3B1-R-2-1

A Distance Between Text Documents based on Topic Models and Ground Metric Learning

JIN Tao
Graduate School of Informatics
Kyoto University
jintao@iip.ist.i.kyoto-u.ac.jp

Marco Cuturi
Graduate School of Informatics
Kyoto University
mcuturi@i.kyoto-u.ac.jp

Akihiro Yamamoto
Graduate School of Informatics
Kyoto University
akihiro@i.kyoto-u.ac.jp

We propose a new distance between text documents that builds upon two techniques. We first represent each document in a database as a histogram of topics using the Latent Dirichlet Allocation (LDA) topic model. We then compare two documents by computing the transportation distance between their respective topic histograms. The transportation distance parameter, which is in that case a metric matrix between topics, is estimated using Ground Metric Learning. We carry out experiments on the 20-newsgroup text messages classification benchmark task that illustrate the interest of our approach.

1. Introduction

Text categorization is the task of classifying a set of documents into categories from a predefined set. Different approaches have been proposed in the past to classify text. We are particularly interested in this work in the bag of words representation [Salton et al., 1975] for text, in which a text is simply summarized as a histogram of words. This framework has been successful in a wide variety of settings [Manning et al., 1999, §15,16], using for instance kernel methods [Joachims, 2002].

We propose to address the problem of text categorization within the simple framework of k-nearest neighbor classification [Hastie et al., 2009, §2,§6] and thus focus on the definition of a flexible and adequate distance between text representations as histograms. We adopt in this work a two step approach: we represent each document in a corpus as a histogram of topics using Latent Dirichlet Allocation Blei et al. [2003]. We then measure the distance between two documents by choosing a distance within the family of transportation distances [Villani, 2003, §7]. Such a transportation distance between histograms of topics is parameterized by a single parameter, known as the ground metric, which can be in that case any metric matrix that describes the distance between the topics themselves . As is the case for all parameterized distances, transportation distance can only prove useful in practice when the topic metric parameter is carefully chosen. We propose to use a set of algorithms proposed recently by Cuturi and Avis [2011] to learn adaptively the ground metric.

This paper is organized as follows: we describe our approach in Section 2., followed by a short review of experimental results in Section 3. and conclude in Section 4.

2. Topic Metric Learning

2.1 Topic Models

Topic models [Blei and Lafferty, 2009, Blei, 2011] are graphical models which can be used to uncover the basic semantic structure of each document within a corpus. Topic models belong to the large family of dimensionality reduction techniques, such as principal component analysis, and are used in practice to represent texts as histograms of abstract topics rather than histogram of words. Since the number of topics is usually set to be much smaller than the size of the set of all words, representing a text as a histogram of topics generally results in a much lighter representation.

We consider in this work the Latent Dirichlet Allocation (LDA)[Blei et al., 2003] model, which is arguably one of the simplest topic models. The aim of LDA is to exhibit common topics in a collection of documents, and describe precisely the weight of each of these topics within each document. A topic is itself a discrete distribution over a fixed vocabulary (or dictionary of terms). Although more advanced topic modeling techniques exist, such as the Chinese restaurant process [Blei et al., 2010], we will consider in this work that the number of topics is fixed beforehand. The initial parameters of the LDA model are: K, the total number of topics; α, the parameter of a Dirichlet distribution with K components; β, a collection of K multinomial distribution over the fixed vocabulary. Given these parameters, the LDA model assumes that each text in a corpus has been generated iteratively following a hierarchical procedure:

1. Choose randomly a document specific multinomial distribution θ over topics following a Dirichlet distribution with parameter α.
2. For each word in the document,
   (a) Choose randomly a topic i from the topic distribution θ
   (b) Choose randomly a word over the vocabulary randomly, following the distribution corresponding to that topic described in β

The topic structure, including the topics themselves; their distribution over the vocabulary; the distribution of topics

specific to each document are all hidden since the only information which is available are the texts themselves. The main problem for topic models is thus to use the observed documents to infer these hidden topic structure, and many algorithmic solutions have been proposed to sample the parameters α and β from a posterior probability as can be seen in [Blei, 2011, §2.2] and references therein.

2.2 Transportation Distances

The family of transportation distances can quantify a discrepancy between histograms. Consider two nonnegative histograms r and c of unit sum and dimension d, that is points in the simplex $\sum_{d-1} = \{u \in \mathbb{R}_+^d | \|u\|_1 = 1\}$. We represent r and c as column vectors $r = (r_1, \ldots, r_d)^T$ and $c = (c_1, \ldots, c_d)^T$. Suppose that we are given a $d \times d$ metric matrix M which quantifies a difference between the d bins of each histogram. A transportation map between r and c is a $d \times d$ nonnegative matrix whose row and column sums are equal to r and c respectively. This set of matrices, $\{X \in \mathbb{R}_+^{d \times d} \,|\, X\mathbf{1}_d = r, X^T\mathbf{1}_d = c\}$ is known as the transportation polytope $U(r, c)$. The cost of a transportation map is defined as the Frobenius dot product of X and M, $\langle X, M \rangle = \mathrm{trace}(XM)$. The transportation distance between r and c parameterized by M is defined as the minimal cost over all transportation maps in $U(r, c)$, namely

$$d_M(r, c) = \min_{X \in U(r,c)} \langle M, X \rangle$$

This quantity is the result of a linear program, more precisely a network flow problem Ahuja et al. [1993]. Transportation distances have been used in computer vision where they are popular under the name of *earth mover's distances* Rubner et al. [1997], see [Pele and Werman, 2009] for a recent survey.

We use transportation distances to quantify a difference between two text documents represented as topic histograms, where topic histograms have been generated by the LDA topic model. We can tune transportation distances between such histograms by selecting a metric matrix M, which is also known as the ground metric.

2.3 Ground Metric Learning

In their seminal work, Rubner et al. [2000] propose to set the ground metric using prior knowledge about the bins of each histogram. Cuturi and Avis [2011] have recently proposed an algorithm to estimate this parameter adaptively, using a labeled training set of histograms, in the same setting used by Mahalanobis metric learning algorithms Xing et al. [2003], Weinberger and Saul [2009], Davis et al. [2007]. Their approach can be applied on histogram data generated from arbitrary features, which fits our setting in which topics themselves are generated automatically by the LDA model. This is a key factor in our approach since transportation distances cannot work efficiently without setting this parameter.

3. Experiments

We consider the 20-Newsgroups dataset*1, which consists in a collection of approximately 19,997 newsgroup posts which have been collected evenly across 20 different newsgroups. The data is organized into 20 classes, each corresponding to a different discussion subject.

We used the the `Gensim` Python framework*2 to carry out simple preprocessing on all the documents by considering only lower case words, which are tokenized and stemmed before producing bag of words. We select randomly 30 documents from within each newsgroup as our training data. Fixing the number of topics to $k = 30$, we define a dictionary and train an LDA model with these documents using the `ldamodel` (100 passes) of `gensim` to estimate the Latent Dirichlet Allocation model parameters. Then we used the LDA model we obtained to infer topic distributions on all other documents in the database, including test documents which haven't been used to define the LDA model.

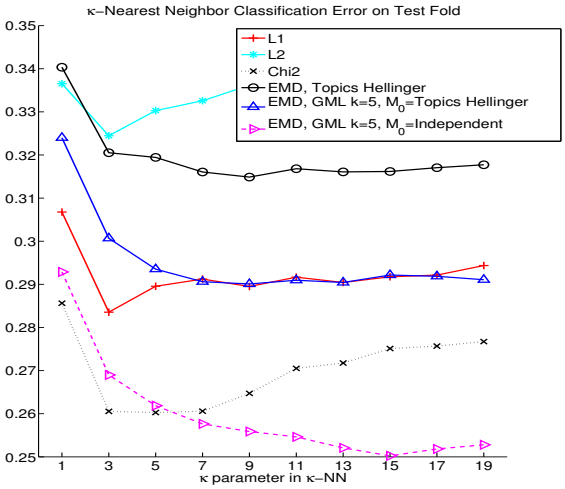

Figure 1: Average classification error in κnearest neighbor classification for pairs of newsgroup discussion subjects. The errors are averaged over 400 test points for each of the 190 considered binary classification tasks.

We then consider all the binary classification tasks that arise by comparing a newsgroup class i against another class j, $1 \leq i < j \leq 20$ namely $20 \times 19/2 = 190$ tasks. For each binary classification task (i, j) we choose randomly 200 test documents in class i and 200 more in class j, thus getting 400 test documents, and classify them using a κ-nearest neighbor classifier defined with the 30 train documents of class i and 30 documents of class j, summing thus to 60 train points.

We consider transportation distances with 3 different ground metrics: the Hellinger distance computed over topics seen as histogram of points; the metric obtained with ground metric learning setting the Hellinger distance as the

*1 `http://people.csail.mit.edu/jrennie/20Newsgroups/`
*2 `http://radimrehurek.com/gensim/`

initial point; the metric obtained with ground metric learning setting the Independence distance ([Cuturi and Avis, 2011, §5]) as the initial point. For the sake of comparison, we also consider the Total variation or Manhattan distance (L_1 norm of the difference of two histograms), the Euclidean distance (L_2 norm of the difference of two histograms) and the Chi-squared distance (L_2 norm of the difference of the component-wise squared roots of two histograms). The average classification error for all test points and all experiments (i, j) is represented in Figure 1 as a function of the neighborhood size κ.

In this figure, transportation distances with a topic metric learnt is shown to provide a competitive performance. The transportation distance seeded with a simple Hellinger distance between topics does not perform well, which underlines the importance of choosing carefully a good ground metric.

4. Conclusion

We have proposed in this paper a two stage approach to carry out text classification on documents seen as bags of words. Topic models can be used to reduce the dimensionality of text data and provide histogram representations. Transportation distances can be effective but need to be well tuned in order to be effective. We have shown that ground metric learning algorithms can be used in that context to compute a topic metric automatically and adaptively.

References

R. Ahuja, T. Magnanti, and J. Orlin. *Network Flows: Theory, Algorithms and Applications*. Prentice Hall, New Jersey, 1993.

D. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, pages 1–16, 2011.

D. Blei and J. Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10:71, 2009.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

D. Blei, T. Griffiths, and M. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2): 7, 2010.

M. Cuturi and D. Avis. Ground metric learning. *Arxiv preprint arXiv:1110.2306*, 2011.

J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216. ACM, 2007.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer Verlag, 2009.

T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer Academic Publishers, 2002.

C. Manning, H. Schütze, and MITCogNet. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009.

Y. Rubner, L. Guibas, and C. Tomasi. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA Image Understanding Workshop*, pages 661–668. Citeseer, 1997.

Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18 (11):613–620, 1975.

C. Villani. *Topics in Optimal Transportation*, volume 58. AMS Graduate Studies in Mathematics, 2003.

K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems*, pages 521–528, 2003.