Homework 6

STAT-GB.4310: Statistics for Social Data
Instructor: Patrick O. Perry

Due April 12, 2016

Application: Sentiment analysis

1. Choose one of the human-labeled text datasets from http://www.crowdflower.com/data-for-everyone/. You do not necessarily have to choose a sentiment analysis dataset; in the discussion that follows, take "orientation" to mean "sentiment", "disastrousness", or whatever makes the most sense depending on your application.

2. If your dataset measures orientation on a numerical scale (e.g., 1–5 or 1–10), choose appropriate cutoffs to code each text as "Postive", "Negative", or "Neutral". This will require some subjective judgment on your part.

3. Divide your dataset into a "training" portion and a "validation" portion, in proportions of 90% and 10%.

4. Use the training dataset to fit at least two of the sentiment analysis methods described in the lecture on "Sentiment Analysis".

5. For each method, use the fitted models to predict the orientations of the texts in the validation set.

6. For each method, report (a) the overall misclassification rate, and (b) a 3-by-3 "confusion" matrix. You should compute these numbers using the validation set, not the training set. Entry $(\mathtt{i}, \mathtt{j})$ of this matrix should report the number of texts in the validation set with true class `i` in the validation set and predicted class `j`, according to your analysis method. (Hint: the `table` command will produce this matrix for you)

7. For the best model you tried, what words have the biggest impact on the predictions?

Theory: Exponential random graph models

Consider the exponential random graph model

$$P_\eta(Y = y) = \frac{1}{\kappa} \exp\Big\{ \sum_A \eta_A g_A(y) \Big\},$$

where

$$\kappa = \kappa(\eta) = \sum_y \exp\Big\{ \sum_A \eta_A g_A(y) \Big\}.$$

The "P_η" notation indicates that the probability depends on the choice of the parameter vector $\eta = (\eta_A)$. The rest of the notation is as described in Robins et. al (2007), "An introduction to exponential random graph (p^*) models for social networks, ", available on the course webpage.

1. For any configuration A, show that

$$\frac{\partial}{\partial \eta_A} [\log \kappa(\eta)] = E_\eta[g_A(y)],$$

   where E_η denotes expectation with respect to the probability measure P_η. (For a general function f, this is given by $E_\eta[f(Y)] = \sum_y P_\eta(Y = y) f(Y)$.)

2. (Optional) For any two configurations A and B, show that

$$\frac{\partial^2}{\partial \eta_A \partial \eta_B} [\log \kappa(\eta)] = \mathrm{cov}_\eta(g_A(Y), g_B(y)).$$