LDA: Technical details

Wouter van Atteveldt

CCS Hannover, Feb 2018

Programme

Tuesday:

- AM: Intro, Text analysis with R
- PM: Topic Modeling: application and validation

Wednesday:

- AM: Technical Details
- PM: Structural Topic Modeling

Thursday:

- AM: Linguistic processing; visualization

Intro: clustering a DTM

- Text is difficult and scary
- But a DTM is just a data matrix, right?
    - Documents are observations (subjects)
    - Terms are measurements
- Can we use 'normal' techniques for clustering text?

Latent Semantic Analysis / Indexing

- LSA/LSI: apply Singular Value Decomposition to DTM
- Similar to factor analysis:
  - Can we find common 'factors' among the terms?
- Found to mimick human generalizations of meaning
- But also found to be problematic:
  - Difficult to interpret (negative values)
  - Not robust to ambiguous terms
  - No theoretical interpretation of mechanism

(Deerwester et.al., 1990, *Indexing by Latent Semantic Analysis*)

SVD and factor analysis

- Any $m \times n$ matrix M can be decomposed into UDV'
  - D is a diagonal matrix with the r *singular values*
- If all singular values are kept, this is lossless ($M = UDV'$)
- By only keeping the first (highest) k values, we effectively reduce the dimensionality:
  - U is $m \times r$ document - factor matrix
  - V' is $n \times r$ term - factor matrix
- U are the PCA factors, and D^2V the factor loadings

See handout graphical interpretation (SVD)

Latent Dirichlet Allocation

- Evolution of LSA
- Full *generative* model:
  - Assume an author draws a mix of topics, and a mix of words from these topics
  - Topics are a probability distribution over words
  - $\rightarrow$ Good interpretability
- Mixture model:
  - Words can be in multiple topics ($\rightarrow$ deals with ambiguity)
  - Documents in multiple topics ($\rightarrow$ deals with mixed content)
    - But skewed towards a couple of topics, depending on α

Dirichlet distribution

- Every topic, document is a probability distribution
    - How likely is word w in topic z, or topic z in document d?
- These are drawn from *dirichlet distribution*

Intermezzo: Restaurant tables

- You walk into the room for the conference dinner and want to sit somehwere
- You prefer not to sit alone, so choose a table with more people
  - $P(t_i) = n_i / \sum_j n_j$
- Everyone does the same
- After a lot of people enter, the distribution converges to equilibrium
- (also known as polya's urn with different colored balls; draw one out and add another one of that color)

http://topicmodels.west.uni-koblenz.de/ckling/tmt/restaurant.html?parameters=3,2,1,5

Dirichlet walked into a restaurant. . .

Dirichlet walked into a restaurant. . .

- The restaurant/urn converges to dirichlet distribution for a given alpha
- The initial number of people at the tables is the alpha hyperparameter
- The resulting final distribution is a single multinomial probability distribution
- Intuitive effect of lower alpha:
    - initial assingments have larger effect
    - likelier that a single table will get all participants
- Alpha's are *hyperparameters*
    - co-determine multinomial *parameters*

Alpha and Dirichlet

(Handout: Dirichlet Distribution and the Alpha)

LDA: what's not to like?

- Interpretable, plausible, elegant
- Generative model gives direct approach to find parameters:
    - a model gives $p(w|m)$, so find m that maximizes this
- Problem: no analytic solution, and no good numerical solution

Enter: gibbs sampling

- Suppose you knew the topics of all words except for one
- That new word *w* in document *d* is a new guest/ball
    - Chance of picking topic z for document d is proportional to existing topics in document plus alpha
    - Similarly, chance of picking topic for the word is proportional to existing topics for word plus alpha
- This gives a probability distribution for z

Iterative gibbs sampling

1. Start with random assignments of topics to words
2. For each word w in document d
    - Compute proportion of topics in word and document
        - (disregarding w itself)
    - Compute probability of each topic z given those proportions
    - Pick a new topic from that probability
    - Update proportions for next iteration
3. Repeat from 2 until converged

(A. Brooks, LDA under the hood,
`https://tinyurl.com/zfpbatb`
Steyvers, M., & Griffiths, T. (2007). *Probabilistic topic models*. Handbook of latent semantic analysis, 427(7), 424-440.)

Gibbs sampling in R

- handout: LDA Animation
- handout: Gibbs sampling in R