Homework 7: Red Brain, Blue Brain

36-402, Section A, Spring 2019

Due at 6:00 pm on Wednesday, 27 March 2017

AGENDA: More practice with classification models, including yet more logistic regression; density estimation; conditional densities.

TIMING: Problems 1–3 and 6 involve fitting models to data, plotting, and interpretation, but no coding. Problem 5 requires doing all that and some bootstrapping, for which you will need to write a little code (along lines you have done before). Problem 7 requires fitting a model and making some plots from it, and you will (probably) need to write a little code, along the lines of examples in the book, to do so. Problem 8 requires comparing models, and you will need to either write some new code, or tweak some example code, to do 8b. The solutions to all problems take about 5 minutes to knit without a cache (and about two seconds with a cache — cache everything!).

The data set n90_pol.csv contains information on 90 British university students who participated in a psychological experiment designed to look for relationships between the size of different regions of the brain and political views. The variables `amygdala` and `acc` indicate the volume of two particular brain regions known to be involved in emotions and decision-making, the amygdala and the anterior cingulate cortex; more exactly, these are residuals from the predicted volume, after adjusting for height, sex, and similar anatomical variables. The variable `orientation` gives the subjects' locations on a five-point scale from 1 (very conservative) to 5 (very liberal). `orientation` is an ordinal but not a metric variable, so scores of 1 and 2 are not necessarily as far apart as scores of 2 and 3.

1. *Marginal density of brain region volumes*

   (a) (5) Using `npudens`, estimate the probability density for the volume of the amygdala. Plot it and report the bandwidth.

   (b) (5) Repeat this for the volume of the ACC.

2. *Joint density of brain regions*

   (a) (5) Using `npudens`, estimate a joint probability density for the volumes of the amygdala and the ACC. What are the bandwidths? Are they the same as the bandwidths you got in problem 1? Should they be?

(b) (5) Plot the joint density. Does it suggest the two volumes are statistically independent? Should they be? You may use three dimensions, color, contours, etc., for your plot, but you will be graded, in part, on how easy to read it is.

Hint: Remember that the random variables X and Y are statistically independent when their joint pdf is the product of their marginal pdfs, $p(x, y) = p(x)p(y)$. Think about what the product of your estimated pdfs from problem 1 would look like.

3. *Predicting brain sizes from political views*

(a) (10) Using `npcdens`, find the conditional density of the volume of the amygdala as a function of political orientation. (Make sure that you are treating `orientation` as an ordinal variable.) Report the bandwidths. Is the bandwidth for the amygdala the same as either of the previous two bandwidths you have found for it? Should it be? Plot the distribution, and comment on whether it suggests any relationship between the size of this brain region and political orientation.

(b) (5) Repeat this for the conditional density of the ACC as a function of orientation.

4. *Creating a binary response variable*

(a) (1) Create a vector, `conservative`, which is 1 when the subject has `orientation` ≤ 2, and 0 otherwise.

(b) (2) Explain why the cut-off was put at an `orientation` score of 2 (as opposed to some other cut-off).

(c) (1) Check that your `conservative` vector has the proper values, *without* manually examining all 90 entries.

(d) (1) Add `conservative` to your data frame. (Creating a new data frame with a new name will only get you partial credit.)

5. *Logistic regression*

(a) (5) Fit a logistic regression of `conservative` (not `orientation`) on `amygdala` and `acc`. Report the coefficients to no more than three significant digits. Explain what the coefficients mean.

(b) (5) Using case resampling, give bootstrap standard errors and 95% confidence intervals for the coefficients. Was the restriction to three significant digits reasonable?

6. (10) *Generalized additive model.* Fit a generalized additive model for `conservative` on `amygdala` and `acc`. (Be sure to smooth both the input variables.) Make sure you are using a logistic link function. Report the

intercept with reasonable precision. Plot the partial response functions, and explain what they mean (be careful!).

7. *Kernel conditional probability estimation*

   (a) (5) Using `npcdens`, find the conditional probability of `conservative` given `amygdala` and `acc`. Make sure `npcdens` treats `conservative` as a categorical variable and not a continuous one. Report the bandwidths.

   (b) (5) Plot the estimated conditional probability that `conservative` is 1, with `acc` set to its median value and `amygdala` running over the range $[-0.07, 0.09]$. (The plotting range for `amygdala` exceeds the range of values found in the data.) *Hint:* your code will need to provide values for `acc`, for `amygdala` *and* for `conservative` (why?).

   (c) (5) Plot the estimated conditional probability that `conservative` is 1, with `amygdala` set to its median value and `acc` running over the range $[-0.04, 0.06]$. (This plotting range also requires extrapolating outside the data.)

8. *Classification* The models from problems 5–7 predict probabilities for `conservative`. If we have to make a definite prediction of whether someone is conservative or not, we should predict 1 if the probability is ≥ 0.5 and 0 otherwise.

   (a) (10) Find such predictions for each subject, under each of the three models. What fraction of subjects are mis-classified? What fraction would be mis-classified by "predicting" that none of them are conservative?

   (b) (5) Re-calculate the classification error rates using leave-one-out cross-validation for each model.

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots and tables are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. When questions ask for a plot or table, the figure is included in the report, and the code which generated it is part of the source file for the report (i.e., all figures can be reproduced by re-knitting the source file). All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.