csvread % for reading files
lassoglm % calculates regularized logistic regression
glmval % for making predictions
```

Listing 2: Resources in R

```
library(glmnet)  # perform logistic regression
read.csv('Name␣of␣file') # read .csv file
```

Listing 3: Resources in Python

```
from sklearn.linear_model import LogisticRegression #
   loading logistic regression
model = LogisticRegression(penalty = 'l1', C = Cval, tol = 1
   e-6) #
lasso penalty for logistic regression
import pandas as pd
pd.read_csv # read csv file
```

End of exercise