STAT 788 - Midterm

Daniel Hartig

May 9, 2018

How to run

The submitted code consists of six different files. The file `main.c` is one of the main executables which calls to the other files. This main has customizable command line arguments. Another main is contained in `longtest.c` which generates test data discussed later in this paper. Reading data from the file, standardizing a dataset, and generating mislabellings are done in `datagen.c` with its header file `datagen.h`. Logistic regression and associated mathematical functions are done in `logistic.c` and associated header file `logistic.h`

The customizable executable is compiled by

```
gcc -o final main.c datagen.c logistic.c -lm
```

and run by

```
./final pima 0.
```

The arguments to this script are

Variable	Description
Data Source	Source of the response data: 'pima': unedited Pima dataset response values 'b0': authors' generated response variables from Hung et al., Section 4.2
Mislabel Type	Strategy for generating mislabels from Hung et al., Section 4.1: 0: No mislabeling 1: Strategy S1 (y-dependent mislabeling) 2: Strategy S2 (x-dependent mislabeling) 3: Strategy S3 (random x, y-dependent mislabeling) 4: Strategy S4 (clustered x, y-dependent mislabeling)

The generation of test data compiled by

```
gcc -o testgen midterm.c datagen.c logistic.c -lm.
```

and run by

```
./testgen.
```

Methodology

Generating samples from the Pima set

The Pima data set is read from its input data file (`pima.dat`) and divided into two parts, the feature and response variable. The features have a column of ones added as the first column to give an intercept. This produces a different ordering of feature numbers compared to the tables in Hung. For example, in Table 1 of Hung, the nine features start with β_{01} = x1 as labeled in the data file; and end with β_{09} = the intercept. In the computational parts of this project, the intercept is pre-pended. For any length 9 vector of dimension of a matrix, the 0th item is the intercept, while the 1st item corresponds to x1 from the data file.

Once the data is read, a sample from the data set is taken. There are 768 data points in the data set, but all samples taken are hardcoded to size 500 following Hung. A random sample isolumn taken using a simplified Reservoir algorithm. Since the total dataset is small, it is used in its entirety as the reservoir.

Algorithm 1 Generate a size SAMPLESIZE sample from DATA

```
output ← empty length(SAMPLESIZE) array
selections ← empty length(DATA) array
numleft ← length(DATA)
for i ∈ (0, 1, ... SAMPLESIZE) do
  n ← RANDOMUNIFORM()
  output[i] ← DATA[selections[n]]
  numleft ← numleft − 1
  selections[n] ← selections[numleft]
end for
```

After the data set is generated, it is standardized. Standardization is done on a columnwise basis to generate a standardized matrix A^* from feature matrix A:

$$\text{Mean for column}j: \quad \mu_j = \frac{1}{n}\sum_{i=1}^{n} A_{i,j}$$

$$\text{Std Dev for column}j: \quad \sigma_j = \sqrt{\frac{\sum_{i=1}^{n}(A_{i,j}-\mu_j)}{n-1}}$$

$$\text{Standarized Matrix } A^*: \quad A^*_{i,j} = \frac{A_{i,j}-\mu_j}{\sigma_j}$$

Generating the Author's response variables

There are two options for response variables. The first is to use the original Pima dataset as is, which represents the incidence of diabetes in females members of the Pima indian tribe, based on eight features describing the individuals. This options is selected by passing `'pima'` as the first command line argument to the `'main.c'` version of the executable. Since the response variables are already read from the data file and selected along with their matching features into a sample of 500, nothing more is done in this case.

The second is to use a 'true' regression parameter set as done by Hung. This option is selected by passing `'b0'` as the first command line argument to the `'main.c'` executable.

Generating mislabeled samples

There are five types of mislabeling schemes. The first scheme is to not mislabel anything, simply passing the dataset through as is.

Logistic Regression

Logistic regression assumes a logit link between the explanatory variables and the response variable, and a binomial distribution of errors. The logit link function is

$$\begin{aligned}\log\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \beta_0 + \beta_1 x_1 + ... \\ &= \beta x\end{aligned}$$

For each observation y_i let x_i be the explanatory variables associated with that observation and $\boldsymbol{\beta}$ be the parameters of the regression function we wish to estimate. Then

$$\pi_i(x|\beta) = \frac{1}{1+e^{-\beta x_i}}.$$

A joint binomial probability mass function is

$$Pr(Y|X;\beta) = \prod_i \pi_i^{y_i}\left(1-\pi_i\right)^{y_i}$$

with negative log likelihood function

$$\mathcal{L}\left(\beta|x,y\right) = -\sum_i y_i \log(\pi_i) + (1-y_i)\log\left(1-\pi_i\right).$$

The derivative of $\pi = 1/(1+\exp(-\beta x))$ can be expressed as

$$\begin{aligned}\frac{d\pi}{d\beta} &= \frac{d}{d\beta}\left(\frac{1}{1+e^{-\beta x}}\right) \\ &= \frac{xe^{-\beta x}}{\left(1+e^{-\beta x}\right)^2} \\ &= x\pi(1-\pi)\end{aligned}$$

Thus the derivative of the log likelihood function is

$$\begin{aligned}\frac{d\mathcal{L}\left(\beta|x,y\right)}{d\beta} &= -\sum_i \left[y_i \frac{1}{\pi_i}\frac{d\pi_i}{d\beta} + (1-y_i)\frac{1}{1-\pi_i}\frac{-d\pi_i}{d\beta}\right] \\ &= -\sum_i y_i x_i\left(1-\pi_i\right) - \sum_i \left(y_i - 1\right)x_i\pi_i \\ &= -\sum_i x_i\left(y_i - \pi_i\right).\end{aligned}$$

The second derivative is

$$\frac{d^2\mathcal{L}\left(\beta|x,y\right)}{d\beta^2} = \sum_i x_i^2 \pi_i\left(1-\pi_i\right)$$

which is no-negative. Thus, the objective function is convex, and a global minimum for the objective function can be obtained by an iterative optimization methodology, such as gradient descent Newton's method.

Solution using Gradient Descent

Solution using Newton-Raphson method

Simulation Studies

Logistic Regression of Mislabel Possibilities

Table 1 contains the results of

Parameter	Mislabel Type	Mean	95% Confidence Minimum	95% Confidence Maximum
Intercept	No Mislabel	-0.877	-0.746	-1.008
	S1	-0.805	-0.667	-0.943
	S2	-0.664	-0.543	-0.786
	S3	-0.645	-0.504	-0.785
	S4	-0.609	-0.440	-0.777
x1	No Mislabel	0.428	0.288	0.567
	S1	0.309	0.154	0.465
	S2	0.297	0.135	0.460
	S3	0.305	0.130	0.480
	S4	0.307	0.158	0.457
x2	No Mislabel	1.158	0.985	1.330
	S1	0.857	0.702	1.011
	S2	0.866	0.686	1.045
	S3	0.853	0.652	1.055
	S4	0.822	0.652	0.992
x3	No Mislabel	-0.258	-0.132	-0.384
	S1	-0.195	-0.044	-0.345
	S2	-0.183	-0.051	-0.315
	S3	-0.193	-0.064	-0.321
	S4	-0.166	0.002	-0.166
x4	No Mislabel	0.013	-0.131	0.158
	S1	0.003	-0.152	0.158
	S2	0.006	-0.137	0.148
	S3	0.021	-0.140	0.181
	S4	-0.002	-0.152	0.148
x5	No Mislabel	-0.151	-0.012	-0.289
	S1	-0.111	0.040	-0.264
	S2	-0.102	0.078	-0.281
	S3	-0.119	0.057	-0.295
	S4	-0.091	0.053	-0.236
x6	No Mislabel	0.706	0.550	0.862
	S1	0.552	0.359	0.746
	S2	0.512	0.352	0.672
	S3	0.498	0.311	0.685
	S4	0.470	0.288	0.652
x7	No Mislabel	0.309	0.165	0.454
	S1	0.239	0.085	0.393
	S2	0.243	0.084	0.400
	S3	0.236	0.093	0.380
	S4	0.225	0.068	0.382
x8	No Mislabel	0.162	0.015	0.310
	S1	0.139	-0.010	0.288
	S2	0.134	-0.042	0.309
	S3	0.120	-0.035	0.275
	S4	0.128	-0.058	0.313

Table 1: Distributions of Standardized Logistic Regression Parameters for 100 samples with sample size $n = 500$