Lecture 17: Probabilistic topic models I: PLSI

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 17

What we'll learn in this lecture

- Review of topic modelling
- Probabilistic version of LSI (pLSI)
- Smoothing pLSI with priors

Clustering

- Clustering partitions terms, or docs, into non-overlapping associations
- Soft clustering allows overlap (term, doc, can be in more than one cluster)
- Bi-clustering simultaneously builds soft clusters on terms and docs, allows associations along both dimensions
- But cluster membership still extrinsic aspect of documents, terms

Topic modelling

In topic modelling:

- Topics represent some "higher-level" associative concept
- Formed by unsupervised learning (like clustering)

But:

- We transform representation of documents (and terms)
- . . . to being intrinsically represented by topics as features

Supports:

- Synonymy: different terms with same meaning will be in same topic
- Polysemy: different meanings of same term will occur in different topics
- Topical analysis of text corpora

Topic modelling with LSA / LSI

$$\mathbf{X}_{t\times d} = \mathbf{T}_{t\times t}\mathbf{\Sigma}_{t\times d}(\mathbf{D}_{d\times d})^T \quad (1)$$

$$\widehat{\mathbf{X}}_{t\times d} = \widehat{\mathbf{T}}_{t\times k}\widehat{\mathbf{\Sigma}}_{k\times k}\left(\widehat{\mathbf{D}}_{d\times k}\right)^T \quad (2)$$

- LSI does SVD then takes k largest singular values from $\mathbf{\Sigma}$
- These k values represent "topics"
- And σ_k gives "importance" of topic
- Search, clustering can be done on $\widehat{\mathbf{X}}$

Topics, documents, terms

- $\widehat{\mathbf{T}}_{.z}$ gives terms associated with topic z
- $\widehat{\mathbf{T}}_{t}$ gives importance of term t to each topic
- $\widehat{\mathbf{D}}_{.z}$ gives docs associated with topic z
- $\widehat{\mathbf{D}}_{d}$ gives importance of document d to each topic
- We can find topics of new document $\mathbf{d}$ by

$$\widehat{\mathbf{d}} = \mathbf{\Sigma}^{-1}\widehat{\mathbf{U}}^{T}\mathbf{d} \tag{3}$$

- But can't find topics of new term $\mathbf{t}$

Weaknesses of LSA for topic modelling

$$\widehat{\mathbf{X}}_{t \times d} = \widehat{\mathbf{T}}_{t \times k} \widehat{\mathbf{\Sigma}}_{k \times k} \left(\widehat{\mathbf{D}}_{d \times k} \right)^T \quad (4)$$

- LSA has poor probabilistic / theoretical foundation
- Difficult to interpret, reason about topic–term and topic–document strengths:
  - If a document has terms t_1 and t_2, how strongly is it associated with topic z?
- Difficult to extend to other forms of evidence
- Difficult to repurpose for other, related problems

(All the problems with geometric models we observed in IR)

Probabilistic LSI (pLSI)

Probabilistic LSI (Hoffman, 1999) casts topic modelling in probabilistic terms.

Works from the following *generative model* for how word w comes to be in document d:

1. Select document d with probability $P(d)$
2. Pick topic z from $i \in \mathcal{Z} = \{z_i, \ldots, z_K\}$ with probability $\theta_{di} = P(z = i|d)$
3. Pick term t with probability $\phi_{iv} = P(w = t|z = i)$

- Introduces *latent topic variable* z to explain relation of w and d.
- We have to select the number of latent topics K

$P(d, w)$

This generative model for observing the pair (w, d) gives the following mixture model for $P(w, d)$:

$$
\begin{aligned}
P(d, w) &= P(d)P(w|d) && (5)\\
P(w|d) &= \sum_{z \in \mathcal{Z}} P(w|z=i)P(z=i|d) && (6)
\end{aligned}
$$

Now all we have to do is estimate $P(d)$, $P(w|z=i)$, and $P(z=i|d)$

Relationship of pLSI with LSI

Can rewrite:

$$P(d, w) = P(d) \sum_{z \in \mathcal{Z}} P(w|z = i)P(z = i|d) \tag{7}$$

as:

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(d|z = i)P(z = i)P(w|z = i) \tag{8}$$

This has similar form to LSI:

- $\widehat{\mathbf{\Sigma}} \Rightarrow P(z = i)$, importance of topic i
- $\widehat{\mathbf{D}} \Rightarrow P(d|z = i)$, relation between document d and topic i
- $\widehat{\mathbf{T}} \Rightarrow P(w|z = i)$, relation between word w and topic i

Solving pLSI

$$P(d, w) = P(d) \sum_{z \in \mathcal{Z}} P(w|z = i) P(z = i|d) \tag{9}$$

How to find $P(d)$, $P(w|z)$, and $P(z|d)$ given corpus $\mathbf{X}$?

- Express as log-likelihood
- Find maximum likelihood values for probabilities

Log-likelihood

Given $P(d)$, $P(z|d)$, and $P(w|z)$, the log likelihood of data $\mathbf{X}$ is:

$$\begin{aligned}
\mathcal{L} &= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} x_{wd} \log P(w, d) \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} x_{wd} \log \sum_{i=1}^{K} P(w = v|z = i) P(z = i|d) P(d)
\end{aligned}$$

Maximum likelihood values for above log-likelihood found using an EM (Expectation–Maximization) algorithm.[1]

[1] See Hofmann, 1999, or Crain et al., 2012, for details.

Interpretating pLSI

Proposed "name"	Top terms
"plane"	plane, airport, crash, flight, safety ...
"shuttle"	space, shuttle, mission, astronauts, launch ...
"family"	home, family, like, love, kids ...
"Hollywood"	film, move, music, new, bets ...

Table : Example topics identifed on TDT-1 corpus (Hofmann, 1999)

- Topic can be represented by its highest-weight terms
- I.e. those having highest $P(w|z)$
- These values interpretable as probabilities (obviously)

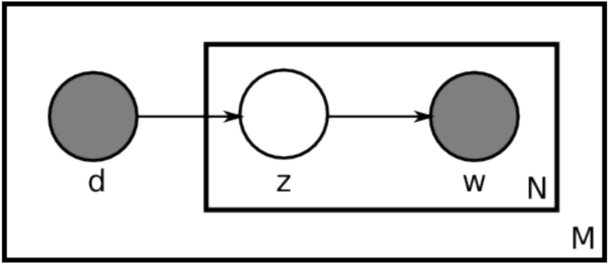
Limitations to pLSI

- pLSI is a maximum likelihood method
- It can therefore only assign probabilities to seen events
  - Can't assign probabilities to new documents
  - Can't assign probabilities to new terms
- Also, risk of "over-fitting" data it observes
- As with LM for IR, both problems can be addressed by *smoothing*
- Or (more formally) assigned *prior probabilities* (prior distributions) to events

Plate notation

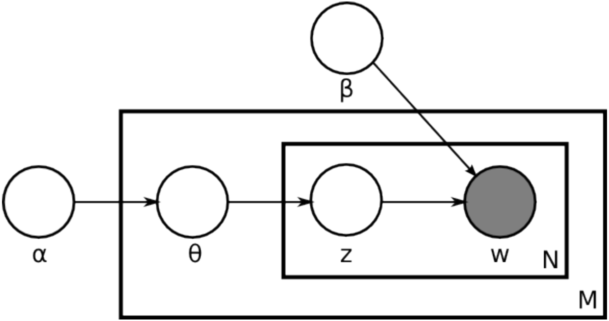
- Complex (e.g. generative, mixture) probabilistic models have multiple, related variables
- Helpful to represent by a graphical notation
- *Plate notation* a commonly use notation:
  - Shows variables, distinguishing between:
    - Latent and seen
    - Discreet and continuous (optionally)
  - Dependencies between variables
  - Cardinality of variables

Plate notation for pLSI



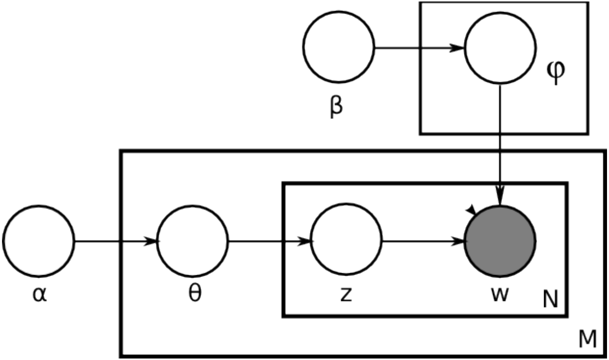
- There are M documents, $\{d_1, \ldots, d_i, \ldots, d_M\}$
- There are N words in document d_i, $\{w_{i1}, \ldots, w_{ij}, \ldots, w_{iN}\}$
- Topic z depends on document d (d generates z)
- Word w depends on topic z (z generates z)
- Word w is conditionally independent of d, given z
- w and d are observable; z is latent (hidden)

Introducing priors



- We want to introduce a prior on documents, and on terms
- Call the term prior β
- Call the document prior α
- And represent the document as its distribution θ_i over topics

Smoothing topics



- In previous model, we smoothed $P(w)$
- Alternatively, we can smooth $P(w|z)$
- i.e. give a different prior to each topic distribution

This is the model of Latent Dirichlet Allocation (LDA) . . . which we'll discuss next lecture

Looking back and forward

Back

- Topic models represent documents as topic mixtures
- Generative model provides probabilistic understanding of document formation
- Probabilistic LSI (pLSI):
  - Pick document
  - Pick topic given document
  - Pick word given topic
- Estimate values using EM
- Gives $P(w|z)$, $P(d|z)$, $P(z)$
- Smoothing to handle unseen words, documents
- Gives LDA (next lecture)

Looking back and forward

Forward

- Latent Dirichlet Allocation (LDA)
  - Current "state-of-the-art" topic model

Further reading

- Thomas Hofmann, "Probabilistic Latent Semantic Indexing", SIGIR 1999 (Original description of pLSI)
- Crain, Zhou, Yang, and Zha, "Dimensionality Reduction and Topic Modeling", Chapter 5 of Aggarwal and Zhai (ed.), *Mining Text Data*, 2012 (good but mathematical summary of topic modeling using LSI, pLSI, and LDA).