Topic Modeling

Andrei Shadrikov

MSU, CMC
Department of Computational Methods of Forecasting

Contents

Matrix Factorization

Given a matrix $Z = \|z_{ij}\|_{n\times m},\ \ (i,j) \in \Omega \subseteq \{1..n\} \times \{1..m\}$

Find matrices $X = \|x_{it}\|_{n\times k}$ and $Y = \|y_{tj}\|_{k\times m}$ such that

$$\big\|Z - XY\big\|_{\Omega,d} = \sum_{(i,j)\in\Omega} d\Big(z_{ij}, \sum_t x_{it} y_{tj}\Big) \to \min_{X,Y}$$

Variety of problems:

- loss function:
  - quadratic: $d(z,\hat{z}) = (z-\hat{z})^2$,
  - Kullback–Leibler: $d(z,\hat{z}) = z\ln(z/\hat{z}) - z + \hat{z}$
- nonnegative matrix factorization: $x_{it} \geqslant 0,\ \ y_{tj} \geqslant 0$
- stochastic matrix factorization: $\sum\limits_i x_{it} = 1,\ \ \sum\limits_t y_{tj} = 1$
- sparse input data: $|\Omega| \ll nm$
- sparse output factorization X, Y

Example applications of Matrix Factorization

1. **Feature Extraction for Image Recognition**

$$z_{ij} = \sum_k w_{ik} h_{kj}$$

given: z_{ij} — set of images;
find: w_{ik} — matrix of basis parts (parts, features);
h_{kj} — matrix of coefficients

2. **The measurement of the expression levels of genes in DNA microarray with cross-hybridization**

$$z_{pk} = \sum_g a_{pg} c_{gk}$$

given: z_{pk} — intensity of probe p on microarray k;
find: a_{pg} — binding affinity of probe p for gene g;
c_{gk} — concentration of gene g on microarray k.

Example applications of Matrix Factorization

3. **Revealing latent interests in recommender system (collaborative filtering)**

$$z_{iu} = \sum_t p_{it} q_{tu}$$

given: z_{iu} — item i rating by user u;
find: p_{it} — interests profile of item i;
q_{tu} — interests profile of a user i.

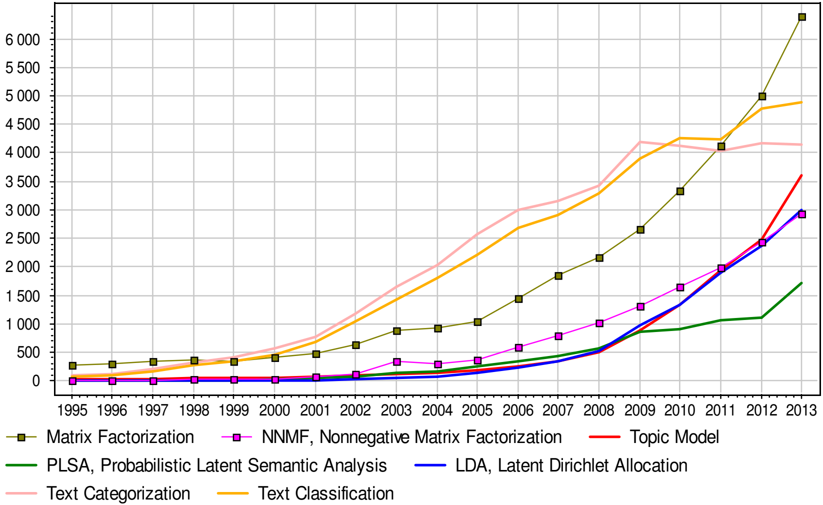
4. **Revealing latent topics in text collection (topic modeling)**

$$z_{wd} = \sum_t \phi_{wt} \theta_{td}$$

given: $z_{wd} = p(w|d)$ — word probabilities for document d;
find: $\phi_{wt} = p(w|t)$ — word probabilities for topic t,
$\theta_{td} = p(t|d)$ — topic probabilities for document d.

Matrix Factorization and Topic Modeling research areas

Google Scholar citation counts



Probabilistic Topic Model (PTM)

W — vocabulary of terms (words or phrases)
D — collection of text documents $d = \big(w_1, \ldots, w_{n_d}\big)$

Assumptions:

- each term in each document refers to some latent topic $t \in T$
- $D \times W \times T$ — discrete probability space, $|T| \ll |D|, |W|$
- $(d_i, w_i, t_i)_{i=1}^{n} \sim p(d, w, t)$ — text collection as an i.i.d. sample
- d_i, w_i are observable, topics t_i are hidden
- $p(w|d, t) = p(w|t)$ — conditional independence assumption

Generative topic model for a text collection:

$$p(w|d) = \sum_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$$

- $\phi_{wt} \equiv p(w|t)$ — distribution over terms for topic t;
- $\theta_{td} \equiv p(t|d)$ — distribution over topics for document d;

Goals and applications of topic modeling

Goals:

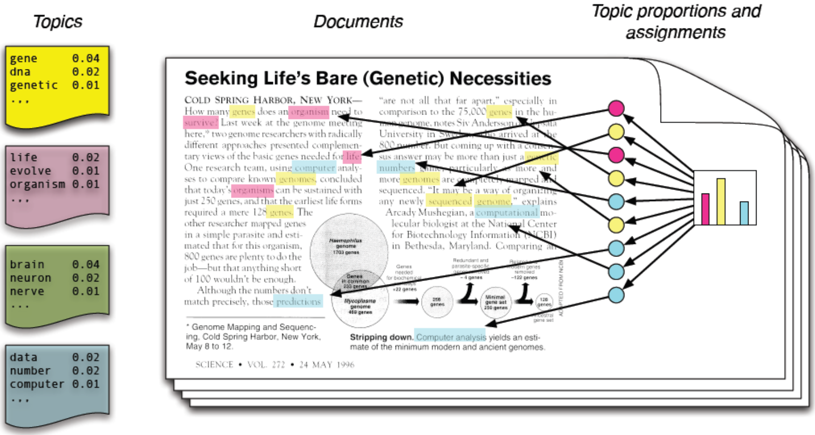
- Uncover a hidden thematic structure of the text collection
- Find a highly compressed representation of each document by a set of its topics

Applications:

- Information retrieval for long-text queries
- Categorization, classification, summarization, segmentation of texts, images, video, signals
- Semantic search in large scientific documents collections
- Revealing research trends and research fronts
- Expert search
- News aggregation
- Recommender systems
- etc...

Direct problem: PTM → document collection

Document $d = (w_1, \ldots, w_{n_d})$ is generated from $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$



Inverse problem: document collection → PTM

Given a document collection:
n_{dw} — how many times term w appears in document d
$\hat{p}(w|d) \equiv \frac{n_{dw}}{n_d}$ — conditional term frequency

Find stochastic matrix factorization

$$\hat{p}(w|d) \approx \sum_{t \in T} \phi_{wt}\theta_{td}$$

or in matrix notation

$$\underset{W \times D}{Z} \approx \underset{W \times T}{\Phi} \cdot \underset{T \times D}{\Theta}$$

$Z = \left\|\hat{p}(w|d)\right\|_{W \times D}$ — known frequency matrix,
$\Phi = \left\|\phi_{wt}\right\|_{W \times T}$ — term–topic matrix, $\phi_{wt} = p(w|t)$,
$\Theta = \left\|\theta_{td}\right\|_{T \times D}$ — topic–document matrix, $\theta_{td} = p(t|d)$.

Matrix Update Rule

Popular iteration methods can be written as:

Input: matrix Z, # of topics $|T|$, # of iterations $i_{\max}$;
Output: matrices Φ and Θ;

initialize Φ_{wt}, Θ_{td} for all w, t, d;
forall the *iterations* $i = 1, \ldots, i_{\max}$ **do**
  $\Phi_{ik}^{new} = F(\Phi^{old}, \Theta^{old})$;
  $\Theta_{kj}^{new} = G(\Phi^{old}, \Theta^{old})$;

Example of Iteration Methods

1. **PLSA — Probabilistic Latent Semantic Analysis [Hoffman, 1999]**

$$n_{dwt} = Z_{ij} \frac{\Phi_{wt}\Theta_{td}}{\sum_{t \in T} \Phi_{ws}\Theta_{sd}}$$

n_{dwt} — counts the number of triples (d, w, t) in D

$$\Phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d \in D} n_{dwt}}{\sum\limits_{d \in D}\sum\limits_{w \in d} n_{dwt}}, \qquad \Theta_{td} = \frac{n_{td}}{n_d} \equiv \frac{\sum\limits_{w \in d} n_{dwt}}{\sum\limits_{w \in W}\sum\limits_{t \in T} n_{dwt}},$$

Short notation via proportionality sign $\propto$:

$$\Phi_{wt} \propto n_{wt}; \qquad \Theta_{td} \propto n_{td};$$

Example of Iteration Methods

2. **MU — Gradient Descent with Multiplicative Update Rule [Lee, Seung, 2001]**

$$\Phi_{wt} = \Phi_{wt} \frac{(Z\Theta^T)_{wt}}{(\Phi\Theta\Theta^T)_{wt}}, \qquad \Theta_{td} = \Theta_{td} \frac{(\Phi^T Z)_{td}}{(\Phi^T \Phi\Theta)_{td}}$$

3. **ALS — Alternating Least Squares [Paatero, Tapper, 1994]**

$$\Phi \leftarrow [solve\, \Theta\Theta^T \Phi^T = \Theta Z^T]_+$$

$$\Theta \leftarrow [solve\, \Phi^T \Phi\Theta = \Phi^T Z]_+$$

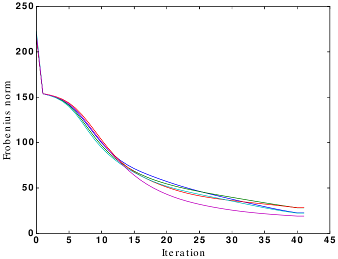
NMF to SMF

How can we use NMF methods in Topic Modeling?

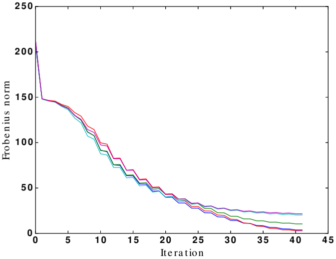
Projection: $Pr_{U_n} v = \frac{1}{\|v\|} v$
U_n — normalization constraints.
v — result of NMF method.

With normalization methods can be used in topic modeling.

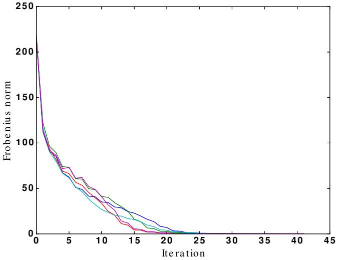
MU with normalization



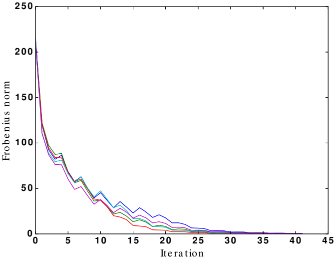
MU + PLSA



MU + ALS (normalized)



MU + ALS (normalized) + PLSA



Problems

For now we have issues like:

- Many local optima, algorithms stuck.
- Slow convergence.
- Not interpretable results.
- ...

Futher Discussion

- Some problems can be solved using regularization:

$$\min_{\Phi, \Theta} D\left(Z - \Phi\Theta\right) + R(\Phi, \Theta)$$

- Can algorithms be paralleled?
- Can there be unique solution?

Bibliography

- *Hofmann T.* Probabilistic Latent Semantic Indexing. SIGIR, 1999.
- *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.
- *Lee D., Seung S.* Algorithms fot nonnegative matrix factorization.
- *Vorontsov K. V.* Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. Pleiades Publisher, 2014. Vol. 88, No. 3.

Questions?