Problem Set 5

D Keck

April 09, 2018

Objectives

- Master the logistic regression method
- Predict 10 year coronary heart disease using a logistic regression library solver
- Also predict with your own gradient descent solver
- Re-compute and verify all statistical and information results from the library solver
- Use a receiver operator characteristic curve to set a prediction threshold
- Implement another method using a libary function to obtain similar or better accuracy and precision (true positive rate)
- Compare linear regression with the method used by the scientists associated with the Framingham project

Predict 10 Year Coronary Heart Disease with the Framingham Heart Study Data

Due April 23

Requirements

- Problem sets are done in groups of 3 (21 groups)
- Each group member must be skilled with the language chosen at the start of problem set
- Each student must do approximately $\frac{1}{3}$ of the development
- The deliverable is single PDF or HTML that meets the standard of a reproducible research report with professional quality communication. Deployable code will not be a deliverable for any problem set.
- All code other than print statements must be displayed
- Identify and explain each answer. Don’t just print a number.
- Numeric output should be easy to read e.g., not 10 decimal places
- Graphics must be easy to read i.e., titles, axis labels, legends, colors, etc.
- The rows and columns of tables and the columns of matrices and data frames must be labeled

Grading

- Each question is worth 2 points of the 20 total points
- 1.5 points are awarded for a correct answer that I can quickly find. I will not spend much time searching for answers
- .5 points will be awarded for the quality, clarity, and insight of the answer

- A single HTML or pdf file is required and must be generated via a notebook or markdown package. I will only grade a single, orderly file from each group.

Reference and Data

Reference (https://www.framinghamheartstudy.org)

Predictions (https://www.framinghamheartstudy.org/fhs-risk-functions/coronary-heart-disease-10-year-risk/)

Technical Paper (http://circ.ahajournals.org/content/97/18/1837)

A simplified Framingham data file is available on canvas: Files/framingham.csv

Note

I will summarize and include the formulas for home grown calculations by April 11. This should not prevent you from proceeding, nor is it actually necessary

Question 1

- Read in the Framingham data set from Canvas
- How many observations? How many features (explanatory variables) ?
- Explore the data (plots, tables, statistics) including feature types.
- Summarize any NA's by feature

Question 2

- Resolve any NA's. Describe your method. I used mice()
- Do you suggest scaling data or not? If so, normal or uniform scaling? If scaling is advisable, then do so. (An iterative solver will be used below.)
- Randomly split data with 80% in the training set and 20% in the test set.
- Provide a baseline accuracy by reviewing the TenYearCHD binary variable
- What is the postive predictive rate (precision) of your baseline prediction ?
- Which error type seems worse and thus should be considered along with accuracy in your predictions below? Why ?

Question 3

- Create a 'null' logistic regression model with glm() using only the intercept feature. Use glm's binomial family with the logit link.
- Create a 'full' logistic regression model with glm() using *all* of the features including the intercept
- (Suggest lableling the model objects as modelNULL and modelFULL for clarity)
- Print summary() for both models
- Print the log likelihood, the deviance, and the AIC using logLik(model), model$deviance, and AIC(model)

for both models
- Compare the models and comment

Question 4

- Create the lowest AIC ('best') model using backward elimination of parameters. (Suggest labeling the model as modelBEST for clarity and use of step(, trace =0)
- Print the summary(), the log likelihood, the deviance, and the AIC for this model and compare with the previous two models.
- Print these 'best' parameters (coefficients) along with their 95% confidence intervals in a single table

Question 5

- Using the 'best' model, "predict" the 10YrCHD on the training set.
- Display the confusion matrix, the accuracy, as well as the true positive and true negative rates. Use a probability threshold (cutoff) of .5 to make the binary decisions from the probabilities.
- Compare to the baseline model from Question 2
- Now predict the 10YrCHD on the test set
- Display the confusion matrix, the accuracy, as well as the true positive and true negative rates. Again, use a probability threshold (cutoff) of .5 again.

Question 6

Now use ROC curves to reconsider a threshold (cutoff) of .5 using the ROCR (or other) package

- Create the prediction object for the training set
- Plot the accuracy v. the threshold (cutoff)
- Plot the true positive rate v. the false positive rate and label the threshold values
- Plot sensitivity v. specificity and label the threshold values
- Plot precision v. recall and label the threshold values
- Comment on your findings for each plot
- Compute AUC for the three models (null, full, and best) and comment

Question 7

Using gradient descent, solve for the parameters from the 'best backward eliminaton' data set and verify relative to the glm() parameters

- Create the $\mathbf{X}$ and $\mathbf{y}$ test and training sets
- Compute the parameters (coefficients) $\boldsymbol{\theta}$
- Compare to glm()'s coefficients in a single table

Question 8

Using your gradient descent model and parameters, compute and verify

- standard errors,
- Z statistics, and
- P values in a single labeled table along with the coefficients
- Note: The standard errors are the diagonals of $\mathbf{C}$
- $\mathbf{C} = \left(\mathbf{X}^T \cdot \mathbf{V} \cdot \mathbf{X}\right)^{-1}$
- $V_{jj} = p_j \cdot \left(1 - p_j\right)$ and $V_{ij} = 0 \;\; i \neq j, \quad 1 \leq i, j \leq p$
- $se\left(\widehat{\theta_j}\right) = C_{jj}$
- $\mathbf{p}$ are the $\widehat{\boldsymbol{\theta}}$ probabilties before thresholding but after the sigmoid transformation
- Compute and verify the log likelihood, deviance, and AIC with those from glm()
- You may exclude verification of the deviance residuals, as well as ignore dispersion and Fisher scoring iterations.

Question 9

- Chose another predictive method that you think might produce similar or better results on the Framingham study.
- Implement using library functions on the training set
- Create confusion tables for predicting on the test set. You may use any solution parameters or tuning options.
- Display the accuracy and the true positive and negative error rates
- Comment on why this method did or did not result in better accuracy or a better true positive rate (precision.) No need to keep trying other methods if you didn't pick a winner initially, just explain why the method was weaker on this data set.

Question 10

- Read the technical paper linked in the References section
- How is that method different from logistic regression? One paragraph is sufficient, full understanding is *not* required, just the essential idea.

Technical Paper (http://circ.ahajournals.org/content/97/18/1837)