arXiv.org > stat > arXiv:1107.2462

Search or Article-id

(Help | Advanced search)

All papers

Go!

Statistics > Machine Learning

Statistical Topic Models for Multi-Label Document Classification

Timothy N. Rubin, America Chambers, Padhraic Smyth, Mark Steyvers

(Submitted on 13 Jul 2011)

Machine learning approaches to multi-label document classification have (to date) largely relied on discriminative modeling techniques such as support vector machines. A drawback of these approaches is that performance rapidly drops off as the total number of labels and the number of labels per document increase. This problem is amplified when the label frequencies exhibit the type of highly skewed distributions that are often observed in real-world datasets. In this paper we investigate a class of generative statistical topic models for multi-label documents that associate individual word tokens with different labels. We investigate the advantages of this approach relative to discriminative models, particularly with respect to classification problems involving large numbers of relatively rare labels. We compare the performance of generative and discriminative approaches on document labeling tasks ranging from datasets with several thousand labels to datasets with tens of labels. The experimental results indicate that generative models can achieve competitive multi-label classification performance compared to discriminative methods, and have advantages for datasets with many labels and skewed label frequencies.

Subjects: **Machine Learning (stat.ML)**; Learning (cs.LG)

Cite as: **arXiv:1107.2462v1 [stat.ML]**

Submission history

From: Timothy Rubin [view email]

[v1] Wed, 13 Jul 2011 04:28:32 GMT (958kb,D)

Which authors of this paper are endorsers?

Download:

- PDF
- Other formats

Current browse context:

stat.ML

< prev | next >

new | recent | 1107

Change to browse by:

cs
cs.LG
stat

References & Citations

- NASA ADS

Bookmark(what is this?)

Science WISE