Homework 6

Advanced Methods for Data Analysis (36-402/36-608)

Due Tues April 1, 2014 at 11:59pm

1 A little calculus plus statistics goes a long way

(a) Suppose that X is a random variable that depends on a parameter θ. Let $L(\theta; X)$ be the likelihood function; i.e., this is the probability density or mass function of X, under the parameter θ. (As a concrete example, think of X as $N(\theta, 1)$, and so $L(\theta; X) = 1/(\sqrt{2\pi}) \exp(-(X-\theta)^2/2)$.) Let $\ell(\theta; X) = \log L(\theta; X)$ denote the log likelihood. Prove that

$$\mathbb{E}\Big[\frac{d}{d\theta}\ell(\theta; X)\Big] = 0.$$

Hint: you may assume that you can swap the order of differentiation and integration, so that

$$\int_a^b \frac{d}{d\theta} f(\theta, x)\, dx = \frac{d}{d\theta} \int_a^b f(\theta, x)\, dx$$

for a function $f(\theta, x)$.

(b) Use part (a) to show that if $L(\theta; X)$ is an exponential family density (or mass) function,

$$L(\theta; X) = \exp\Big(\frac{X\theta - b(\theta)}{a(\phi)} + c(X, \phi)\Big), \tag{1}$$

then $\mathbb{E}(X) = b'(\theta)$ (where b' is the derivative of b).

(c) Now return to considering a general likelihood $L(\theta; X)$. Prove that

$$\mathbb{E}\Big[\frac{d^2}{d\theta^2}\ell(\theta; X)\Big] = -\mathbb{E}\Big[\frac{d}{d\theta}\ell(\theta; X)\Big]^2.$$

Hint: assume again that you can swap the order of differentiation and integration, any number of times that you want.

(d) Returning to the exponential family form for $L(\theta; X)$ in (1), prove that $\mathrm{Var}(X) = b''(\theta)a(\phi)$ (here b'' denotes the second derivative of b).

(e) Using parts (b) and (d), derive (i.e., verify) the means and variances of the $N(\mu, \sigma^2)$, Bern(p), and Poisson(λ) distributions. Hint: it will be helpful to go back to the lecture notes to see precisely how these can be written in exponential family form.

2 Red brain, blue brain

(credit to Cosma Shalizi)

The data set `n88_pol.csv` contains information on 88 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 body-type 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.

2.1 Correlating brain sizes and political views

(a) Ignoring the fact that `orientation` is an ordinal variable, what is the correlation between it and the volume of the amygdala? Between `orientation` and the volume of the ACC?

(b) Give 95% bootstrap confidence intervals for these correlations (using the usual, i.e., pairs bootstrap).

(c) The function `rank`, applied to a data vector, returns the vector of ranks, where 1 indicates the smallest value, 2 the next-smallest, etc. What are the correlations between the ranks of `orientation` and the ranks of `amygdala`? Between `orientation` and `acc`? Hint: What does `cor(x,y,method="spearman")` do?

(d) Give 95% bootstrap confidence intervals for the rank correlations.

2.2 Creating a binary response variable

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

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

(c) Check that your `conservative` vector has the proper values, *without* manually examining all 88 entries.

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

2.3 Logistic regression

(a) 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) Give bootstrap standard errors and 95% confidence intervals for the coefficients. Was the restriction to three significant digits reasonable?

2.4 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!).

2.5 Classification

The models from Sections 2.3, 2.4 predict probabilities for `conservative`. If we have to make a point prediction of whether someone is conservative or not, we should predict 1 if the probability is ≥ 0.5 and 0 otherwise. Find such predictions for each subject, under each of the two models. What fraction of subjects are mis-classified? What fraction would be mis-classified by simply "predicting" that none of them are conservative?

2.6 Summing up

Explain what you can conclude from this data about the relationship between brain anatomy and political orientation. Refer to your answers to earlier problems.