Probabilistic latent semantic analysis

Probabilistic latent semantic analysis (PLSA), also known as **probabilistic latent semantic indexing** (**PLSI**, especially in information retrieval circles) is a statistical technique for the analysis of two-mode and co-occurrence data. In effect, one can derive a low dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis. PLSA evolved from latent semantic analysis.

Compared to standard latent semantic analysis which stems from linear algebra and downsizes the occurrence tables (usually via a singular value decomposition), probabilistic latent semantic analysis is based on a mixture decomposition derived from a latent class model.

Model

Considering observations in the form of co-occurrences (w,d) of words and documents, PLSA models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions:

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

The first formulation is the *symmetric* formulation, where w and d are both generated from the latent class c in similar ways (using the conditional probabilities $P(d|c)$ and $P(w|c)$), whereas the second formulation is the *asymmetric* formulation, where, for each document d, a latent class is chosen conditionally to the document according to $P(c|d)$, and a word is then generated from that class according to $P(w|c)$. Although we have used words and documents in this example, the co-occurrence of any couple of discrete variables may be modelled in exactly the same way.

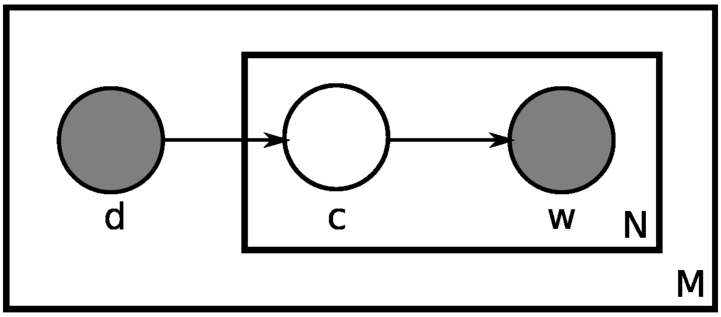

Plate notation representing the PLSA model ("asymmetric" formulation). d is the document index variable, c is a word's topic drawn from the document's topic distribution, $P(c|d)$, and w is a word drawn from the word distribution of this word's topic, $P(w|c)$. The d and w are observable variables, the topic c is a latent variable.

So, the number of parameters is equal to $cd + wc$. The number of parameters grows linearly with the number of documents. In addition, although PLSA is a generative model of the documents in the collection it is estimated on, it is not a generative model of new documents.

Their parameters are learned using the EM algorithm.

Application

PLSA may be used in a discriminative setting, via Fisher kernels.[1]

PLSA has applications in information retrieval and filtering, natural language processing, machine learning from text, and related areas.

It is reported that the aspect model used in the probabilistic latent semantic analysis has severe overfitting problems.

In 2012, pLSA has also been used in the bioinformatics context, for prediction of Gene Ontology biomolecular annotations.[2]

Extensions

- Hierarchical extensions:
  - Asymmetric: MASHA ("Multinomial ASymmetric Hierarchical Analysis") [3]
  - Symmetric: HPLSA ("Hierarchical Probabilistic Latent Semantic Analysis") [4]
- Generative models: The following models have been developed to address an often-criticized shortcoming of PLSA, namely that it is not a proper generative model for new documents.
  - Latent Dirichlet allocation - adds a Dirichlet prior on the per-document topic distribution
- Higher-order data: Although this is rarely discussed in the scientific literature, PLSA extends naturally to higher order data (three modes and higher), i.e. it can model co-occurrences over three or more variables. In the symmetric formulation above, this is done simply by adding conditional probability distributions for these additional variables. This is the probabilistic analogue to non-negative tensor factorisation.

History

It was introduced in 1999 by Jan Puzicha, who later founded together with Derek Schueren the company Recommind, and Thomas Hofmann,[5] and it is related [6] to non-negative matrix factorization.

References and notes

[1] Thomas Hofmann, *Learning the Similarity of Documents : an information-geometric approach to document retrieval and categorization* (http://www.cs.brown.edu/people/th/papers/Hofmann-NIPS99.ps), Advances in Neural Information Processing Systems 12, pp-914-920, MIT Press, 2000

[2] *"Probabilistic Latent Semantic Analysis for prediction of Gene Ontology annotations"*, Marco Masseroli, Davide Chicco, Pietro Pinoli. IEEE WCCI 2012 - the 2012 IEEE World Congress on Computational Intelligence proceedings. Brisbane, Australia, June 2012. (.pdf) (http://home.dei.polimi.it/chicco/Wcci2012_DavideChicco_et_al.pdf)

[3] Alexei Vinokourov and Mark Girolami, A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections (http://citeseer.ist.psu.edu/rd/30973750,455249,1,0.25,Download/http://citeseer.ist.psu.edu/cache/papers/cs/22961/http:zSzzSzcis.paisley.ac.ukzSzvino-ci0zSzvinokourov_masha.pdf/vinokourov02probabilistic.pdf), in *Information Processing and Management*, 2002

[4] Eric Gaussier, Cyril Goutte, Kris Popat and Francine Chen, A Hierarchical Model for Clustering and Categorising Documents (http://www.xrce.xerox.com/Research-Development/Publications/2002-0044), in "Advances in Information Retrieval -- Proceedings of the 24th BCS-IRSG European Colloquium on IR Research (ECIR-02)", 2002

[5] Thomas Hofmann, *Probabilistic Latent Semantic Indexing* (http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf), Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999

[6] Chris Ding, Tao Li, Wei Peng (2006). " Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence Chi-Square Statistic, and a Hybrid Method. AAAI 2006 (http://www.aaai.org/Papers/AAAI/2006/AAAI06-055.pdf)

External links

- Probabilistic Latent Semantic Analysis (http://www.cs.brown.edu/people/th/papers/Hofmann-UAI99.pdf)
- Complete PLSA DEMO in C# (http://www.semanticsearchart.com/researchpLSA.html)

Article Sources and Contributors

Probabilistic latent semantic analysis *Source*: http://en.wikipedia.org/w/index.php?oldid=573293522 *Contributors*: A3 nm, Arcenciel, Bkkbrad, CheekyMonkey, Chiccodoro, EduardoValle, Efsunselin, Erniesgrove, Fnielsen, Jitse Niesen, Johnchallis, Jonsafari, Keretapi, Kku, Larry.europe, Maghnus, Mbell, Mcld, Mehdiym, Melcombe, Michael Hardy, Oleg Alexandrov, Pmj005, Rama, Seo01, Ste1n, Sunny house, Sylenius, Transcendence, Vishvas vasuki, Xin Alan Rong, 18 anonymous edits

Image Sources, Licenses and Contributors

Image:Plsi 1.svg *Source*: http://en.wikipedia.org/w/index.php?title=File:Plsi_1.svg *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: Xin Alan Rong

License