The Last Homework

11/22/2015

Due: Sunday, December 6.

Part 1

Your goal is to classify emails into spam vs ham (not spam). Download file `email.csv`, which contains the dataset. The dataset contains just two fields:

- text: The text of the email.
- spam: A binary variable indicating if the email was spam.

There is a starter script to get you going. Follow the code in `NB_reviews.R` that you saw in the classroom in order to preprocess data and create the document term matrix.

- Use Naive Bayes classifier and report accuracy on the test set.
- How does the accuracy change if you are also using stemming to preprocess emails compared to preprocessing that does not involve stemming?
- One of the parameters that you have to set in order to remove infrequent words is the sparsity parameter. Change the sparsity parameter in some range, say 0.9 and 0.99, and report how does this parameter affect accuracy of the Naive Bayes classifier on the test set.

Part 2

Take a look at part 2 of `starterScript.R` that provides an example code for sampling a random graph according to the stochastic block model. This is an example we saw in the classroom.

Modify the code and generate a network that resembles the eBay fraudster network (see slide 40 of lecture notes on networks). The network should have three communities. Generate a network that has 5 fraudulent nodes, 5 accomplice nodes and 90 honest nodes. Report which stochastic block matrix you used to generate a random graph. Note that the stochastic block matrix needs to be symmetric if you are generating an undirected graph.

Take the random graph you generated and apply the `mmsb.collapsed.gibbs.sampler` function from the `lda` package on the adjacency matrix. Simply calling the code from the starter script on the newly generated adjacency matrix should work. Are you able to recover the three distinct communities?

Optional

Apply the `mmsb.collapsed.gibbs.sampler` on Zachary karate club graph to identify 2 communities. What two communities do you find? Try characterizing them briefly. What do you think went wrong?