Gmail **Andrew Ho <kironide@gmail.com>**

Week 2 Day 2 (morning)

2 messages

Jonah Sinick <jsinick@gmail.com> Tue, Feb 23, 2016 at 9:48 AM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

Today we'll be looking at Raw data from online personality tests, and predicting gender in terms of answers to personality inventory items.

- If you need a refresher, go through ISLR Chapter 4, sections 1-3, as well as the the sections of the lab 4.6.1 & 4.6.2 again. We'll be talking more about the theory in class
- Start with the ***Rosenberg Self-Esteem Scale*** http://personality-testing.info/_rawdata/RSE.zip
- All of the datasets have the same form. When you download one,

  (1) Filter for gender %in% 1:2 (males correspond to 1's and females correspond to 2's). Subtract off 1 from the gender column, so that 0 corresponds to male and 1 corresponds to female.

  (2) Missing values for personality inventory items correspond to 0's. Convert these to NAs and then remove the rows containing them.
- Scale the personality inventory items. Aggregate by gender, then compute gender differences for response to each question in standard deviations.
- Fit a logistic regression model for gender using glm() with argument family = "binomial." How do the signs of the coefficients compare with the signs of the gender differences in average values of answers?
- For more intuition on what's going on above do *stepwise regression* via the "step" function in R http://www.inside-r.org/r-doc/stats/step. Start with the model

  glm(gender ~ 1,df, family = "binomial")

  with scope given by

  formula(glm(gender ~ 1,df, family = "binomial"))

  Write a loop which in each iteration, replaces the current model with the result of using stepwise regression with 1 step, storing the summary of each resulting model in a list. Compare the coefficients of these models.
- Use CVbinary from the DAAG package to compute the accuracy of the model, as well as the cross-validated accuracy.

  Download the pROC package, and compute the ROC curve for actual vs. predicted as outputted by the model. Graph it.
- Do the same for other personality inventories. If the sample size and number of features are large, glm() can be very slow, because it's not well optimized. In such instances, using cv.glmnet() will speed things up.

cv.glmnet will also handle instances in which there are too many features relative to the number of examples for usual logistic regression to not overfit.

Jonah Sinick <jsinick@gmail.com> Tue, Feb 23, 2016 at 5:06 PM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

Questions to investigate if you'd like:

- Take a look at other datasets at http://personality-testing.info/_rawdata/ of your choosing, fitting logistic regression for gender in terms of the survey questions (after cleaning the data appropriately).

  Pay attention to the meanings of the questions, and the possible answer choices.

  As above, if the number of features is large relative to the number of examples, you may prefer to use glmnet. I suggest using the L^1 penalty (alpha = 1) to reduce the number of features that you need to keep in mind and more easily interpret what's going on.

- Aggregate the speed dating dataset by person. Use logistic regression to predict gender in terms of activities.

  If you want to predict career_c or race, restrict to a subset of the data with only two categories (e.g. academia vs. finance/business, or Caucasian vs. Asian).

- If you'd like, you can look at a cleaned version of LW survey dataset (link to code) and predict gender, sexual orientation, involvement in the community, or another variable of your choice. Here too, restrict to the subset of examples falling into two categories so that you can use logistic regression.

[Quoted text hidden]