Additive Regularization for Probabilistic Topic Modeling

Konstantin Vorontsov

MIPT • CC RAS • HSE • MSU • Yandex

15 May 2014 PreMoLab

• Advances in Optimization and Statistics •

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

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

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: $\sum_i x_{it} = 1$, $\sum_t y_{tj} = 1$
- sparse input data: $|\Omega| \ll nm$
- sparse output factorization X, Y

Example applications of Matrix Factorization

1. **Separation of a mixture of chemical substances in High Performance Liquid Chromatography**

$$z_{t\lambda} = \sum_i x_{ti} y_{i\lambda}$$

given: $z_{t\lambda}$ — output of a scanning ultraviolet detector;
find: x_{ti} — chromatogram of i-th substance, t — time;
$y_{i\lambda}$ — spectrum of i-th substance, λ — wavelength.

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.

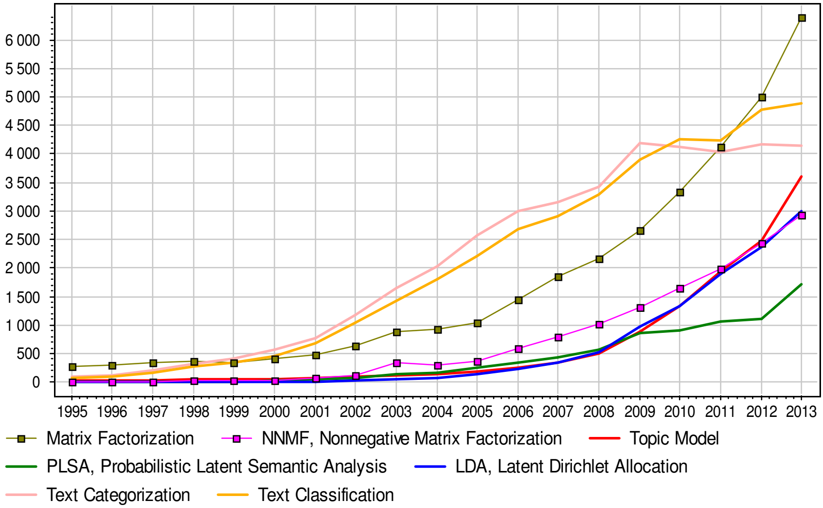
4. Revealing latent topics in text collection (topic modeling)

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

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



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...

Probabilistic Topic Model (PTM)

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

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

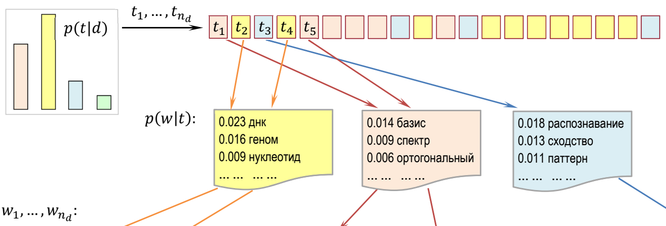
Generative topic model for a text collection:

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

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

Direct problem: PTM $\to$ document collection

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



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

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, $\theta_{td} = p(t|d)$.

PLSA — Probabilistic Latent Semantic Analysis [Hofmann 1999]

Likelihood maximization: $\ln \prod_{d,w} p(d,w)^{n_{dw}} \to \max_{\Phi,\Theta}$

The problem of log-likelihood maximization:

$$\mathscr{L}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta},$$

under non-negativeness and normalization restrictions

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t\in T} \theta_{td} = 1$$

$\Longleftrightarrow$ minimize a weighted sum of KL-divergences:

$$\sum_{d\in D} n_d \underbrace{\sum_{w\in d} \hat{p}(w|d) \ln \frac{\hat{p}(w|d)}{p(w|d)}}_{\mathrm{KL}(\hat{p}\|p)} \;\to\; \min_{\Phi,\Theta},$$

EM-algorithm for likelihood maximization

Theorem

Maximum of $\mathscr{L}(\Phi,\Theta)$ satisfies the system of equations with basic variables ϕ_{wt}, θ_{td} and auxiliary variables p_{tdw}, n_{wt}, n_{td}

$$\begin{array}{ll} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\ \phi_{wt} = \dfrac{n_{wt}}{\sum_{w'} n_{w't}}; \quad n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw}; \\ \theta_{td} = \dfrac{n_{td}}{\sum_{t'} n_{t'd}}; \quad n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw}; \end{array} \right. \\ \text{M-step:} & \end{array}$$

EM-algorithm alternates E-step and M-step until convergence.

It is a simple iteration method for solving this system of equations [Hofmann 1999], [Asuncion 2009].

Probabilistic interpretation of E-step and M-step

E-step is equivalent to Bayes' formula:

$$p_{tdw} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}$$

$n_{dwt} = n_{dw}\,p(t|d,w)$ counts the number of triples (d,w,t) in D

M-step is a frequency estimation of conditional probabilities:

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

Short notation via proportionality sign $\propto$:

$$\phi_{wt} \propto n_{wt}; \qquad \theta_{td} \propto n_{td};$$

The efficient implementation of EM-algorithm

The idea is to incorporate E-step into M-step. No 3D-arrays!

Input: collection D, num. of topics $|T|$, num. of iterations $i_{\max}$;
Output: matrices Φ and Θ;

```
initialize φ_wt, θ_td for all d ∈ D, w ∈ W, t ∈ T;
for all iterations i = 1, ..., i_max
    n_wt, n_td, n_t, n_d := 0 for all d ∈ D, w ∈ W, t ∈ T;
    for all documents d ∈ D and terms w ∈ d
        p_tdw = φ_wt θ_td / Σ_s φ_ws θ_sd  for all t ∈ T;
        n_wt, n_td, n_t, n_d += n_dw p_tdw  for all t ∈ T;
    φ_wt := n_wt / n_t  for all w ∈ W, t ∈ T;
    θ_td := n_td / n_d  for all d ∈ D, t ∈ T;
```

Usually $i_{\max} = 20..50$ iterations are sufficient. Time is $O(n|T|i_{\max})$.

LDA — Latent Dirichlet Allocation [Blei 2003]

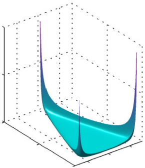
Assumption. Column vectors $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ are generated from Dirichlet distributions, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod\limits_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \; \beta_t \geqslant 0;$$

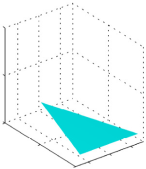
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \; \alpha_t \geqslant 0;$$

Example:

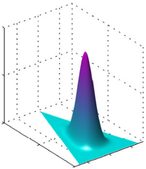
$\mathrm{Dir}(\theta|\alpha)$
$|T| = 3$
$\theta, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$



$\alpha_1 = \alpha_2 = \alpha_3 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = 10$

The main difference between LDA and PLSA

The estimates of conditionals $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- in PLSA — unbiased maximum likelihood estimates:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- in LDA — smoothed Bayesian estimates:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

The difference is only significant for small n_{wt}, n_{td}.
Robust LDA and robust PLSA produce almost identical models.

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.

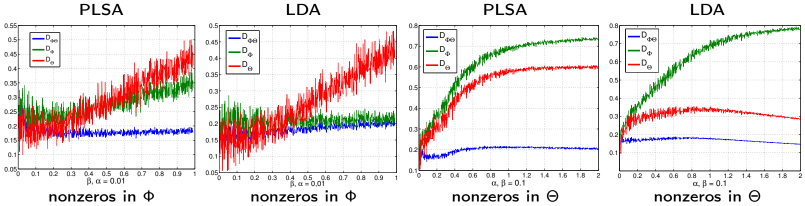
Potapenko A. A., Vorontsov K. V. Robust PLSA Performs Better Than LDA. ECIR-2013, Moscow, Russia, 24-27 March 2013. LNCS, Springer. Pp. 784–787.

Topic Modeling as an ill-posed inverse problem

The *nonuniqueness* and *instability* of matrix factorization:
$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$ for all S such that Φ', Θ' are stochastic.

Experiment: recovering known Φ, Θ on model dataset,
$|D| = 500, \;\; |W| = 1000, \;\; |T| = 30, \;\; n_d \in [100, 600]$.

Result: product $\Phi\Theta$ is always recovered well, however
matrix Φ and matrix Θ are recovered if being highly sparse only:



Conclusion: regularization is needed to ensure uniqueness!

Additive Regularization of Topic Model

Suppose that along with the likelihood we want to maximize n more criteria $R_i(\Phi, \Theta)$, $i = 1, \dots, n$ called *regularizers*.

Scalarization is a standard technique for multi-criteria optimization.

The problem of regularized log-likelihood maximization:

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

under non-negativeness and normalization restrictions

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1$$

where $\tau_i > 0$ are *regularization coefficients*.

ARTM: EM-algorithm with regularized M-step

Theorem

The maximum of $\mathscr{L}(\Phi, \Theta) + R(\Phi, \Theta)$ satisfies the system of equations with auxiliary variables p_{tdw}, n_{wt}, n_{td}

$$\begin{aligned} &\text{E-step:} \quad && \begin{cases} p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\ \phi_{wt} \propto \left(n_{wt} + \phi_{wt} \dfrac{\partial R}{\partial \phi_{wt}} \right)_+ ; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \\ \theta_{td} \propto \left(n_{td} + \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} \right)_+ ; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \end{cases} \\ &\text{M-step:} && \end{aligned}$$

where $(x)_+ \overset{df}{=} \max(x, 0)$ is a positive cutoff.

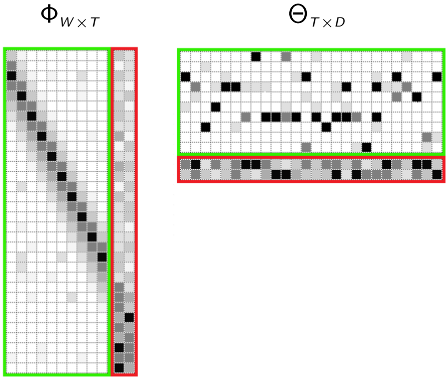
PLSA: $R(\Phi, \Theta) = 0$

LDA: $R(\Phi, \Theta) = \sum\limits_{t,w} \beta_w \ln \phi_{wt} + \sum\limits_{d,t} \alpha_t \ln \theta_{td}$

Assumptions: what topics would be well-interpretable?

Topics $S \subset T$ contain domain-specific terms
$p(w|t)$, $t \in S$ are sparse and different (weakly correlated)

Topics $B \subset T$ contain background terms
$p(w|t)$, $t \in B$ are not sparse and contain common lexis words



Smoothing regularization (rethinking LDA)

The non-sparsity assumption for background topics $t \in B$:
ϕ_{wt} are similar to a given distribution β_w;
θ_{td} are similar to a given distribution α_t.

$$\sum_{t \in B} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d \in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

We maximize the sum of these KL-divergences to get a regularizer:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td} \to \max.$$

The regularized M-step applied for all $t \in B$ coincides with LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_0 \alpha_t,$$

which is new non-Bayesian interpretation of LDA [Blei 2003].

Sparsing regularizer (further rethinking LDA)

The sparsity assumption for domain-specific topics $t \in S$: distributions ϕ_{wt}, θ_{td} contain many zero probabilities.

We minimize the sum of KL-divergences $\mathrm{KL}(\beta \| \phi_t)$ and $\mathrm{KL}(\alpha \| \theta_d)$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \to \max .$$

The regularized M-step gives "anti-LDA", for all $t \in S$:

$$\phi_{wt} \propto \big(n_{wt} - \beta_0 \beta_w\big)_+, \qquad \theta_{td} \propto \big(n_{td} - \alpha_0 \alpha_t\big)_+ .$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Regularization for topics decorrelation

The dissimilarity assumption for domain-specific topics $t \in S$: if topics are interpretable then they must differ significantly.

We maximize covariances between column vectors ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \backslash t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max .$$

The regularized M-step makes rows of Φ more distant:

$$\phi_{wt} \propto \Bigl(n_{wt} - \tau \phi_{wt} \sum_{s \in S \backslash t} \phi_{ws} \Bigr)_+ .$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Regularization for topic selection

Assumption: infrequent topics can not be well-interpretable.

We maximize KL-divergence $\mathrm{KL}\big(\frac{1}{|T|} \,\|\, p(t)\big)$ to make distribution over topics $p(t) = \sum_d p(d)\theta_{td}$ sparse:

$$R(\Theta) = -\tau \sum_{t\in S} \ln \sum_{d\in D} p(d)\theta_{td} \to \max .$$

The regularized M-step formula results in Θ rows sparsing:

$$\theta_{td} \propto \left(n_{td} - \tau \frac{n_d}{n_t}\theta_{td}\right)_+ .$$

Effect:
if n_t is small then all values in the t-th row could turn into zeros.

Regularization for topic coherence maximization

Assumption: if topic is well-interpretable then its top words are *coherent* i. e. frequently appear nearby in the documents.

$C_{uw} = \hat{p}(w|u) = \frac{N_{uw}}{N_u}$ — *coherence* of a word pair $u, w \in W$,
N_u, N_{uw} are document frequency of word w and word pair u, w.

Bring together ϕ_{wt} and its coherent words estimate $\hat{p}(w|t)$:

$$\hat{p}(w|t) = \sum_u \hat{p}(w|u) p(u|t) = \frac{1}{n_t} \sum_u C_{uw} n_{ut};$$

$$R(\Phi, \Theta) = \tau \sum_{t \in T} n_t \sum_{w \in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

The regularized M-step gives a kind of smoothing:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} n_{ut}.$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Regularization for semi-supervised learning

Assumption: experts have provided us with topic labeling data:

- each document $d \in D_0 \subseteq D$ belongs to a subset of topics $T_d \subset T$;
- each topic $t \in T_0 \subseteq T$ contains a subset of words $W_t \subset W$.

ϕ^0_{wt} — uniform distribution over subset of terms W_t
θ^0_{td} — uniform distribution over subset of topics T_d

We minimize the sum of KL-divergences $\mathrm{KL}(\phi^0_t \| \phi_t)$ and $\mathrm{KL}(\theta^0_t \| \theta_t)$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi^0_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \ln \theta_{td} \to \max .$$

The regularized M-step results in LDA-like smoothing:

$$\phi_{wt} \propto n_{wt} + \beta_0 \phi^0_{wt} \qquad \theta_{td} \propto n_{td} + \alpha_0 \theta^0_{td}$$

Nigam K., McCallum A., Thrun S., Mitchell T. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, no. 2–3.

Experiment with additive combinations of regularizers

The goal of the experiment:
Can we improve interpretability without a loss of the likelihood?

The set of regularizers:

- smoothing background topics — Φ columns, Θ rows
- sparsing domain-specific topics — Φ columns, Θ rows
- decorrelation of domain-specific topics — Φ columns
- topic selection — Θ rows

Dataset: NIPS (Neural Information Processing System)

- $|D| = 1566$ papers from NIPS conference;
- collection length $n \approx 2.3 \cdot 10^6$,
- vocabulary size $|W| \approx 1.3 \cdot 10^4$.
- testing collection length: $|D'| = 174$.

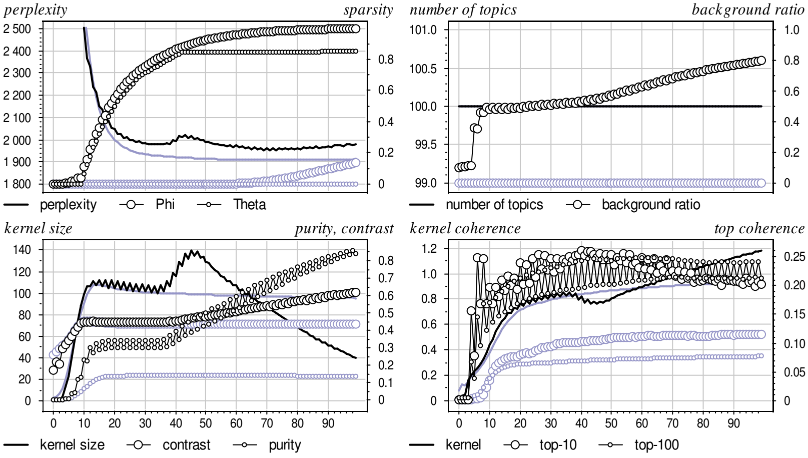
Topic model quality measures

Multi-criteria optimization requires multiple quality measures.

- Hold-out *perplexity*: $\mathcal{P} = \exp(-\frac{1}{n}\mathscr{L})$
- *Sparsity* — the number of zero elements in Φ and Θ
- Interpretability measures for each topic t:
  - topic *coherence* [Newman, 2010]
  - topic *kernel size*: $|W_t|$, kernel $W_t \stackrel{df}{=} \{w\colon p(t|w) > 0.25\}$
  - topic *purity*: $\sum\limits_{w\in W_t} p(w|t)$
  - topic *contrast*: $\frac{1}{|W_t|} \sum\limits_{w\in W_t} p(t|w)$
- Model degeneracy:
  - number of non-zero topics: $|T|$
  - the fraction of background words: $\frac{1}{n} \sum\limits_{d\in D} \sum\limits_{w\in d} \sum\limits_{t\in B} p(t|d,w)$

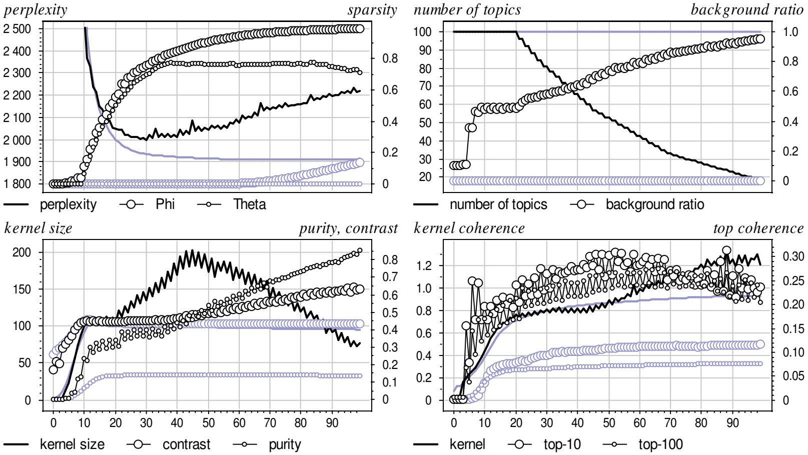
Sparsing + Smoothing + Decorrelation

Quality measures as functions of the iteration step
(grey lines — PLSA, black lines — ARTM)



Sparsing + Smoothing + Decorrelation + Topic Selection

Quality measures as functions of the iteration step
(grey lines — PLSA, black lines — ARTM)



Conclusions from experiments

ARTM provides a multi-objective model improvement:

- *sparsity* augments from 0 to 95%–98%
- *coherence* augments from 0.1 to 0.3
- *purity* augments from 0.15 to 0.8
- *contrast* augments from 0.4 to 0.6
- *kernel size* augments from 0 to 150 terms
- almost without any loss of the *perplexity*

Recommendations for choosing regularization path:

- turn *sparsing on* gradually after first 10-20 iterations
- turn *topic selection on* after turning on sparsing
- turn *sparsing off* as soon as kernel size begins to decrease
- turn *background smoothing on* from the beginning
- turn *decorrelation on* as much as possible from the beginning
- make *topic selection* and *decorrelation* at different iterations

Variety of regularizers for ARTM

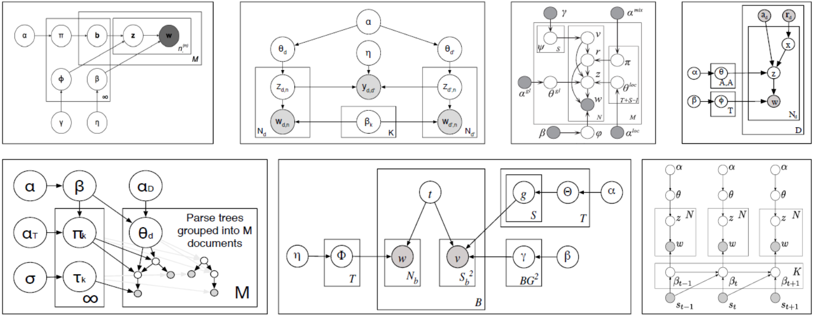
Understood and implemented:

1. smoothing
2. sparsing
3. topic decorrelation
4. topic selection

Understood but not implemented yet:

5. semi-supervised learning
6. coherence maximization
7. using links or cites between documents
8. using document categories or classes
9. using time-stamped data
10. ...

Graphical Models and Bayesian Inference



Topic Modeling Bibliography: http://mimno.infosci.cornell.edu/topics.html

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.
Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(русский перевод на www.MachineLearning.ru)

ARTM vs. Bayesian Inference

Bayesian Inference for Topic Modeling

1. Fully probabilistic generative model of data
2. Dirichlet distribution plays a central role in the theory
3. Complicated maths for combined and multi-objective models
4. High barrier to entry into PTMs research field

Additive Regularization for Topic Modeling

1. Semi-probabilistic approach
2. No Dirichlet prior, no integration, no graphical models
3. Simple maths for combined and multi-objective models
4. Very short way from an idea to the algorithm

Conclusions on ARTM approach

ARTM advantages:

- ARTM is much simpler that Bayesian Inference
- ARTM focuses on formalizing task-specific requirements
- ARTM simplifies the multi-objective PTMs learning
- ARTM reduces barriers to entry into PTMs research field
- ARTM encourages the development of regularization library

ARTM restrictions:

- Choosing a regularization path is a new open issue for PTMs

Further research work:

- More linguistically motivated regularizations
- BigARTM — open source project for Large-Scale Parallel Distributed Multi-Objective Topic Modeling

Bibliography

- *Hofmann T.* Probabilistic Latent Semantic Indexing. SIGIR, 1999.
- *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.
- *Teh Y. W., Newman D., Welling M.* A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. NIPS, 2006, Pp. 1353–1360.
- *Porteous I., Newman D., Ihler A., Asuncion A., Smyth P., Welling M.* Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation. KDD 2008.
- *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.
- *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic Evaluation of Topic Coherence // Human Language Technologies, HLT-2010, Pp. 100–108.
- *Yi Wang.* Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.
- *Sato I., Nakagawa H.* Rethinking Collapsed Variational Bayes Inference for LDA. Int'l Conf. on Machine Learning ICML, 2012.
- *Vorontsov K. V.* Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. Pleiades Publisher, 2014. Vol. 88, No. 3.
- *Vorontsov K. V., Potapenko A. A.*, Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'14. Springer. 2014. (to appear)

Vorontsov Konstantin
voron@forecsys.ru

Wiki www.MachineLearning.ru (in Russian):

- User:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование

LDA. Принцип максимума апостериорной вероятности

$$\ln \prod_{d\in D} \prod_{w\in d} p(d,w)^{n_{dw}} \prod_{t\in T} \mathrm{Dir}(\phi_t;\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d;\alpha) \to \max_{\Phi,\Theta}$$

Задача максимизации регуляризованного правдоподобия:

$$\mathscr{L}'(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\,[\phi_{wt}>0] \ln \phi_{wt} + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\,[\theta_{td}>0] \ln \theta_{td} \;\to\; \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t\in T} \theta_{td} = 1$$

EM-алгоритм для максимизации апостериорной вероятности

Теорема

Максимум $\mathscr{L}'(\Phi, \Theta)$ удовлетворяет системе уравнений со вспомогательными переменными p_{tdw}, n_{wt}, n_{td},

$$
\begin{cases}
\text{E-шаг:} & p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\
\text{M-шаг:} & \phi_{wt} = \dfrac{\left(n_{wt} + \beta_w - 1\right)_+}{\sum_{w'} \left(n_{w't} + \beta_{w'} - 1\right)_+}; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \\
 & \theta_{td} = \dfrac{\left(n_{td} + \alpha_t - 1\right)_+}{\sum_{t'} \left(n_{t'd} + \alpha_{t'} - 1\right)_+}; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw};
\end{cases}
$$

где $(x)_+ = \max(x, 0)$ — операция положительной срезки.

EM-алгоритм — это чередование E- и M-шага до сходимости.
Это метод простых итераций для решения системы уравнений.

Доказательство Теоремы о регуляризации М-шага

1. Условия ККТ для ϕ_{wt}:

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geqslant 0; \quad \lambda_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Учтём ограничение $\phi_{wt} \geqslant 0$ и предположение $\lambda_t > 0$:

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

4. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w\in W} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

5. Подставим λ_t из (4) в (3), получим требуемое. ■

Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_{x}; \\ g_i(x) \leqslant 0, \quad i = 1, \dots, m; \\ h_j(x) = 0, \quad j = 1, \dots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \dots, m$, λ_j, $j = 1, \dots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j g_j(x); \\ g_i(x) \leqslant 0; \; h_j(x) = 0; \; \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \; \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \; \text{(условие дополняющей нежёсткости)} \end{cases}$$

Дивергенция Кульбака–Лейблера

Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

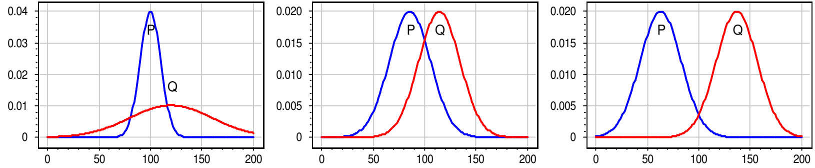
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}.$$

1. $\mathrm{KL}(P\|Q) \geqslant 0$; $\quad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q$;

2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_{\alpha} \quad \Longleftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_{\alpha}.$$

3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$ $\quad$ $\mathrm{KL}(P\|Q) = 0.444$ $\quad$ $\mathrm{KL}(P\|Q) = 2.969$

$\mathrm{KL}(Q\|P) = 2.966$ $\quad$ $\mathrm{KL}(Q\|P) = 0.444$ $\quad$ $\mathrm{KL}(Q\|P) = 2.969$

Проблема $\ln 0$ в дивергенции Кульбака–Лейблера

Почему в регуляризаторе разреживания

$$R(\Phi) = -\beta_0 \sum_{t\in S} \sum_{w\in W} \beta_w \ln \phi_{wt} \to \max$$

не возникает проблем с $\ln \phi_{wt}$ при $\phi_{wt} \to 0$?

Подправим регуляризатор, при сколь угодно малом ε:

$$R(\Phi) = -\beta_0 \sum_{t\in S} \sum_{w\in W} \beta_w \ln(\phi_{wt} + \varepsilon) \to \max$$

Подставив в формулу М-шага, получим для всех $t \in S$:

$$\phi_{wt} \propto \left(n_{wt} - \beta_0 \beta_w \tfrac{\phi_{wt}}{\phi_{wt}+\varepsilon}\right)_+$$

Если $\phi_{wt} = 0$, то разреживания не будет, и оно уже не нужно.

Регуляризатор для учёта связей между документами

Гипотеза: чем больше n_{dc} — число ссылок из d на c, тем более близки тематики документов d и c.

Минимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Phi, \Theta) = \tau \sum_{d,c \in D} n_{dc} \mathrm{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} \propto n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Регуляризатор для классификации документов

Пусть C — множество классов документов
(категории, авторы, ссылки, годы, пользователи,...)

Гипотеза:
классификация документа d объясняется его темами:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}.$$

Минимизируем дивергенцию между моделью $p(c|d)$
и «эмпирической частотой» классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct}\theta_{td} \to \max.$$

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

Регуляризатор для классификация документов

EM-алгоритм дополняется оцениванием параметров ψ_{ct}.

E-шаг. По формуле Байеса:

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T} \phi_{ws}\theta_{sd}} \qquad p(t|d,c) = \frac{\psi_{ct}\theta_{td}}{\sum\limits_{s\in T} \psi_{cs}\theta_{sd}}$$

M-шаг. Максимизация регуляризованного правдоподобия:

$$\phi_{wt} \propto n_{wt} \qquad n_{wt} = \sum_{d\in D} n_{dw}\, p(t|d,w)$$

$$\theta_{td} \propto n_{td} + \tau m_{td} \qquad n_{td} = \sum_{w\in W} n_{dw}\, p(t|d,w) \qquad m_{td} = \sum_{c\in C} m_{dc}\, p(t|d,c)$$

$$\psi_{ct} \propto m_{ct} \qquad m_{ct} = \sum_{d\in D} m_{dc}\, p(t|d,c)$$

Регуляризатор для категоризации документов

Снова регуляризатор для классификации:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \to \max$$

Недостаток: для «эмпирической частоты классов» приходится необоснованно брать равномерное распределение:

$$m_{dc} = n_d \frac{1}{|C_d|} [c \in C_d]$$

Ковариационный регуляризатор:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \sum_{t \in T} \psi_{ct} \theta_{td} \to \max$$

приводит к естественному аналитическому решению

$$\psi_{ct} = \big[c = c^*(t)\big], \quad c^*(t) = \arg\max_{c \in C} \sum_{d \in D} m_{dc} \theta_{td}$$

Эффект: Каждая категория c распадается на свои темы.

Регуляризаторы для динамической тематической модели

Y — моменты времени (например, годы публикаций),
$y(d)$ — метка времени документа d,
$D_y \subset D$ — все документы, относящиеся к моменту $y \in Y$.

Гипотеза 1: распределение $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ разрежено:

$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \sum_{t \in T} \ln p(t|y) \to \max.$$

Эффект — разреживание тем t с малым $p(t|y(d))$:

$$\theta_{td} \propto \left(n_{td} - \tau_1 \frac{\theta_{td} p(d)}{p(t|y(d))} \right)_+.$$

Гипотеза 2: $p(t|y)$ меняются плавно, с редкими скачками:

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} \bigl| p(t|y) - p(t|y{-}1) \bigr| \to \max.$$