Electronic Journal of Statistics > Vol. 3 (2009) open journal systems

Separating populations with wide data: A spectral analysis

Avrim Blum, *Carnegie Mellon University*
Amin Coja-Oghlan, *University of Edinburgh*
Alan Frieze, *Carnegie Mellon University*
Shuheng Zhou, *ETH Zurich*

Abstract

In this paper, we consider the problem of partitioning a small data sample drawn from a mixture of k product distributions. We are interested in the case that individual features are of low average quality $gamma$, and we want to use as few of them as possible to correctly partition the sample. We analyze a spectral technique that is able to approximately optimize the total data size—the product of number of data points n and the number of features K—needed to correctly perform this partitioning as a function of $1/gamma$ for $K>n$. Our goal is motivated by an application in clustering individuals according to their population of origin using markers, when the divergence between any two of the populations is small.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.

Keywords: mixture of product distributions, clustering, small sample, spectral analysis.

Full Text: PDF

Blum, Avrim, Coja-Oghlan, Amin, Frieze, Alan, Zhou, Shuheng, Separating populations with wide data: A spectral analysis, Electronic Journal of Statistics, 3, (2009), 76-113 (electronic). DOI: 10.1214/08-EJS289.

References

Achlioptas, D. and McSherry, F. (2005). On spectral learning of mixtures of distributions. In Proceedings of the 18th Annual COLT. (Version in http://www.cs.ucsc.edu/~optas/papers/). MR2203280

Arora, S. and Kannan, R. (2001). Learning mixtures of arbitrary gaussians. In Proceedings of 33rd ACM Symposium on Theory of Computing. MR2120323

Chaudhuri, K., Halperin, E., Rao, S. and Zhou, S. (2007). A rigorous analysis of population stratification with limited data. In Proceedings of the 18th ACM-SIAM SODA.

Coja-Oghlan, A. (2006). An adaptive spectral heuristic for partitioning random graphs. In Proceedings of the 33rd ICALP. MR2305568

Cryan, M. (1999). Learning and approximation Algorithms for Problems motivated by evolutionary trees. Ph.D. thesis, University of Warwick.

Cryan, M., Goldberg, L. and Goldberg, P. (2002). Evolutionary trees can be learned in polynomial time in the two state general markov model. SIAM J. of Computing 31 375–397. MR1861281

Dasgupta, A., Hopcroft, J., Kleinberg, J. and Sandler, M. (2005). On learning mixtures of heavy-tailed distributions. In Proceedings of the 46th IEEE FOCS.

Dasgupta, S. (1999). Learning mixtures of gaussians. In Proceedings of the 40th IEEE Symposium on Foundations of Computer S cience. MR1917603

Dasgupta, S. and Schulman, L. J. (2000). A two-round variant of em for gaussian mixtures. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI).

Feldman, J., O'Donnell, R. and Servedio, R. (2005). Learning mixtures of product distributions over discrete domains. In Proceedings of the 46th IEEE FOCS.

Feldman, J., O'Donnell, R. and Servedio, R. (2006). PAC learning mixtures of Gaussians with no separation assumption. In Proceedings of the 19th Annual COLT.

Fiedler, M. (1973). Algebraic connectivity of graphs. Czechoslovak Mathematical Journal 298–305. MR0318007

Fjallstrom, P. (1998). Algorithms for graph partitioning: a survey. Tech. rep., Linkoping University Electroni Press.

Freund, Y. and Mansour, Y. (1999). Estimating a mixture of two product distributions. In Proceedings of the 12th Annual COLT.

Kannan, R., Salmasian, H. and Vempala, S. (2005). The spectral method for general mixture models. In Proc. of the 18th Annual COLT. MR2203279

Kearns, M., Mansour, Y., Ron, D., Rubinfeld, R., Schapir, R. and Sellie, L. (1994). On the learnability of discrete distributions. In Proceedings of the 26th ACM STOC.

Latala, R. (2005). Some estimates of norms of random matrices. In Proceedings of the American Mathematical Society, vol. 133. MR2111932

McSherry, F. (2001). Spectral partitioning of random graphs. In Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science. MR1948742

Meckes, M. (2004). Concentration of norms and eigenvalues of random matrices. J. Funct. Anal. 211 508–524. MR2057479

Mossel, E. and Roch, S. (2005). Learning nonsinglar phylogenies and hidden markov models. In Proceedings of the 37th ACM STOC. MR2181638

Patterson, N., Price, A. and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2. Doi:10.1371/journal.pgen.0020190.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. nature genetics 38 904–909.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155 954–959.

Spielman, D. (2002). The behavior of algorithms in practice. Lecture notes.

Vempala, V. and Wang, G. (2002). A spectral algorithm of learning mixtures of distributions. In Proceedings of the 43rd IEEE FOCS.

Vu, V. (2005). Spectral norm of random matrices. In Proceedings of 37th ACM STOC. MR2181644

Zhou, S. (2006). Routing, Disjoint Paths, and Classification. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA. CMU Technical Report, CMU-PDL-06-109.

Electronic Journal of Statistics. ISSN: 1935-7524