Dimensionality reduction in text mining.

Victor Kitov

v.v.kitov@yandex.ru

Common dimensionality reduction methods

- LSA
- non-negative matrix factorization
- pLSA
- LDA
- more advanced topic models

LSA

- LSA=latent semanyic analysis
  - also called latent semantic indexing or LSI
- SVD decomposition: $X = U\Sigma V^T$, $U^T U = I$, $V^T V = I$, $\Sigma = \text{diag}\left\{\sigma_1^2, ...\sigma_R^2\right\}$, $U, V \in \mathbb{R}^{NxD}$, $\Sigma \in \mathbb{R}^{DxD}$, $R = \text{rg}\, X$
- $\widehat{X}_K = U_K \Sigma_K V_K^T$, U_K, V_K-first K columns of U,V; Σ_K-first K columns&rows of Σ
- $U_K, V_K \in \mathbb{R}^{NxK}$, $\Sigma_K \in \mathbb{R}^{KxK}$, $K \le R$, usually $K \in [200, 500]$.
- $\widehat{X}_K = \arg\min_{B: rg\, B \le K} \|X - B\|_{Fr}^2$
- $U = XV\Sigma^{-1}$ => for new $x \in \mathbb{R}^{1xD}$:$u = xV\Sigma^{-1}$ (*folding in* of new observations).

pLSA[1]

- pLSA = probabilistic latent semantic analysis
- probabilistic generative model for words in documents
  - words in replica
  - genes in DNA sequences
  - other properties in property sequences
- Each document is associated some distribution on topics $z \sim p(z|d)$
- Each topic is associated a distribution on words $w \sim p(w|z)$

[1] Thomas Hofmann, Probabilistic Latent Semantic Indexing, SIGIR-99, 1999.

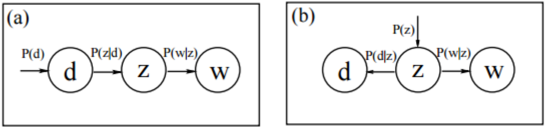
pLSA generation

- For each word position:
  - Document is sampled with $p(d)$
  - Exact topic $z \sim p(z|d)$ is sampled.
  - Exact word $w \sim p(w|z)$ is sampled on currect word position.

$$p(d,w) = p(d)p(w|d) = p(d)\sum_z p(z|d)p(w|z) \quad (1)$$

$$= \sum_z p(d,z)p(w|z) = \sum_z p(z)p(d|z)p(w|z) \quad (2)$$

graphical representation for pLSA: asymmetric (a) and symmetric (b)



Connection of pLSA to LSA

- In matrix form $X = U\Sigma V^T$, where
  - $X \in \mathbb{R}^{DxW}$, $U \in \mathbb{R}^{DxK}$, $\Sigma \in \mathbb{R}^{KxK}$, $V \in WxK$
  - U, V - are stochastic, not orthogonal matrices
  - U, Σ, V are estimated with maximum likelihood, not Frobenius norm minimization.
- pLSA - more interpretable
  - document-topics distribution
  - topic-word distribution
  - We can truncate this representation by taking only topics with $p(z) \geq threshold$.
  - allows finding semantically close words and documents
  - segmentation into topics of running text

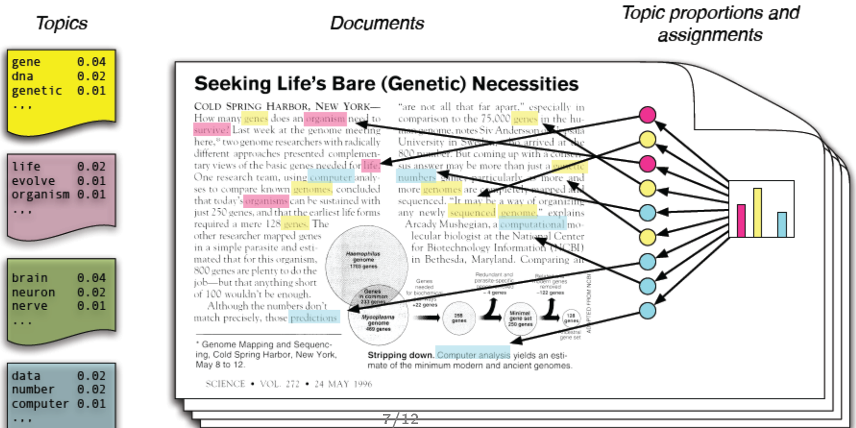
Dimensionality reduction with pLSA

- Define $x_{dw} := p(w|d), \quad a_{dz} := p(z|d), \quad b_{zw} := p(w|z)$
- $X = \{x_{dw}\} \in \mathbb{R}^{DxW}, \quad A = \{a_{dz}\} \in \mathbb{R}^{DxK}, \quad B = \{b_{zw}\} \in \mathbb{R}^{KxW}$
- $p(w|d) = \sum_z p(z|d)p(w|z)$
- In matrix form $X = AB$
- $a_{d,:} \in \mathbb{R}^K$-low dimensional representation of document d
- $b_{:,w} \in \mathbb{R}^K$-low dimensional representation of word w
- Allows to find similar/dissimilar documents and words.

Segmentation into topics of running text

Label words with

$$\arg\max_z p(z|d,w) = \arg\max_z \frac{p(z,d,w)}{p(d,w)} = \arg\max_z p(z)p(d|z)p(w|z)$$



Probabilistic model with latent variables

Suppose objects have observed features x and unobserved (latent) features z[2].

- $[x, z] \sim p(x, z, \theta)$, $x \sim p(x, \theta)$
- denote $X = [x_1, x_2, ...x_N]$, $Z = [z_1, z_2, ...z_N]$.

To find $\widehat{\theta}$ we need to solve

$$L(\theta) = \ln p(X|\theta) = \ln \sum_{Z} p(X, Z|\theta) \to \max_{\theta}$$

- This is intractable for unknown Z.
- We need to fallback to iterative optimization, such as SGD.
- Alternatively, we may use EM algorithm, which "averages" over different fixed variants of Z.

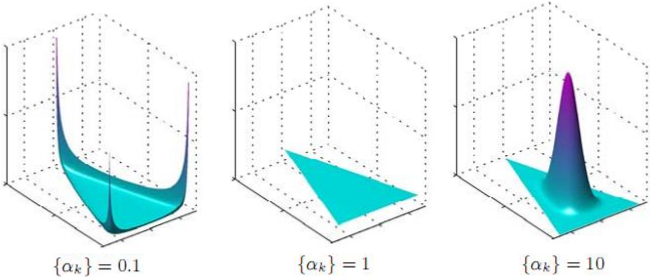
[2] They are considered discrete here. Everything holds true for continious latent variables if everywhere you replace summation over Z with integration

LDA method

- Bayesian extension of pLSA
- Distributions $p(z|d)$ and $p(w|z)$ are «inner random parameters» with prior distributions:

$$p(z|d) \sim Dir(\alpha), \quad p(w|z) \sim Dir(\beta)$$

Probability density function of Dirichlet(α), $\alpha = \{\alpha_k\}_{k=1}^{K}$



LDA variables

Parameters:

- α-Dirichlet prior on topics distributions $p(z|d)$
- β-Dirichlet prior on words distributions $p(w|z)$

Estimated values:

- $\varphi_z = p(w|z)$, $w = \overline{1, W}$, $z = \overline{1, Z}$
- $\theta_d = p(z|d)$, $z = \overline{1, Z}$, $d = \overline{1, D}$

Latent variables:

- topics at each word-position:

$$z_i^d, \quad d = \overline{1, D}, \; i = \overline{1, n_d}$$

Observed variables:

- words at each word-position:

$$w_i^d, \quad d = \overline{1, D}, \; i = \overline{1, n_d}$$

LDA-data generation process

1. generate $\theta_d \sim Dir(\alpha), \quad d = \overline{1, D}$
2. generate $\varphi_z \sim Dir(\beta), \quad z = \overline{1, Z}$
3. for each document d and each word-position $n = \overline{1, n_d}$:
    1. generate topic $z_n^d \sim Multinomial(\theta_d)$
    2. generate word $w_n^d \sim Multinomial(\varphi_{z_n^d})$

Extensions of topic models

- Automatically select number of topics (e.g. HDP)
  - still need to specify «willingless to make new topic»
- hierarchical set of topics
  - greedy layerwise optimization
  - joint optimization for whole hierarchy
- incorporate other rich text information:
  - authors, images, links, titles etc.