Latent variable modeling

Ella Bingham
Neural Networks Research Centre
Lab of Computer and Information Science
ella@iki.fi

Problem setting

Task: represent a large data set in a compressed format.
Data not completely random but contains regularities, forming a kind of internal structure
→ Find out this latent structure and thus obtain a simple representation of the data.

latent: hidden, unobserved, unknown

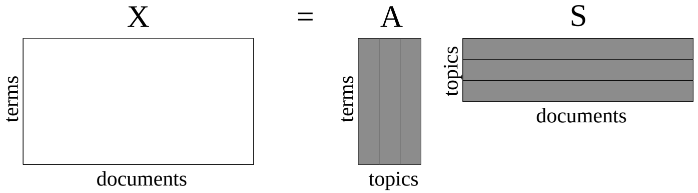
Problem formulation for those who like . . .

- applications:
  topics in text documents, brain activities in MEG, . . .
- clustering:
  multi-way clustering into overlapping groups
- matrices:
  decompose an observed matrix into $X = AS$
- mixture models:
  data is a mixture of latent variables. Some differences remain!

- Bayesian data analysis:
  the prob distr of data is a (convex) combination of distrs of latent variables.
  Alternatively, the (mean) parameter of the distr of data is a convex combination of the (mean) parameters of the distrs of latent vars.
  In any case, something like $P = AS$ can be written.
- Bayesian networks: the latent variables are independent of each other

Data representation

Bag of words: The observed matrix X contains term counts or tf-idf weights or other.



Could also be customers and transactions, or users and web pages, etc.

Methods

- **Principal component analysis (PCA)**: The latent variables are uncorrelated with each other, and capture most of the variance in the data. Solved by eigenvalue decomposition of the covariance matrix of observed data. Suitable for continuous (Gaussian) data. Also called SVD or LSA.
- **Factor analysis**
- **Nonnegative matrix factorization (NMF)**: (Lee and Seung) All matrices have nonnegative entries

- **Independent component analysis (ICA)**: (Hyvärinen+ 2001, Bingham+ 2003) The observed data is generated by a combination of non-Gaussian, statistically independent latent variables (= topic activities in documents). Solved by approximating information theoretic measures of independence. Fast algorithms exist for continuous data. For non-continuous cases, Bayesian approaches have been presented.

- **Mixture models** are different in that they usually assume a multidimensional observation (a doc) being generated by one latent variable (topic) although generation probabilities are given to all latent variables; and generation of different dimensions of the observation (terms in the doc) is not analyzed

$$p(\mathbf{x}) = \sum_k \pi_k p_k(\mathbf{x}|\theta_k)$$

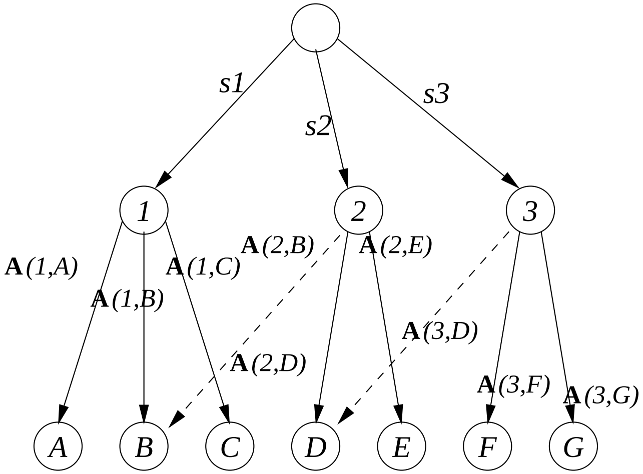
- **Probabilistic latent semantic analysis (PLSA)**: (Hofmann 2001) An occurrence of a term in a document results from first picking a topic and then generating a term from it.
  Can be seen as a matrix decomposition: matrices in $P = AS$ are prob("term appears in a doc"), prob("topic generates a term") and prob("topic is active in a doc"), of which none is observed, only the term by doc matrix X of multinomial counts.

  $$P(\text{term,doc}) = \sum_{\text{topic}} A(\text{topic,term})S(\text{doc,topic})$$

  Solved by the EM algorithm.

- **Latent Dirichlet allocation (LDA) / Multinomial PCA**: (Blei+ 2001/2003, Minka+ 2002, Buntine 2002) Can assign probabilities to unseen documents, and less parameters than in PLSA.
  Heavily Bayesian! EM algorithm needs variational approaches
- Topic models

Topic model



Joint work with Heikki Mannila and Jouni Seppänen.

- Model for binary data. An observed doc vector lists the presence (1) or absence (0) of each term
- In a given document, some topics are active. If a topic is active, it generates some terms with some probabilities.
- We study the probabilities s_i of topics in docs, and the probabilities $\mathbf{A}(i, \cdot)$ of terms in topics
- The model is a Bayesian network: given a topic, the terms are independent. Acyclic, directed graph.
- The task is to infer the topics, given observed term frequencies and pairwise term frequencies

Estimation of the topic structure: Lift

(Bingham+ 2002 (outdated), Seppänen+ 2003).

$$\text{lift}(A,B) = \frac{P(A \mid B)}{P(A)} = \frac{P(A,B)}{P(A)P(B)} \qquad (1)$$

which equals 1 if terms A and B are independent (that is, they belong to different topics) and the larger the lift statistic is, the more dependent the occurrences of A and B are.

Use a "soft" clustering algorithm to find overlapping groups of terms – these are now the topics.

Estimation of the topic structure: Probe dist.

- If two terms A and B belong to the same topic, they behave similarly with respect to any third term
- The information that A's occurrence gives is about the same as the information that B's occurrence gives
- probe distance: $d(A,B) = \sum_{C \neq A,B} |prob(C|A) - prob(C|B)|$
- Terms with a small probe distance typically belong to the same topic. Use soft clustering to find the topic structure.

Experimental results

Simulated data
Probe and ratio algorithms perform quite well compared to NMF, ICA and PLSA that are computationally heavier.

Bibliographical data on computer science
Probe distances clustered. Some topics are names of journals/conferences, some are research areas. Several topics may apply to one document

topic	terms
1	algorithm algorithms efficient fast graph graphs matching optimal parallel problem set simple
2	actainf beatcs damath dmath focs geometry icalp infctrl ipl jacm jcss libtr mfcs sicomp stacs stoc tcs tr
3	complexity functions machines probabilistic
4	applications problems some
5	approach de logic model programming programs system systems van
6	network networks routing sorting
7	computational information theory
8	linear new two
9	binary search tree trees
10	polynomial time
11	algebraic automata finite languages note properties sets theorem
12	data structures
13	analysis design distributed using
14	computation computing
15	bounds lower
16	computer science
17	from learning
18	cacm crypto ieeetc lncs
19	number random
20	abstract extended
21	finding minimum planar

References

(1) Ella Bingham, Ata Kabán, and Mark Girolami. Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*, 17(1):69–83, 2003.

(2) Ella Bingham, Heikki Mannila, and Jouni K. Seppänen. Topics in 0-1 data. In *Proc SIGKDD 2002*, pages 450–455.

(3) David M. Blei, Andrew Y. Ng, and Michael I.

Jordan. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14*, 2001.

(4) David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol 3, pages 993–1022, 2003.

(5) Wray Buntine. Variational extensions to EM and multinomial PCA. In *ECML 2002*, pages 23–34.

(6) Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

(7) Aapo Hyvärinen, Juha Karhunen, and Erkki Oja.

Independent component analysis. Wiley Interscience, 2001.

(8) Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, October 1999.

(9) Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proc 18th Conf on Uncertainty in Artificial Intelligence*, 2002.

(10) Jouni K. Seppänen, Ella Bingham and Heikki Mannila. A simple algorithm for topic identification in 0-1 data. In *Proc PKDD 2003*, pages 423–434.