Robust PLSA Performs Better Than LDA

Anna Potapenko[1] and Konstantin Vorontsov[2]

[1] MSU, CMC, Moscow, Russia
anya_potapenko@mail.ru
[2] MIPT, Moscow, Russia, CC RAS, Moscow, Russia
voron@forecsys.ru

Abstract. In this paper we introduce a generalized learning algorithm for probabilistic topic models (PTM). Many known and new algorithms for PLSA, LDA, and SWB models can be obtained as its special cases by choosing a subset of the following "options": regularization, sampling, update frequency, sparsing and robustness. We show that a robust topic model, which distinguishes specific, background and topic terms, doesn't need Dirichlet regularization and provides controllably sparse solution.

Keywords: topic modeling, Gibbs sampling, perplexity, robustness.

1 Generalized Learning Algorithm for PTMs

Topic modeling is a rapidly developing application of machine learning to text analysis. A topic model of a text corpus determines what terms characterize each topic and what topics are associated with each document. Each document d from a text corpus D is a sequence of terms $(w_1, \ldots, w_{n_d})$ from a vocabulary W, where n_d is the length of the document. Let n_{dw} denote the number of term w occurrences in document d. According to probabilistic topic models PLSA [4] and LDA [2] a finite set of latent topics T exists and each document $d \in D$ is a set of terms, drawn independently from the following distribution:

$$p(w \,|\, d) = \sum_{t \in T} \phi_{wt} \theta_{td}, \tag{1}$$

where $\phi_{wt} \equiv p(w \,|\, t)$ and $\theta_{td} \equiv p(t \,|\, d)$ are discrete distributions to be found.

In *Probabilistic Latent Semantic Analysis* (PLSA) parameters of the model $\Phi = (\phi_{wt})_{W \times T}$ and $\Theta = (\theta_{td})_{T \times D}$ are estimated through likelihood maximization, given non-negativity and normality constraints for vectors ϕ_t and θ_d:

$$L(\Theta, \Phi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Theta, \Phi}. \tag{2}$$

In *Latent Dirichlet Allocation* (LDA) parameters are assumed to be drawn from a prior Dirichlet distribution: $\theta_d \sim \mathrm{Dir}(\alpha)$, $\alpha \in \mathbb{R}^T$, $\phi_t \sim \mathrm{Dir}(\beta)$, $\beta \in \mathbb{R}^W$ helping to reduce overfitting [2]. Although PLSA and LDA have different generative models, the differences between their learning algorithms are not so significant [1]. Both of them use an iterative process originating from the EM-algorithm. Each

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 784–787, 2013.

iteration is a linear pass through the corpus. For each document–term pair (d, w) current values of ϕ_{wt}, θ_{td} are used to estimate discrete distribution over topics $H_{dwt} = p(t \mid d, w)$ from Bayes' theorem; then vice-versa conditional probabilities ϕ_{wt}, θ_{td} are estimated from counters $n_{dwt} = n_{dw} H_{dwt}$:

$$H_{dwt} = \phi_{wt}\theta_{td}\Big(\sum_s \phi_{ws}\theta_{sd}\Big)^{-1}. \tag{3}$$

$$\begin{aligned} \phi_{wt} &= \big(\hat{n}_{wt} + \beta_w\big)\big(\hat{n}_t + \sum_u \beta_u\big)^{-1}, \quad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt}, \quad \hat{n}_{wt} = \sum_{d \in D} n_{dwt}; \\ \theta_{td} &= \big(\hat{n}_{dt} + \alpha_t\big)\big(\hat{n}_d + \sum_s \alpha_s\big)^{-1}, \quad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}, \qquad \hat{n}_{dt} = \sum_{w \in d} n_{dwt}. \end{aligned} \tag{4}$$

We propose a set of mutually compatible "options" for this iterative process that being combined give a variety of learning algorithms for PTMs.

1. *Dirichlet regularization* with fixed [5] or optimized [6] smoothing parameters α_t, β_w gives LDA. Turning them off ($\alpha_t = 0$, $\beta_w = 0$) gives PLSA.

2. *Sampling* uses Monte-Carlo estimate $\hat{p}(t \mid d, w)$ instead of $p(t \mid d, w)$. Sampling n_{dw} topics for each pair (d, w) gives Gibbs Sampling (GS) algorithm [5]. However this contradicts the *H-sparsity hypothesis* "each term in the document is typically associated with one topic". Our experiments show that *reduced sampling* of $s = 1, \ldots, 5$ topics per pair (d, w) makes the algorithm more speed- and memory-efficient without significant loss of quality.

3. *Frequent update* of ϕ and θ parameters per each of n_{dw} occurrences of a term is used in GS. Rare update per iteration is used in original PLSA [4] and in recent collapsed GS and variational Bayesian (VB) algorithms including highly competitive CVB0 algorithm. We also tested per-k-terms and per-document update strategies. Our experiments show that the increase of frequency speeds up convergence but does not influence the model quality. Per-occurrence update used in GS is too intensive; per-term update seems to be optimal.

4. *Sparsing* heuristic follows the hypotheses of *Θ-sparsity*: "a document typically refers to a few topics" and *Φ-sparsity*: "a topic is typically characterized by a small part of terms". We perform sparsing by setting to zero the fraction σ of the smallest probabilities θ_{td} for each d and the smallest probabilities ϕ_{wt} for each t at the end of each i-th iteration if $i > i_0$ and i is divisible by k. The parameters σ, k, i_0 provide a way to trade off sparsity against quality.

2 Robust PLSA and LDA Topic Models

Robust PTM named *Specific Words and Background* (SWB) [3] introduces a very realistic assumption that each document d can be represented by a mixture of topic terms distribution (1), now rewritten as Z_{dw}, *noise terms* distribution $\pi_{dw} \equiv p_{\mathrm{n}}(w \mid d)$ that models specific aspects of the document, and *background terms* distribution $\pi_w \equiv p_{\mathrm{b}}(w)$ that models common aspects of the whole corpus:

$$p(w \mid d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1 + \gamma + \varepsilon}, \qquad Z_{dw} = \sum_{t \in T} \phi_{wt}\theta_{td},$$

Algorithm 1. Robust PLSA and LDA learning algorithm.

1: $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, n_{dwt}, ν_{dw}, ν_d, ν, ν'_{dw}, ν'_w, $\nu' := 0$; $\pi_{dw} := n_{dw}/n_d$; $\pi_w := n_w/n$;
2: **repeat**
3: **for all** $d \in D$, $w \in d$ **do**
4: **if** not first pass through the corpus **then**
5: update ϕ_{wt}, θ_{td} for all $t \in T$ according to (4);
6: $\pi_w := \nu'_w/\nu'$; $\pi_{dw} := (n_{dw}/\nu_d - Z_{dw}/\gamma - \varepsilon\pi_w/\gamma)_+$;
7: $Z := Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w$;
8: $\delta := n_{dw}\phi_{wt}\theta_{td}/Z$; increase $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ by $(\delta - n_{dwt})$; $n_{dwt} := \delta$; $\forall t \in T$;
9: $\delta := n_{dw}\gamma\pi_{dw}/Z$; increase ν_d, ν by $(\delta - \nu_{dw})$; $\nu_{dw} := \delta$;
10: $\delta := n_{dw}\varepsilon\pi_w/Z$; increase ν'_w, ν' by $(\delta - \nu'_{dw})$; $\nu'_{dw} := \delta$;
11: **until** Φ, Θ, Π converge.

where π_{dw} and π_w are unknown distributions, γ and ε are given fixed parameters. A modified LDA-GS has been proposed in [3] to train the SWB model. We use our generalized PTM learner to combine robustness with other options not shown in the sketch Algorithm 1 because of volume limitation. Note that step 6 uses a maximum likelihood estimate for π_{dw} as opposed to recurrent formulas in [3].

3 Experiments and Conclusions

To evaluate PTMs the *hold-out perplexity* is commonly used:

$$\mathscr{P}(D') = \exp\left(- \sum_{d \in D'} \sum_{w \in d''} n_{dw} \ln p(w \mid d) \Big/ \sum_{d \in D'} \sum_{w \in d''} n_{dw} \right),$$

where each test document d from the document set D' is randomly divided into two halves d' and d''. The parameters ϕ_{wt} and π_w are learned from the training set D. The document-related parameters θ_{td} and π_{dw} are learned from d'. Then the perplexity is computed using the second halves d'' of the test documents.

We use two different datasets. The NIPS corpus is standard. The RuDis corpus contains 2000 Russian-language synopses of theses of the total length about $8.7 \cdot 10^6$ and the vocabulary size about $3 \cdot 10^4$ after lemmatization and stop-words removal. The test set contains $|D'| = 200$ documents for both corpora.

The parameters are as follows: number of topics $|T| = 100$; Dirichlet prior for LDA models: $\alpha_t = 0.5$, $\beta_w = 0.01$; robustness parameters: $\gamma = 0.3$, $\varepsilon = 0.1$.

Fig. 1–3 represent the graphs of $\mathscr{P}(D')$ from the number of iterations.

Fig. 1 shows that PLSA and LDA perform almost identically if the test set doesn't contain the terms that haven't occurred in the training set. Thus LDA does not reduce overfitting but only describes the probability of new terms better. However robust models describe new terms even more accurately, see Fig. 3.

Sparsing may deteriorate PLSA and LDA models, which are not intrinsically sparse. Robust models are more suitable for sparsing due to the compensative role of the noise component π_{dw}. Sparsing with $\sigma = 0.05$, $i_0 = 10$, $k = 2$ gives about 90% of zeros in Φ and Θ matrixes with no loss of quality, see Fig. 2.

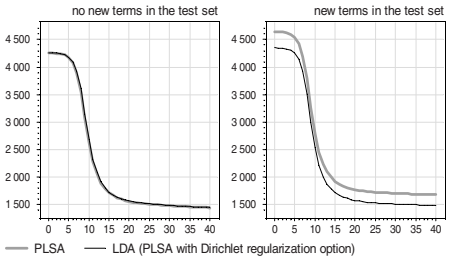

Fig. 1. Regularization has an advantage if only there are new terms in a test set

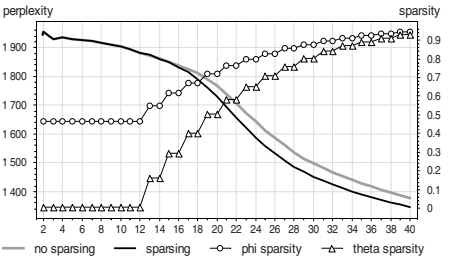

Fig. 2. Sparsing Φ and Θ up to 90% of zero values does not worsen perplexity (RuDis)

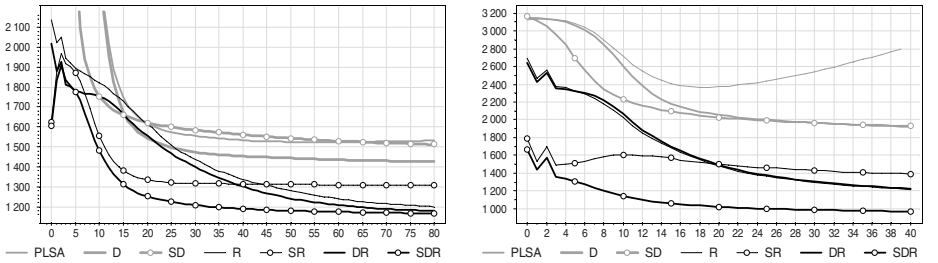

Fig. 3. Robustness reduces the hold-out perplexity more effectively than regularization does (options: D–Dirichlet prior, S–sampling, R–robustness; left: RuDis, right: NIPS)

The most surprising result is that robust models perform well without Dirichlet prior, see Fig. 3. Robust PLSA gives a better hold-out perplexity than non robust LDA. Robustness, sparsing and reduced sampling together make PTMs learning algorithms more scalable to large text collections.

Acknowledgments. The work is supported by the Ministry of Education and Science of the Russian Federation, State Contract 07.524.11.4002.

References

1. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Int'l Conf. on Uncertainty in Artificial Intelligence (2009)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling general and specific aspects of documents with a probabilistic topic model. In: Advances in Neural Information Processing Systems, vol. 19, pp. 241–248. MIT Press (2006)
4. Hofmann, T.: Probabilistic latent semantic indexing. In: 22nd Int'l Conf. SIGIR, pp. 50–57. ACM (1999)
5. Steyvers, M., Griffiths, T.: Finding scientific topics. In: Proceedings of the National Academy of Sciences, vol. 101(suppl. 1), pp. 5228–5235 (2004)
6. Wallach, H., Mimno, D., McCallum, A.: Rethinking LDA: Why priors matter. In: Advances in Neural Information Processing Systems, pp. 1973–1981 (2009)