Lecture 20: Probabilistic topic models II: LDA (part 3)

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 20

Solving LDA

Directly solving LDA would involve finding parameters that maximize the empirical likelihood $\mathcal{L}$ of the observed documents $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$:

$$\mathcal{L} = \prod_{j=1}^{m} \prod_{i=1}^{n} P(w_{i,j}|z_{i,j}, \phi) P(z_{i,j}|\theta_i) P(\theta_d|\alpha) P(\phi|\beta) \tag{1}$$

- Note: parameters found are not α, β
- Rather, they are parameters to ϕ, θ

These parameters cannot be directly solved as $z_{i,j}$ not observed.

Instead, an approximation method must be used.

Gibbs sampling

A common approach is to use *Gibbs sampling*

- A Monte-Carlo Markov Chain method from statistical physics
  - "Monte Carlo" means based on random simulation
  - "Markov Chain" describes a random process in which each state depends only on the previous state
- Basic idea is in a complex model with many dependent variables:
  - Sequentially sample each variable, dependent upon state of all other variables
  - Observe averages over very large number of samples as probability estimates

Collapsed Gibbs sampling

Method developed by Griffiths and Steyvers, 2006:

- Marginalize out θ, ϕ
- Instead, estimate $P(\mathbf{z}|\mathbf{w})$ (that is, $P(z_{i,j}|w_{i,j})$ for all i, j)
- Using Bayes' Theorem and the Law of Total Probability:

$$\begin{aligned} P(\mathbf{z}|\mathbf{w}) &= \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z})}{P(\mathbf{w})} \\ &= \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}|\mathbf{z})P(\mathbf{z})} \end{aligned}$$

Collapsed Gibbs sampling

By intergrating out (marginalizing, "collapsing") θ, ϕ, derive:

$$P(\mathbf{w}|\mathbf{z}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=i}^{T} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)} \quad (2)$$

$$P(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n_\cdot^{(d)} + T\alpha)} \quad (3)$$

- Γ is the Gamma function
- W, T are number of words, topics
- $n_j^{(w)}$ is number of times word w is assigned to topic j

Don't worry about formula; just note that it involves only:

- actual word assignments to topics, $n_j^{(w)}$
- the Dirichlet priors α, β

Φ, Θ have been marginalized out.

- Equations 2 and 3 cannot be directly solved
  - T^n terms involved (n is number of word instances in corpus)
- However, probability of topic assignment to word instance i can be calculated given topic assignments to all other terms $(\mathbf{z}_{-i}, \mathbf{w})$ as:

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \tag{4}$$

- $n_{-i,j}^{(w_i)}$ is number of times topic j has been assigned to term w_i (excluding word instance i)
- $n_{-i,j}^{(d_i)}$ is number of times topic j has occurred in document d_i (excluding word instance i)

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \tag{4}$$

---

(4) easily (if expensively) approximated using Gibbs sampling:

- Init. topic assignments z_i, by sequentially applying Equation 4
- Resample assignment of each z_i based upon $\mathbf{z}_{-i}$
- Iterate a large number of times
- Estimate $P(\mathbf{w}|\mathbf{z})$, $P(\mathbf{z})$ from set of samples
- Estimate Φ, Θ as:

$$\widehat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta}$$

$$\widehat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}$$

Results of LDA

- Topics characterized by list of high-weight terms
- Terms ordered by decreasing $P(w|z)$
- Unlike LSA, there is no natural ordering between topics

Evaluation of LDA and other topic models

Three main approaches:

Fit of test data Divide test / train split; how well does trained model fit test data?

Application performance Employ topic model in application (e.g. web search); does it improve application performance?

Interpretability Are models interpretable to humans?

Fit of test data: perplexity

In probabilistic language models (e.g. LDA):

- Calculate prob of each test set word instance given model:

$$P(w_{dn}|\text{model}) \tag{5}$$

- Calculate *entropy* ("surprise") across test set:

$$H(\mathbf{w}|\text{model}) = \frac{1}{N}\sum_{d=1}^{M}\sum_{n=1}^{N_d}\log_2(w_{dn}|\text{model}) \tag{6}$$

  Higher the entropy, greater the "surprise", worse the fit

- Often expressed as *perplexity*:

$$\text{Perplexity}(\mathbf{w}|\text{model}) = 2^{H(\mathbf{w}|\text{model})} \tag{7}$$

Application performance

Wei and Croft (2006) examine perforamnce of LDA in IR.

- $P(w|d) = \sum P(w|z)P(z|d)$
- Linearly combined with standard language model
- Performs
  - better than LM alone
  - worse than LM with PRF ("relevance model", RM)
  - slightly better in combination with PRF than PRF alone

Collection	LM	LDA+LM	RM	LDA+RM
AP	0.2179	0.2651	0.2745	0.2869
FT	0.2589	0.2807	0.2835	0.2907
SJMN	0.2032	0.2307	0.2633	0.2603
LA	0.2468	0.2666	0.2614	0.2715
WSJ	0.2958	0.3253	0.3422	0.3606

Table : Average precision, Wei and Croft, 2006)

Human interpretability

- Run LDA on LYRL30k collection, with 100 topics
- Randomly select 5 topics to display, with 10 top words

Topic	Terms
17	bon wilson germ von horton curv steep spd german garcia
29	drachm bangladesh nedlloyd intact eei india fujimor chittagong cbi ftk
76	imf drug russian yeltsin russia murd investig polic moscow sentenc
37	nz shi wellington ite signatur shield janeir trichet underscor unreal
84	heinek actual buyout guild florent lloyd anthon meespierson uefa unilev

How "interpretable" are these topics?

Human interpretability

Tpc	Terms
17	bon wilson germ von horton curv steep spd german garcia
29	drachm bangladesh nedlloyd intact eei india fujimor chittagong cbi ftk
37	nz shi wellington ite signatur shield janeir trichet underscor unreal
76	imf drug russian yeltsin russia murd investig polic moscow sentenc
84	heinek actual buyout guild florent lloyd anthon meespierson uefa unilev

Table : Top 10 words from 10 topics, LYRL30k, LDA

Tpc	Terms
5	ton pct shar wheat stock oil export point aug index
48	japan crown austral sugar zloty carg index gas air singapor
49	crown sugar wheat rupee japan cop austral afric pric eu
70	hung austral dole quart gas afric story rand clinton pound
74	hung taiw quart roman export brazil contract ship crop cotton

Table : Top 10 words from 10 topics, LYRL30k, LSA (negative in red)

- Does LDA (top) give more interpretable topics than LSA (bottom)?

Human intepretability

Chang et al. (2009) take a more controlled approach to measuring human interpretability:

- Intrude a random word into list of words for topics
  - Ask humans to identify the anomalous word
- Intrude a random topic into list of topics describing document
  - Ask humans to identify the anomalous topic
- Comparing LDA, pLSI, and another topic model called "Correlated Topic Models" (CTM).
- Find that human measure uncorrelated or negatively correlated with automated fit measures

Further reading

- Griffiths and Steyvers, "Finding Scientific Topics", *PNAS*, 2004 (collapsed Gibbs sampling for solving LDA)
- Chang, Boyd-Graber, Gerrish, Wang, and Blei, "Reading Tea Leaves: How Humans Interpret Topic Models", *NIPS*, 2009 (finds human interpretability anti-correlated with statistical fit for topic models).
- Wallach, Murray, Salakhutdinov, and Mimno, "Evaluation Methods for Topic Models", *ICML* 2009 (examines various automatic evaluation models based around probabilistic fit).
- Wei and Croft, "LDA-Based Document Models for Ad-hoc Retrieval", *SIGIR* 2006 (uses LDA for information retrieval).