TMA4315 Generalized Linear Models

Assignment 3:

GLMs for binomial data

Deadline: Monday, November 21, 2016 (week 47)

Reports should be submitted to Jacob Skauvold, either in the mailbox (7th floor, Central Building II), or by e-mail to `skauvold@math.ntnu.no`.

Guidance will be by appointment. If you need help with this project, contact the teaching assistant and set up a meeting time.

In the first assignment, you created the myglm package and provided basic functionality for fitting linear models with a Gaussian response. In the second assignment, you extended this package to make it applicable to Poisson count data. In this third assignment, you will extend the package even further, so it can be used for logistic regression with a binomial response.

Problem 1: Logistic Regression

Wikipedia's *List of highest mountains*[1] lists 118 of the world's highest mountains, along with some properties of each one, including the number of suc-

[1] `https://en.wikipedia.org/wiki/List_of_highest_mountains`

cessful and failed attampts at reaching the summit as of 2004. In this problem, we will consider a data set consisting of the height (in meters), topographic prominence (also in meters), number of successful ascents and number of failed attempts for 114 of the mountains on the list. The following four mountains are excluded from the dataset because of incomplete data:

- Mount Everest (ranked 1st in height)
- Muztagh Ata (ranked 43rd)
- Ismoil Somoni Peak (ranked 50th)
- Jengish Chokusu/Tömür/Pk Pobeda (ranked 60th)

The data set is available at `http://www.math.ntnu.no/emner/TMA4315/2016h/Assignment3/wikimountains.txt`.

In the following, let y_i be the number of successful ascents, and let n_i be the total number of attempts (successful and failed) of the ith mountain. Use a binomial model for the number of successful ascents, and a linear model for the logit transform of the probability of success,

$$Y_i \sim \mathrm{Bin}(n_i, p_i) \quad \text{for } i = 1, \ldots, 114$$

and

$$\mathrm{logit}(p_i) = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}$ is the model matrix and $\boldsymbol{\beta}$ is the vector of model parameters.

a) Write down the log-likelihood function $\ell(\boldsymbol{\beta})$ for this model. Extend your `myglm` package so it can fit this type of model.

b) Fit the model, with height and prominence as predictors, to the data using your `myglm` package.

Hint: To fit this model using the built-in `glm` function, you could write `model = glm(cbind(ascents, failed) ∼ height + prominence, data = wikimountains, family = "binomial")`.

c) Make plots of the estimated probability of successfully ascending a mountain as a function of the predictors. Interpret the model parameters and discuss their significance. Assess goodness of fit.

d) The height and prominence of Mount Everest are both equal to 8848 meters. Based on the fitted model, predict the probability of successfully ascending Mount Everest. Give both a point estimate and an interval estimate. Is the prediction reasonable? Is the degree of uncertainty reasonable? Discuss potential problems with the prediction.

Problem 2: Logistic regression and Cross validation

In credit scoring the goal is to determine if a customer should be given credit or not. If credit is given to someone who can not service it, the company loses money. On the other hand, if credit is withheld unnecessarily, less money is earned. The file `credit.txt` contains 30 000 observations of 23 predictors labeled `X1` through `X23` and an indicator `Y` for whether the customer is good (0) or bad (1). The goal of this exercise is to make a model for predicting whether a customer should be given credit (is good) or should not be given credit (is bad). A description of the data set is available at `http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`, and the data set itself is available at `http://www.math.ntnu.no/emner/TMA4315/2016h/Assignment3/credit.txt`.

This problem is focused on prediction, and the goodness of a model is determined by its predictive power. You should consider the misclassification of a bad customer as good to be five times as expensive as the misclassification of a good customer as bad.

You should evaluate predictive power based on 10-fold cross validation. As cross validation is not covered extensively in lectures, you are encouraged to find relevant material online.

In addition to the code for fitting the binomial GLM, which is needed in problems 1 and 2 both, a full solution to problem 2 should contain the following.

- Estimation of the classification error by cross validation.
- A description of how you selected the final model (e.g. comparisons of deviances and classification errors).
- A description of the chosen model, including interpretation of the parameters.
- A discussion of the quality and performance of the chosen model.