Multinomial probit latent dirichlet allocation

Måns Magnusson, Leif Jonsson, Mattias Villani

Linköping University

2014-08-26

Text in statistics

- Abundance of textual data today
- Not many statisticians in this field
- NLP (natural language processing): area of computer science, computional linguistics and machine learning
- Mostly algorithmic solutions (dynamic programming etc.)

Some definitions

- **Words**: words or n-grams (can be stemmed/lemmatized)
- **Documents**: collection of words
- **Corpus**: collection of documents

What is special about modeling of text?

- Unstructured information (but highly structured)
- What are the observations? Words? Sentences? Chapters?
- Vocabulary sizes are often HUGE (Zipf distributed) - high-dimensional
- Many words (usually a lot of **dirty** data - meaningless words)
- One way is to try to summarize the text as topics
  - Latent dirichlet allocation, topic models
  - Think of PCA, but for discrete multinomial data

What is Latent dirichlet allocation?

- Latent dirichlet allocation (LDA) is a hierarchical probabalistic model
- The basic model is used for unsupervised learning (derive topics)
  - Proportions per topic
- Mainly used is modeling textual data
- The basic model assumes that a document is just a "bag of words"

What is Latent dirichlet allocation?

1. For each topic k:
    1. Sample topic-word distribution $\phi_i \sim \text{Dir}(\beta)$
2. For each document D:
    1. Sample the topic proportions $\theta \sim \text{Dir}(\alpha)$
    2. For each word in document d:
        1. Sample topic indicator $z \sim \text{Multinomial}(\theta)$
        2. Sample word $w \sim \textit{Multinomial}(\phi_z)$

Example

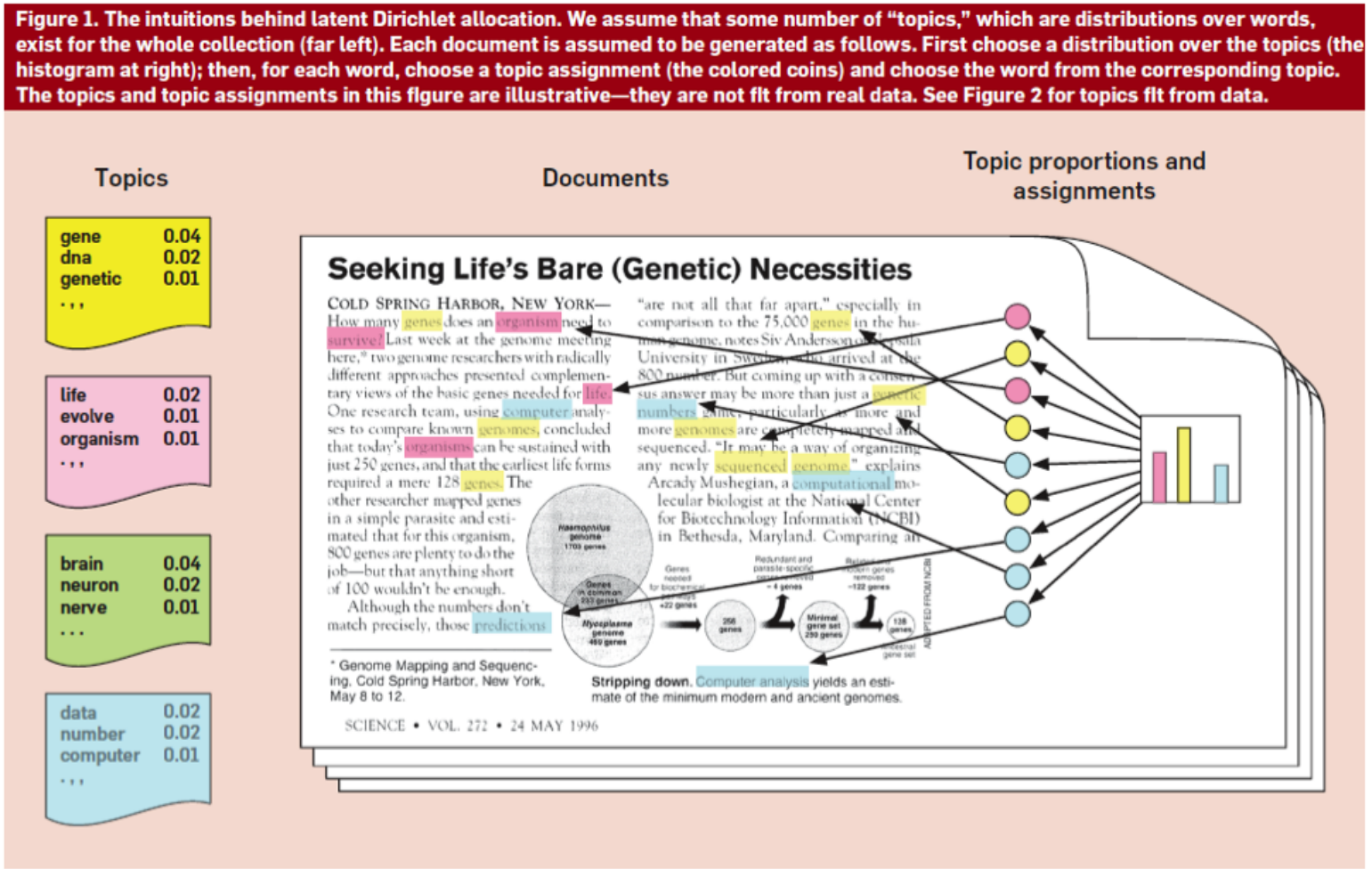

Figure : LDA example (Blei et al. (2010))

Inference methods

- Two main approaches:
  1. Variational bayes (Blei et al. (2003))
  2. Collapsed gibbs sampling (Griffiths and Steyvers (2004))
     - Integrate out Φ and θ

$$p(z_i = j | w_i, \mathbf{z}_{-i}) \propto \underbrace{\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta}}_{word-topic} \cdot \underbrace{\frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + K\alpha}}_{topic-document}$$

     - Sample topic indicator for each word!
- Spectral methods (Arora et al. (2012))

Extensions of LDA

- Supervised topic models (partial least squares analogy)
- Dynamic topic models
- Author topic models
- Incorporating bayesian non-parametrics
- ... etc.

Our problem: Fault localization at Ericsson

- Predict where in the system a fault is located:
  - The parts of the system are linked in a network - classes are not independent
  - Contains both text and structured information
  - Needed for decision (where to look for the fault)
- Our first try: Multinomial probit model

The generative model

1. **For each label** $l \in L$
   1. **Sample a coefficient matrix** $\eta \sim \mathcal{N}(\eta_0, E^{-1})$
2. **Sample a covariance matrix** $\Sigma_{(L-1)\times(L-1)} \sim p(\cdot)$
3. For each topic k:
   1. Sample topic-word distribution $\phi_i \sim \text{Dir}(\beta)$
4. For each document D:
   1. Sample the topic proportions $\theta \sim \text{Dir}(\alpha)$
   2. For each word in document d:
      1. Sample topic indicator $z \sim \text{Multinomial}(\theta)$
      2. Sample word $w \sim \textit{Multinomial}(\phi_z)$
   3. **Sample a latent normal vector:**
      $\mathbf{a}_d \sim \mathcal{N}_{(L-1)}((\ \bar{\mathbf{z}}\quad \mathbf{x}\ )_d^\top \eta, \Sigma_{(L-1)\times(L-1)})$
   4. **Apply class** l **to observation according to** $y_d = \arg\max(\mathbf{a}_d)$

MNP LDA: Gibbs sampler

- Sampler is almost the same as MNP model except for sampling **z**:s:

$$p(z_i = j | w_i, \mathbf{z}_{-i}) \quad \propto \quad \underbrace{\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + K\alpha}}_{LDA} \cdot \underbrace{p(\mathbf{a}|\eta,)}_{MNP}$$

- We need to sample this **for every word**

Problems with MNP LDA

- Problems:
  - A LOT of modes in the posterior
  - Identifiability issues - what is a latent topic?
  - bad mixing with MNP and LDA sampling
  - overweight on the LDA part of the model (class only counts as one other word)

Referenser

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M., 2012. A practical algorithm for topic modeling with provable guarantees. arXiv preprint arXiv:1212.4777.

Blei, D., Carin, L., Dunson, D., Nov. 2010. Probabilistic Topic Models. IEEE Signal Processing Magazine, 77–84.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5563111

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022.

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. . . . academy of Sciences of the United . . . .
URL http://www.pnas.org/content/101/suppl.1/5228.short