Bayesian Clustering and Topic Discovery in Gene Expression Data

Karren Dai Yang, Skanda Koppula {KARREN, SKOPPULA}@MIT.EDU
Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Abstract

We report the results using various Bayesian models to analyze gene expression data. Our exploration includes parallelized LDA, mixture models, dynamic-time topic models, topic-clustering models, and non-parametric models, implemented in `numpy`, `C++`, and `Edward`. We evaluate our methods using held-out likelihood, posterior predictive checks, and biological meaningfulness testing.

1. Introduction

Tumors are composed of different sub-populations of cells, and these sub-populations often exhibit shared patterns of gene expression. With contemporary sequencing machines, it is possible to obtain the expression levels of 10,000+ genes for 1000+ cells in a single experiment.

In this project, we applied Bayesian methods for clustering and topic discovery to discover meaningful cell clusters and gene modules from single-cell RNA-sequencing (scRNA-seq) datasets from tumors. We first intend to explore the use of a hierarchical topic model, Latent Dirichlet Allocation (LDA) (Mitchell, 1980), as well as two non-parametric topic models, Hierarchical Dirichlet Processes (HDP) (Kearns, 1989) and the Indian Buffet Process Compound Dirichlet Process (IBP-CDP) (Langley, 2000) for analyzing scRNA-seq data. We will evaluate these models based on their ability to discover meaningful functional gene modules. Subsequently, we plan to explore the use of mixture models to cluster cells based on the output of the topic models, which we will evaluate based on their ability to differentiate cell types (e.g. immune, cancer, non-cancerous). Finally, we will train a combined clustering-topic model (e.g. MGCTM (Michalski et al., 1983)) to see if it outperforms the individual models. We will compare the results of these methods with standard methods used in prior work, and evaluate these methods' robustness to noise. Time permitting, we will study extensions of these models appropriate for time-series scRNA-seq data (Samuel, 1959).

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

2. Prior work

Prior research has applied basic measures of statistical distance to cluster cell groups or gene modules based on samples of single-cell RNA-sequencing (scRNA-seq) data taken from tumors (Borenszstein, 2017). Good clustering of scRNA-seq data often has biologically meaningful results, and as such, a wide variety of correlation based methods have been used in prior work to derive meaning from scRNA-seq data (Author, 2011; Duda et al., 2000; Newell & Rosenbloom, 1981). We believe that Bayesian methods for clustering and topic modeling can be more illuminating than these previous methods, due to their consideration of higher-order relationships within the data.

3. Format of the Paper

3.1. Partitioning the Text

3.2. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

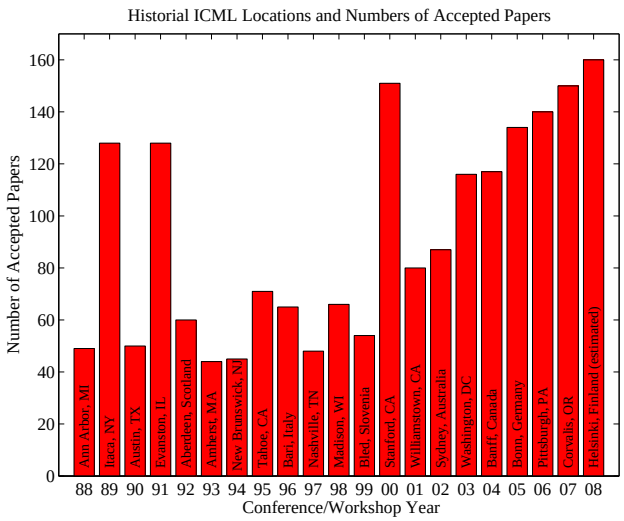

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

3.3. Citations and References

Acknowledgments

support.

References

Author, N. N. Suppressed for anonymity, 2011.

Borenszstein, et al. Xist-dependent imprinted x inactivation and the early developmental consequences of its failure. *Nature Structural & Molecular Biology*, 2017.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I.* Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.