CS229: Meeting Notes 5

October 30, 2013

Emaad Ahmed Manzoor

This week I did a survey of latent dirichlet allocation, its variants, and some applications of LDA. I've summarised the interesting points from my survey, and I've proposed 2 ideas for a graphical model similar to LDA that we can use for our classification problem.

1 Latent Dirichlet Allocation

Assume a fixed number of topics.

Each topic is a distribution over the vocabulary.

Generative Process. For each document:

– Randomly choose a distribution over topics for the document.

– For each word in the document:

  - Randomly choose a topic from the distribution of topics.
  - Randomly choose a word from this topic.

Properties of LDA. Some points I found important to keep in mind.

– All documents in the collection share the same set of topics. Each document exhibits these topics in different proportions.

– The order of the words is not important.

– The order of the documents is not important.

– Topics are mutually independent.

2 Variants of LDA

2.1 Supervised LDA[5]

This is the same as LDA, but it adds one additional step to the generative process. After generating all the words in the document, sample the response from a binomial (or normal, etc.) distribution.

2.2 Labelled LDA[3]

This introduces, in addition to the latent topics, a set of known topics called labels. A subset of the documents are labelled with these topics, providing some kind of weak supervision.

The goal of labelled LDA is, given the labels, to associate each word in the document with the most appropriate labels and vice versa.

For each labelled document, words are generated only from that document's labels. For unlabelled documents, words may be generated from any of the labels and the latent topics.

2.3 Dynamic, Spherical and Correlated Topic Models

Dynamic topic models assume that the order of documents is important.

Correlated topic models assume some dependencies between topics.

Spherical topic models assume that some words are unlikely to be part of certain topics.

3 LDA Applications

3.1 TopicSpam[1]

This approach tries to classify deceptive/fake hotel reviews. It assumes the following topics:

- Background topic B.
- Deceptive review topic D.
- Truthful review topic T.

– Hotel-specific topics, for each hotel H_j.

The interesting point is the background topic, which includes words that are not specific to hotels or deceptive or truthful reviews. Introducing this background topic boosted their accuracy from 88% to 94%.

3.2 The Author-Topic Model[6]

This approach aims to model the topics of interest for each author.

Each author is associated with a multinomial distribution of topics.

Each topic is a multinomial distribution over words.

A document with multiple authors is modeled as a distribution of topics that is a mixture of the distributions associated with the authors.

Generative process. For a document d written by authors a_d:

– For each word in the document

  - Choose an author uniformly at random.
  - Choose a topic from the distribution of topics for that author.
  - Choose a word from that topic.

3.3 Characterizing Microblogs with Topic Models[2]

This approach uses labelled LDA on Twitter tweets to understand the topics on which people tweet about. 200 latent topics are assumed, and 504 known topics (labels). The non-latent topics (labels) are derived from the tweet content where applicable; for example, hashtags are labels, so a tweet containing "#jobs" has the label "#jobs".

The interesting point here is that using known labels, you can derive the words that describe this label, and you can also interpret the meaning of this label when it appears as a topic on another document.

4 My Ideas

The broad goal is to come up with a generative model of Android applications that will help detect malware.

Each of these ideas can be evaluated against the baseline of LDA with 50 topics + SVM. The model can be evaluated on perplexity on a held-out set of apps.

Each of the ideas can be extended to jointly model the response (malware or non-malware), as in supervised LDA.

4.1 Idea 1

Each category is a Dirichlet distribution over app types.

Each app type as a multinomial distribution over permissions.

Generative process. For every permission in an app in a specific category, randomly choose an app type, and randomly choose if that permission occurs based on the multinomial distribution over permissions associated with the app type.

This is equivalent to training a topic model for every category. Once we have these per-category models, given an app and a category, we can calculate the probability of generating this app from the model for this category. For malware, we should expect to see a low probabilty.

4.2 Idea 2

This is based on the approach section 3.1. We apply plain LDA with a particular setting of topics.

We can assume each permission is generated from one of the following topics:

- M, the malware topic.
- G, the good apps topic (non-malware).
- C_i, a topic for each category i.
- B, the background topic (optional)

The priors for the Dirichlet distribution that generates these topics can be set based on the data.

We train the malware and non-malware separately. When training the malware, the permissions can be generated only from M, B and C_j. When training the non-malware, the permissions can be generated only from G, B and C_j.

On a new app, we mark each permission with the topic we believe it is generated from. We label it as malware/non-malware based on the proportion of permissions labelled as coming from the malware/non-malware topics.

References

1. Li, Jiwei, Claire Cardie, and Sujian Li.: TopicSpam: a Topic-Model-Based Approach for Spam Detection. Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics. 2013.
2. Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling.: Characterizing Microblogs with Topic Models. ICWSM. 2010.
3. Ramage, Daniel, et al.: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009.
4. Blei, David M.: Probabilistic topic models. Communications of the ACM 55.4 (2012): 77-84.
5. Blei, David M., and Jon D. McAuliffe.: Supervised topic models. arXiv preprint arXiv:1003.0783 (2010).
6. Rosen-Zvi, Michal, et al.: The author-topic model for authors and documents. Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004.