PROBABILISTIC LATENT SEMANTIC ANALYSIS

Lingjia Deng

Revised from slides of Shuguang Wang

CS3750

Outline

- Review of previous notes
  - PCA/SVD
  - HITS
- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Applications
  - pHITS
- Limitations of Probabilistic Latent Semantic Analysis

Outline

- Review of previous notes
  - PCA/SVD
  - HITS
- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Applications
  - pHIST
- Limitations of Probabilistic Latent Semantic Analysis

Review of PCA/SVD

- Vector Space Model
  - e.g. term-document co-occurrence matrix A
  - limitation: high dimension
  - solution: principal component of the original matrix, to lower the dimension
- PCA
  - subtract the mean
  - get the co-variance matrix
  - calculate the eigenvalues and eigenvectors of co-variance matrix
- SVD
  - $A = USV^T$
  - U: left eigenvectors
  - V: right eigenvectors
  - S: diagonal matrix of eigenvalues

Review of HITS

- Authority-Hub Webpages
  - each webpage has an authority score *x* and a hub score *y*
  - authorities: most definitive information sources (on a specific topic)
  - hubs: most useful compilation of links to authoritative webpages
- A good hub is a page that points to many good authorities;
- A good authority is a page that is pointed to by many good hubs

Iteration Solution of HITS

- Translate mutual relationship into iterative update equations
- Adjacency Matrix A

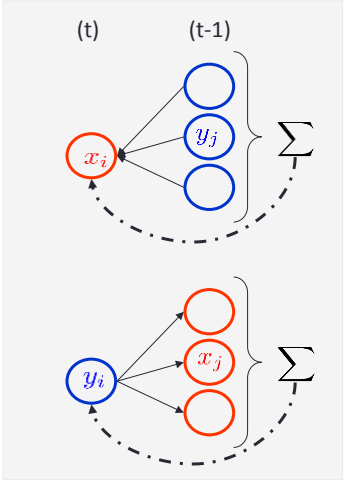
$$\mathbf{A} = (a_{ij}), \quad a_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

x: a vector of authority scores

$$\mathbf{x}^{(t)} \propto \mathbf{A}^T \mathbf{y}^{(t-1)},$$

y: a vector of hub scores

$$\mathbf{y}^{(t)} \propto \mathbf{A}\, \mathbf{x}^{(t-1)}$$

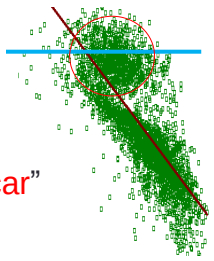


SVD Solution of HITS

- Another solution: Apply SVD to Adjacency Matrix A
  - A = XSY
  - x_k: hub, eigenvector of A^TA
  - y_k: authority, eigenvector of AA^T
- In the iteration solution,
  - x: a vector of authority scores
  - y: a vector of hub scores
  - x and y converge to the largest principle component of the adjacency matrix.

Community On the Web

- There are multiple communities on the web
  - webpage around red line is in "car" community
  - webpage around blue line is in "music" community
- In the iteration solution, x and y are about "car"
  - most dominant community



- In the SVD solution, we have multiple eigenvectors
  - The largest correspond to authority and hub webpges in the red "car" community
  - The second eigenvector correspond to authority and hub webpages in the blue "music" community
  - The third ……

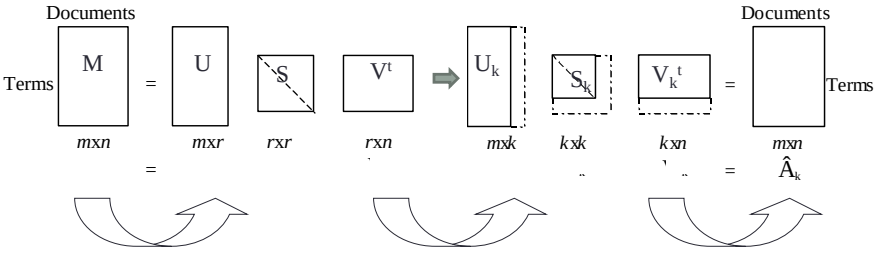
Outline

- Review of previous notes
  - PCA/SVD
  - HITS
- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Applications
  - pHITS
- Limitations of Probabilistic Latent Semantic Analysis

Latent Semantic Analysis

- Latent Semantic Analysis/Indexing (LSA/LSI)
  - LSA is a technique of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.
  - In the context of its application to information retrieval, it is called LSI.
- Perform a low-rank approximation of term-document matrix
- Represent documents (and terms) as vectors in a lower-dimensional space whose axes are concepts that effectively group together similar words
- These axes are the Principal Components from PCA

LSA Process



SVD
Convert term-by-document matrix into 3matrices U, S and V

Dimension Reduction
Ignore zero and low-order rows and columns
r → k

Reconstruct Matrix
Use the new matrix to process queries
OR, map query to reduced space

LSA Pros

- Low-dimensional document representation is able to capture synonyms. Synonyms will fall into same/similar concepts.
- Noise removal and robustness by dimension reduction.
- Correlation analysis and Query expansion (with related words)
- Empirical study shows it outperforms naïve vector space model
- Language independent
- High recall: query and document terms may be disjoint
- Unsupervised/completely automatic

LSA Cons

- No probabilistic model of term occurrences.
- Implicit Gaussian assumption, but term occurrence is not normally distributed.
- Euclidean distance is inappropriate as a distance metric for count vectors (reconstruction may contain negative entries).

- Directions are hard to interpret.
- Ad hoc selection of the number of dimension k
- Computational complexity is high: O(min(mn2,nm2)) for SVD, and it needs to be updated as new documents are found/updated.

- Problem of polysemy (multiple meanings for the same word) is not addressed.

Outline

- Review of previous notes
  - PCA/SVD
  - HITS
- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Applications
  - pHITS
- Limitations of Probabilistic Latent Semantic Analysis

Probabilistic LSA: a statistical view of LSA

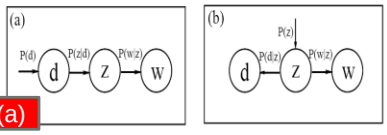
- In LSA: "concept", a dimension in the reduced space
- Aspect Model (generative)
  - For co-occurrence data which associated with a latent class variable,
  - *d* is document, *w* is term, *z* is concept,
  - *d* and *w* are independent conditioned on *z*.

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

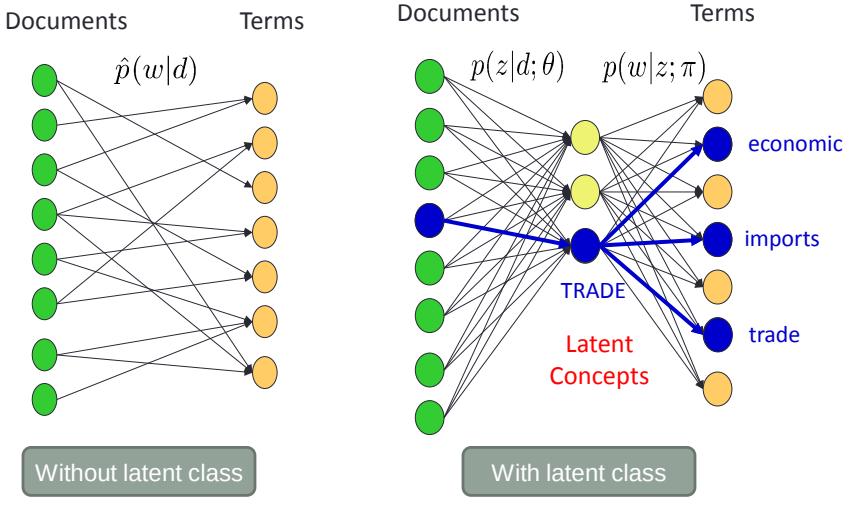
$$= \sum_{z\in Z} P(d)P(w|z)P(z|d) \quad \text{(a)}$$

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

$$= \sum_{z\in Z} P(z)P(w|z)P(d|z) \quad \text{(b)}$$



PLSA Illustration



Why Latent Concept?

- Sparseness problem, terms not occurring in a document get zero probability
- Probabilistic dimension reduction
- No prior knowledge about concepts required

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

An interpretation: KL Projection

- Maximizing the log likelihood

$$L = \sum_{d \in D, w \in W} n(d,w) \log P(d,w)$$

$$\mathcal{L} = \sum_{d \in \mathcal{D}} n(d) \left[\sum_{w \in \mathcal{W}} \frac{n(d,w)}{n(d)} \log P(w|d) + \log P(d) \right]$$

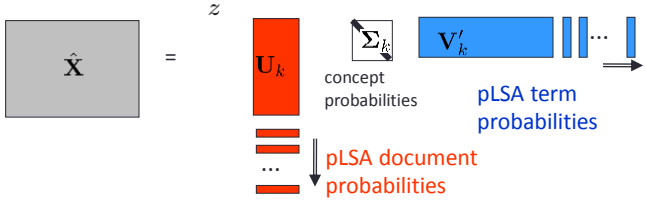
- Minimizing the KL Divergence
  - Empirical word distribution $P = \hat{P}(w|d) = \frac{n(d,w)}{n(d)}$
  - Model distribution $Q = P(w|d)$
  - KL divergence $-P \log \frac{1}{Q}$

$$D_{\mathrm{KL}}(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

PLSA vs. LSA/SVD

- PLSA is based on mixture decomposition derived from latent class model.

$$\hat{p}_{\mathrm{LSA}}(d,w) = \sum_{z} p(d|z)\, p(z)\, p(w|z)$$



- Different from LSA/SVD:
  - pLSA is always non-negative
  - pLSA has no Gaussian distribution requirement

PPCA vs. PLSA

- pPCA is also a probabilistic model.
- pPCA assume normal distribution, which is often not valid.
- pLSA models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions.
- Multinomial distribution is a better alternative in this domain.

PLSA via EM

- E-step: estimate posterior probabilities of latent variables, ("concepts")

$$P(z \mid d, w) = \frac{P(d \mid z)P(w \mid z)P(z)}{\sum_{z'} P(d \mid z')P(w \mid z')P(z')}$$

Probability that the occurence of term w in document d can be "explained" by concept z

- M-step: parameter estimation based on expected statistics.

$$P(w \mid z) \propto \underbrace{\sum_{d} n(d, w)P(z \mid d, w)}$$

how often is term w associated with concept z

$$P(d \mid z) \propto \underbrace{\sum_{w} n(d, w)P(z \mid d, w)}$$

how often is document d associated with concept z

$$P(z) \propto \underbrace{\sum_{d,w} n(d, w)P(z \mid d, w)}$$

probability of concept z

PLSA Summary

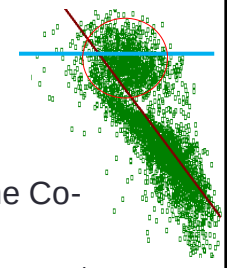
- Optimal decomposition relies on likelihood function of multinomial sampling, which corresponds to a minimization of KL divergence between the empirical distribution and the model.
- Problem of polysemy is better addressed.
- EM approach gives local solution.

- Number of parameters increases linearly with number of documents.
- Not a generative model for new documents.

Outline

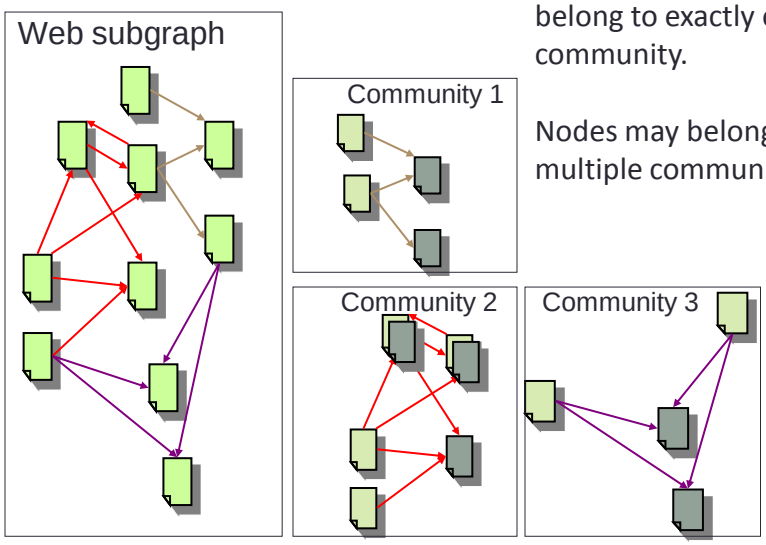
- Review of previous notes
  - PCA/SVD
  - HITS
- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Applications
  - pHITS
- Limitations of Probabilistic Latent Semantic Analysis

Probabilistic HITS



- Probabilistic version of HITS.
- We try to find out the web communities from the Co-citation matrix.
- Loading on eigenvector in the case of HITS does not necessarily reflect the authority of document in community.
- HITS uses only the large eigenvector and this is not necessary the principal community.
- What about smaller communities? (smaller eigenvectors) They can be still very important.
- Mathematically equivalent as pLSA

Web Communities



Links (probabilistically) belong to exactly one community.

Nodes may belong to multiple communities.

Finding Latent Web Communities

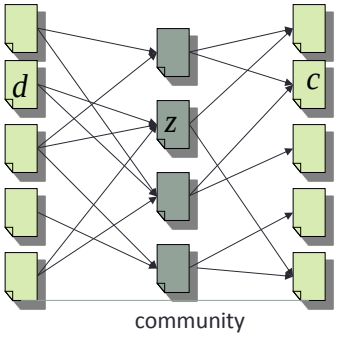
- Web Community: densely connected bipartite subgraph
- document d (hub)
- citation c (authority)

$$P(d,c) = \sum_z P(z)P(d \mid z)P(c \mid z)$$

Source nodes

$P(d \mid z)$

probability that a random out-link from d is part of the community z



Target nodes

$P(c \mid z)$

probability that a random in-link toward c is part of the community z

PHITS via EM

- E-step: estimate the expectation of latent "community".

$$P(z \mid d,c) = \frac{P(d \mid z)P(c \mid z)P(z)}{\sum_{z'} P(d \mid z')P(c \mid z')P(z')}$$

Probability that the particular document –citation pair is "explained" by community z

- M-step: parameter estimation based on expected statistics.

$$P(c \mid z) \propto \underbrace{\sum_{d} n(d,c)P(z \mid d,c)}$$

how often is citation c associated with community *z*

$$P(d \mid z) \propto \underbrace{\sum_{w} n(d,c)P(z \mid d,c)}$$

how often is document *d* associated with community *z*

$$P(z) \propto \underbrace{\sum_{d,w} n(d,c)P(z \mid d,c)}$$

probability of community *z*

Interpreting the PHITS Results

- Simple analog to authority score is P(c|z).
  - How likely a citation *c* is to be cited from within the community *z*.
- P(d|z) serves the same function as hub score.
  - The probability that document *d* points to (i.e., contains a citation to) a given community *z*.
- Document classification using P(z|c).
  - Classify the citation according its community membership.
- Find characteristic citaions of a community with P(z|c) * P(c|z).

Problems with Link-only Approach (e.g. pHITS)

- Because link-only approach only considers the link's relation between webpages.

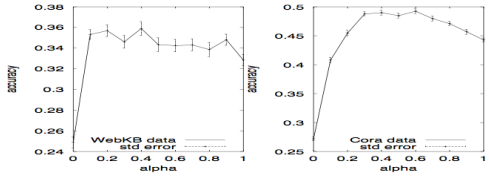
- Not all links are created by human.
- The top ranked authority pages may be irrelevant to the query if they are just well connected.
- Web Spam.

Joint Model of PLSA and PHITS

- Joint probabilistic model of document content (pLSA) and connectivity (pHITS).
- Able to answer questions on both structure and content.
- Likelihood is

$$\mathcal{L} = \sum_j \left[\alpha \sum_i \frac{N_{ij}}{\sum_{i'} N_{i'j}} \log \sum_k P(t_i|z_k) P(z_k|d_j) + (1-\alpha) \sum_l \frac{A_{lj}}{\sum_{l'} A_{l'j}} \log \sum_k P(c_l|z_k) P(z_k|d_j) \right]$$

- EM approach to estimate the probabilities.

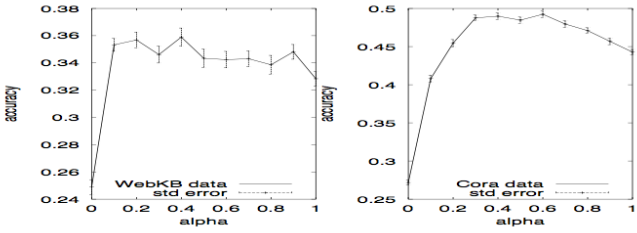


Joint Model of PLSA and PHITS

- Likelihood is

$$\mathcal{L} = \sum_j \left[\alpha \sum_i \frac{N_{ij}}{\sum_{i'} N_{i'j}} \log \sum_k P(t_i|z_k) P(z_k|d_j) + (1-\alpha) \sum_l \frac{A_{lj}}{\sum_{l'} A_{l'j}} \log \sum_k P(c_l|z_k) P(z_k|d_j) \right]$$

- EM approach to estimate the probabilities.



Outline

- Review of previous notes
  - PCA/SVD
  - HITS
- Latent Semantic Analysis
- Probability Latent Semantic Analysis
- Applications
  - pHITS
- Limitations of Probabilistic Latent Semantic Analysis

Tempered EM

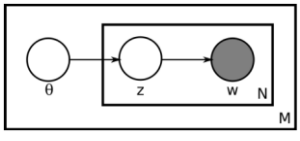
- One of limitations of pLSA: over-fitting to the training set
- Tempered EM
  - Reduce the effect of fitting as we do more EM steps
- Split training data into training set and validation set
- Conduct EM on the training set until the performance on the validation set decreases
- Decrease β in the E-step from 1 to ηβ, where η <1

$$P(z \mid d, w) = \frac{[P(d \mid z)P(w \mid z)P(z)]^{\beta}}{\sum_{z'} [P(d \mid z')P(w \mid z')P(z')]^{\beta}}$$

- As long as the performance on the validation set increases, continue EM with the β
- If the performance decreases, decreas β again.
- Stop if no more performance improvement is achieved even decreases β

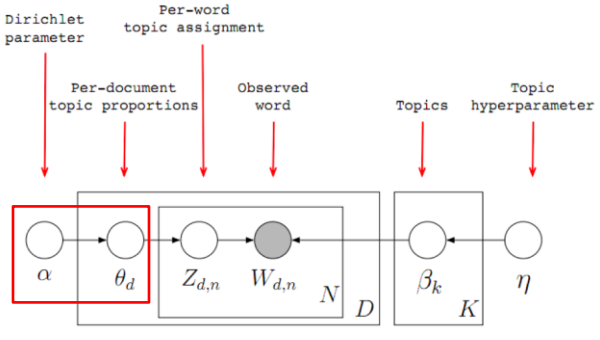
A Glimpse of LDA

- pLSA



No generation of the topic distribution per document from the corpus.

- LDA



this part is missing from pLSA

Thank you.

- Any question?

- Reference
  - Thomas Hoffman. Probabilistic Latent Semantic Analysis. UAI-99, 1999.
  - Thomas Hofmann. Probabilistic Latent Semantic Indexing. SIGIR-99, 1999.
  - David Cohn and Huan Chang. Learning to probabilistically identify Authoritative documents
  - David Cohn and Thomas Hoffman. The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity
  - David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. JMLR, 2003.
  - Slides form Shuguang