CMSC 25025 / STAT 37601

Machine Learning and Large Scale Data Analysis

Assignment 4

Due: Thursday, May 9, 2013

This assignment consists of four problems. Three of them are "pencil and paper" problems to help solidify your understanding of topic modeling. The fourth asks you to calculate topic models on a portion of the wishes data from LSD Project 1, and on the State of the Union data. For extra credit, you can do the fifth problem, which is to implement collapsed Gibbs sampling for LDA. In all of these problems we adopt the notation used in class.

1. *Local Topic Variable Sampling* (20 points)

   (a) Let $z_{1:N}$ denote the topic indicator variables for document d in a K-topic LDA model. The topics are denoted β_k, for $k = 1, \ldots, K$; each of these is a multinomial over a V-word vocabulary.

   Derive the conditional probability distribution

   $$\mathbb{P}(z_n \,|\, z_{-n}, \beta_{1:K}, w_{1:N}, \alpha)$$

   where the topic mixture proportions $\theta_d \sim$ Dirichlet(α) are integrated out. Give a detailed derivation and explanation of each step. Include all normalizing constants.

   (b) Explain how the distribution above can be used to approximate $\mathbb{E}(\theta_d \,|\, \beta_{1:K}, w_{1:N}, \alpha)$ with a Gibbs sampling algorithm.

2. *Big Data, Big Integrals* (20 points)

   Assuming the same latent Dirichlet allocation model as above, derive a closed form expression for the integral

   $$\mathbb{P}(w_{1:N}, z_{1:N} \,|\, \beta_{1:K}, \alpha) = \int \mathbb{P}(w_{1:N}, z_{1:N} \,|\, \theta_d, \beta_{1:K})\, p(\theta_d \,|\, \alpha)\, d\theta_d.$$

3. *Global Topic Variable Sampling* (20 points)

   (a) Let $\boldsymbol{z}$ denote the topic indicator variables for entire collection of words $\boldsymbol{w}$ across all documents in the collection. Derive the conditional probability distribution

   $$\mathbb{P}(z_n \,|\, \boldsymbol{z}_{-n}, \boldsymbol{w}, \alpha, \eta)$$

   of the topic z_n for a specific word w_n in a document, where the topics $\beta_{1:K}$ are integrated out with respect to a prior Dirichlet(η). Give a detailed derivation and explanation of each step. Include all normalizing constants.

(b) Explain how the distribution above can be used to approximate $\mathbb{E}(\beta_k \mid \boldsymbol{w}, \alpha, \eta)$ with a Gibbs sampling algorithm.

4. *Running LDA* (40 points)

For this problem you are asked to build topic models on the SOTU data and part of the Twitter data. The goal is for you to get some feeling for how these models work on different types of data. You should run this on a local computer, not on Amazon AWS. You will need to have R installed (and for some of the graphics you may need R version 2.14 or 2.15). If you want to use Python, please install the RPy package `http://rpy.sourceforge.net/rpy_download.html` or use `sudo apt-get install python-rpy` if you are on Debian Linux or one of its variants.

For the SOTU data, you will be provided with two files `SOTU_stopword_removed.txt` and `SOTU_filenames.txt`. The first file contains the SOTU speeches with stop words removed while the second file contains the names of the speeches. Use the R package `lda`, installed via `install.packages("lda")` to build the topics models.

Recall the parameter α of the Dirichlet distribution generating θ for each document. This parameter is passed in to the function `lda.collapsed.gibbs.sampler`. You can change the value of α and the number of topics K to get different fits.

Comment on the topics you get with different settings of the parameters. Give examples of the topics, and of the posterior distribution over topics for given documents. Do the models make sense? How are the assumptions appropriate or inappropriate for these data?

We have made available some R scripts that you may wish to look at as examples. These generate graphics using the package `ggplot2`. The script `lda_SOTU.R` fits the model and then shows the topic proportions in each speech via the colored bars. An example of running `lda` on the wishes dataset is `lda_wishes.R`. For those using Python, you can execute

```
import rpy
from rpy import *
r.library(<package_name>)
r.source("lda_SOTU.R")
```

5. *Collapsed Gibbs Sampling* (EC points)

For extra credit, implement the collapsed Gibbs sampler for LDA in R or Python. Use it to estimate a topic model on the State of the Union data, and compare the results to those obtained above.