Robust PLSA Performs Better Than LDA


Anna Potapenko, Konstantin Vorontsov

potapenko@forecsys.ru, voron@forecsys.ru

MOSCOW STATE UNIVERSITY
COMPUTATIONAL MATHEMATICS AND CYBERNETICS

MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

FORECSYS
Forecasting & Recognition Systems Center

Problems

1. Modern Probabilistic Topic Models (PTM) theory is too complicated to make a model fast, robust, sparse, online, hierarchical, semi-supervised, multilingual, etc. all at once.
2. Dirichlet prior in Latent Dirichlet Allocation (LDA) is mathematically convenient but poorly motivated linguistically.
3. Sparsity vs. smoothness contradiction.

Contributions

1. We modify the learning algorithm, rather than the underlying generative model, to combine robustness, smoothing, sampling, sparsing, and online semi-supervised learning.
2. We show that LDA improves estimates for rare and new terms (which are useless for topics), rather than prevents overfitting.
3. We show that robustness and sparsity are more effective than Dirichlet smoothing.

Formulation

Given:
D — a set of documents,
W — a vocabulary of terms,
n_{dw} — frequency data for all $(d, w) \in D \times W$,
n — collection size in terms.

Find:
$p(w \mid d)$ — a *topic model* depending on:
$\phi_{wt} \equiv p(w \mid t)$ — terms for each topic $t \in T$,
$\theta_{td} \equiv p(t \mid d)$ — topics for each document $d \in D$.

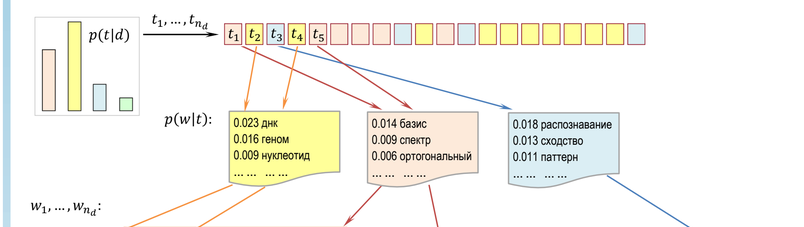
Likelihood maximization:

$$\mathscr{L}(\Theta, \Phi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w \mid d) \;\to\; \max_{\Theta, \Phi}$$

Hypotheses:
1. bag of terms, bag of documents
2. conditional independence: $p(w \mid t, d) = p(w \mid t)$
3. distributions $p(w \mid t)$, $p(t \mid d)$ are sparse
4. not all terms in a document are topical

Models

Generative probabilistic topic model:



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).

PLSA (Probabilistic Latent Semantic Analysis):

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

Robust PLSA: noise π_{dw} and background π'_w

$$p(w \mid d) = \frac{1}{1+\gamma+\varepsilon} \left(\sum_{t \in T} \phi_{wt} \theta_{td} + \gamma \pi_{dw} + \varepsilon \pi'_w \right)$$

EM-algorithm

1: **repeat**
2: **E-step:**
estimate distribution over topics for all (d, w) using Bayes' theorem:
$p(t \mid d, w) := \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$;
3: **M-step:**
count terms belonging to topics:
$n_{wt} := \sum\limits_{d \in D} n_{dw} p(t \mid d, w)$;
$n_t := \sum\limits_{w \in W} n_{wt}$;
count documents belonging to topics:
$n_{dt} := \sum\limits_{w \in d} n_{dw} p(t \mid d, w)$;
$n_d := \sum\limits_{t \in T} n_{dt}$;
estimate conditionals as frequencies:
$\phi_{wt} := \dfrac{n_{wt}}{n_t}$;
$\theta_{td} := \dfrac{n_{dt}}{n_d}$;
4: **until** Φ, Θ converge.

Smoothing

Dirichlet prior leads to smoothed frequency estimations for conditional probabilities:

$$\phi_{wt} := \frac{n_{wt} + \beta_w}{n_t + \beta_0}; \quad \theta_{td} := \frac{n_{dt} + \alpha_t}{n_d + \alpha_0}.$$

Sampling

Stochastic EM-algorithm:
generate topics t_{dwi}, $i = 1, \dots, s$ randomly from $p(t \mid d, w)$, then use empirical distribution $\hat{p}(t \mid d, w) = \frac{1}{s} \sum\limits_{i=1}^{s} \left[t_{dwi} = t\right]$ instead of $p(t \mid d, w)$.

Gibbs sampling is a special case at $s = n_{dw}$.

Fixing $s = 3..5$ works faster and gives sparse $\hat{p}(t \mid d, w)$ without significant loss of quality.

Sparsing

Forced sparsing: set the smallest 5% of probabilities θ_{td}, ϕ_{wt} to zero at the end of each $(10 + 2k)$-th iteration, $k = 1, 2, \dots$

Iteration steps reordering

1. Rational EM-algorithm:
M-step is a pass through the collection D.
E-step is performed inside the M-step.

2. Generalized EM-algorithm:
Updating ϕ_{wt}, θ_{td} more frequently than after each pass through D accelerates convergence.

3. Online EM-algorithm:
E-step and θ_{td} updates are combined together to form per-document iterations.
M-step updates ϕ_{wt} after each pass through D.
(The analog is Online-LDA in Vowpal Wabbit)

Semi-supervised learning

Labeled data:
$p_0(w \mid t)$ — terms labeled by topics;
$p_0(t \mid d)$ — documents labeled by topics;

$$\phi_{wt} := \lambda p_0(w \mid t) + (1 - \lambda) \frac{n_{wt}}{n_t};$$

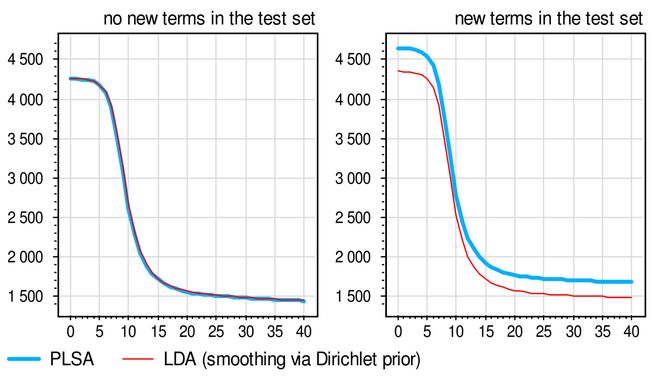
$$\theta_{td} := \mu p_0(t \mid d) + (1 - \mu) \frac{n_{dt}}{n_d}.$$

Results

Data set: $|D| = 2000$ Russian-language synopses of theses ($n = 8.7 \cdot 10^6$, $|W| = 3 \cdot 10^4$ terms). Graphs below show the hold-out perplexity by iteration.

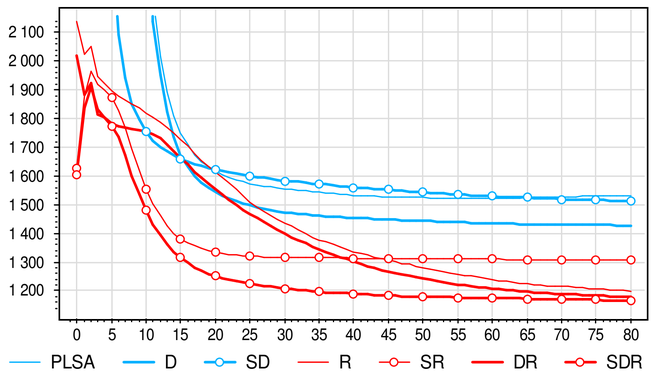
Perplexity $= \exp\left(-\frac{1}{n}\mathscr{L}(\Theta, \Phi)\right)$.

Smoothing (Dirichlet prior) reduces perplexity only if there are new terms in a test set.



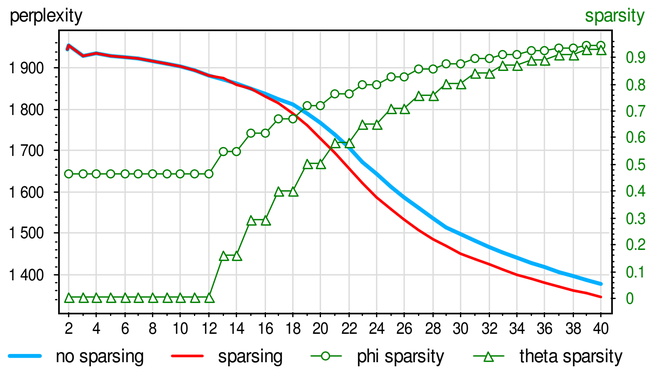
Robustness reduces the hold-out perplexity more effectively than Dirichlet smoothing does.

Sampling accelerates convergence.



D–Dirichlet prior, S–sampling, R–robustness

Forced sparsing of small probabilities ϕ_{wt}, θ_{td} gives about 95% of zeros without loss of quality.



Explanation: sparsing may improve perplexity due to the sparse nature of data itself.

References

[1] *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование*, 2012, Т. 4, № 4, С. 693–706.

Future directions

Developing a well-interpretable, semi-supervised, hierarchical, fine-grained, multilingual, temporal topic model and a fast, online, parallelized, distributed learning algorithm for it.

Developing a search engine for academic publications.