

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 semantic analysis
 - also called latent semantic indexing or LSI
- SVD decomposition: $X = U\Sigma V^T$, $U^T U = I$, $V^T V = I$,
 $\Sigma = \text{diag} \{ \sigma_1^2, \dots, \sigma_R^2 \}$, $U, V \in \mathbb{R}^{N \times D}$, $\Sigma \in \mathbb{R}^{D \times D}$, $R = \text{rg } X$
- $\hat{X}_K = U_K \Sigma_K V_K^T$, U_K, V_K -first K columns of U, V ; Σ_K -first K columns&rows of Σ
- $U_K, V_K \in \mathbb{R}^{N \times K}$, $\Sigma_K \in \mathbb{R}^{K \times K}$, $K \leq R$, usually $K \in [200, 500]$.
- $\hat{X}_K = \arg \min_{B: \text{rg } B \leq K} \|X - B\|_{Fr}^2$
- $U = X V \Sigma^{-1} \Rightarrow$ for new $x \in \mathbb{R}^{1 \times D}$: $u = x V \Sigma^{-1}$ (folding in of new observations).

pLSA¹

- 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 $w \sim p(w|z)$

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

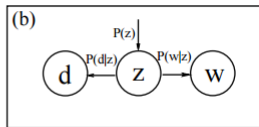
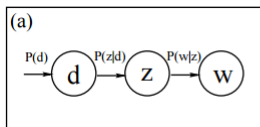
pLSA generation

- 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 current 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
 - $X \in \mathbb{R}^{D \times W}$, $U \in \mathbb{R}^{D \times K}$, $\Sigma \in \mathbb{R}^{K \times K}$, $V \in W \times K$
 - 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 $p(z) \geq \text{threshold}$.
 - allows finding semantically close words and documents
 - segmentation into topics of running text

Dimensionality reduction with pLSA

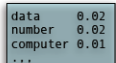
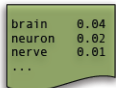
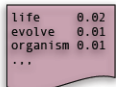
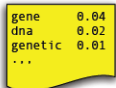
- Define $x_{dw} := p(w|d)$, $a_{dz} := p(z|d)$, $b_{zw} := p(w|z)$
- $X = \{x_{dw}\} \in \mathbb{R}^{D \times W}$, $A = \{a_{dz}\} \in \mathbb{R}^{D \times K}$, $B = \{b_{zw}\} \in \mathbb{R}^{K \times W}$
- $p(w|d) = \sum_z p(z|d)p(w|z)$
- In matrix form $X = AB$
- $a_{d,:} \in \mathbb{R}^K$ -low dimensional 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

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

Topics



Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those **predictions**

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

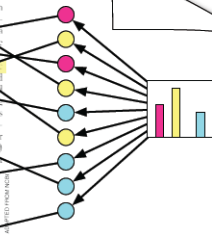
SCIENCE • VOL. 272 • 24 MAY 1996

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a geneticist at the University in Sweden, who arrived at the 800 number. But coming up with a **compact** answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely sequenced and analyzed. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Topic proportions and assignments



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 $X = [x_1, x_2, \dots, x_N]$, $Z = [z_1, z_2, \dots, z_N]$.

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

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

- 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 .

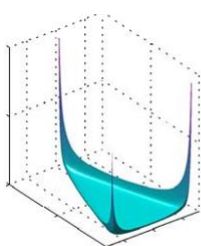
²They are considered discrete here. Everything holds true for continuous latent variables if everywhere you replace summation over Z with integration

LDA method

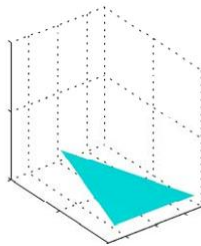
- Bayesian extension of pLSA
- Distributions $p(z|d)$ and $p(w|z)$ are «inner random parameters» with prior distributions:

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

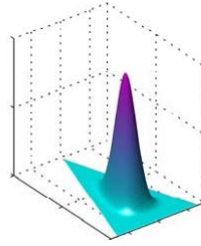
Probability density function of Dirichlet(α), $\alpha = \{\alpha_k\}_{k=1}^K$



$$\{\alpha_k\} = 0.1$$



$$\{\alpha_k\} = 1$$



$$\{\alpha_k\} = 10$$

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

Extensions of topic models

- Automatically select number of topics (e.g. HDP)
 - still need to specify «willingness 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.