Towards EM-style Algorithms for *a posteriori* Optimization of Normal Mixtures

Eric Sven Ristad
Mnemonic Technology, Inc.
50 Western Way
Princeton, NJ 08540 USA
Email ristad@mnemonic.com

Peter N. Yianilos
NEC Research Institute
4 Independence Way
Princeton, NJ 08540 USA
Email pny@research.nj.nec.com

Abstract — **Expectation maximization (EM) provides a simple and elegant approach to the problem of optimizing the parameters of a normal mixture on an unlabeled dataset. To accomplish this, EM iteratively reweights the elements of the dataset until a locally optimal normal mixture is obtained. This paper explores the intriguing question of whether such an EM-style algorithm exists for the related and apparently more difficult problem of finding a normal mixture that maximizes the *a posteriori* class probabilities of a labeled dataset.**

We expose a fundamental degeneracy in the relationship between a normal mixture and the *a posteriori* class probability functions that it induces, and use this degeneracy to prove that reweighting a dataset can almost always give rise to a normal mixture exhibiting any desired class function behavior. This establishes that EM-style approaches are sufficiently expressive for *a posteriori* optimization problems and opens the way to the design of new algorithms for them.

I. Technical Summary

Normal mixtures have proven useful in several areas including pattern recognition [3] and speech recognition – along with vector quantization and many others. The problem of finding a k-component normal mixture M that maximizes the likelihood $\prod_i p(s_i|M)$. of an unlabeled dataset $s_1, \ldots, s_n$ may be approached using the well-known expectation maximization (EM) algorithm [2, 1]. EM iteratively reweights each sample's membership in each of the k mixture components by the posteriori probability of each component given the sample.

When normal mixtures are applied to pattern classification problems each mixture component corresponds to a pattern class. Given a class label $\omega(s_i)$ for each element s_i in the dataset, the goal is to maximize the mixture's *a posteriori* performance $\prod_i p(\omega(s_i)|s_i, M)$, i.e.to predict well the correct labels — not model the observation vectors themselves.

We define two normal mixtures to be *class equivalent* if they induce identical *a posteriori* class functions $p(\omega|x, M)$, i.e. perform identically as probabilistic classifiers. Theorem 1 shows that the relationship between mixtures and their class behavior is highly degenerate and, we suggest, somewhat strange and counterintuitive. As a positive result of this degeneracy one can search the entire space of class functions without considering all possible mixtures. So to solve the *a posteriori* maximization problem above, it suffices to find *any* normal mixture that induces optimal class functions.

Theorem 1 *Let p be a d-dimensional normal mixture with k components. For any $x \in \mathbb{R}^d$ and $\epsilon > 0$, there exists a d-dimensional k-component normal mixture p', such that for $1 \leq i \leq k$:*

1. $\|\mu'_i - x\| < \epsilon$
2. $\|\Sigma'_i\| < \epsilon$
3. *p' and p are class-equivalent*

The following theorem shows that by reweighting a dataset in a way that differs slightly from that of EM, arbitrary class behavior may be induced.

Theorem 2 *Given any d-dimensional normal mixture M with k components, and $s_1, \ldots, s_n \in \mathbb{R}^d$ such that $n \geq d(d+3)/2$ with some subset of size $d(d+3)/2$ satisfying a mild orthodoxy condition, then there exists a $k \times n$ table of nonnegative values $\{\gamma_{i,j}\}$, and nonnegative values $m_1, \ldots, m_{k-1}$ with $\sum m_k \leq 1$, such that the normal mixture M' generated as described below, is class-equivalent to M.*

1. *The mixing parameters of M' are $m_1, \ldots, m_{k-1}, 1 - \sum_{j=1}^{k-1} m_j$.*
2. *Each mean μ'_i within M' is given by:*

$$\mu_i = \frac{1}{\sum_{j=1}^n \gamma_{i,j}} \sum_{j=1}^n \gamma_{i,j} s_j$$

3. *Each covariance Σ'_i within M' is given by:*

$$\Sigma_i = \sum_{j=1}^n \gamma_{i,j} (s_j - \mu_i)(s_j - \mu_i)^t$$

Our work also shows that reweighting in exactly the fashion of EM, fails to be universally expressive in the sense above. That is, where $\Sigma_i = \frac{1}{\sum_{j=1}^n \gamma_{i,j}} \sum_{j=1}^n \gamma_{i,j} (s_j - \mu_i)(s_j - \mu_i)^t$.

Acknowledgements

The authors thank Leonid Gurvits and Joe Kilian for helpful discussions.

References

[1] L. E. Baum and J. E. Eagon, *An inequality with application to statistical estimation for probabalistic functions of a Markov process and to models for ecology*, Bull. AMS, 73 (1967), pp. 360–363.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum-likelihood from incomplete data via the EM algorithm*, J. Royal Statistical Society Ser. B (methodological), 39 (1977), pp. 1–38.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, Inc., 1973.