Stable Mixing of Complete and Incomplete Information

Adrian Corduneanu
adrianc@mit.edu

Tommi Jaakkola
tommi@ai.mit.edu

MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139

Abstract

An increasing number of parameter estimation tasks involve the use of at least two information sources, one complete but limited, the other abundant but incomplete. Standard algorithms such as EM (or em) used in this context are unfortunately not stable in the sense that they can lead to a dramatic loss of accuracy with the inclusion of incomplete observations. We provide a more controlled solution to this problem through differential equations that govern the evolution of locally optimal solutions (fixed points) as a function of the source weighting. This approach permits us to explicitly identify any critical (bifurcation) points leading to choices unsupported by the available complete data. The approach readily applies to any graphical model in $O(n^3)$ time where n is the number of parameters. We use the naive Bayes model to illustrate these ideas and demonstrate the effectiveness of our approach in the context of text classification problems.

1 Introduction

Many modern application areas such as text classification involve estimating generative probability models under incomplete or partial information. The incomplete nature of the available data can be often thought of as arising from two data sources, one complete but limited and the other abundant but incomplete. The estimation problem is consequently cast as one with two (or more) heterogeneous data sources. Estimating generative models is not the only way of exploiting such data sources, especially in the context classification problems with labeled and unlabeled examples. While transductive algorithms have been demonstrated as viable (or sometimes even superior) alternatives in classification contexts, we focus here only on likelihood based estimation of generative models but envision a useful application of (some of) the ideas described in this paper also to transductive methods.

We identify two related but fundamental estimation issues: 1) how to allocate or balance the two sources of information to maximize the accuracy of the resulting model and 2) how to ensure that the estimation algorithm itself remains stable (e.g., small changes in the source allocation results in small changes in the accuracy of the solution). Neither question is adequately answered by the usual application of the EM (or em) algorithm. Indeed, the two algorithms are not stable, nor do they provide any help in determining the appropriate balance of the sources. Indeed, incorporating more incomplete data for use with the EM (or em) algorithm can result in a dramatic loss of accuracy.

The main contribution of this paper is to provide a stable alternative to the EM and em algorithms. The idea is to evolve differential equations that govern the fixed points of the two algorithms, where the source allocation parameter takes the role of time in the evolution process. The differential equations are initialized with the solution obtained from the complete data source alone. The significant feature of this approach is that it provides an explicit way of identifying critical points that arise in these estimation problems. In case where the incomplete data sources is overwhelmingly more abundant, our approach yields a close approximation to the optimal allocation parameter and an interval estimate for other cases.

The paper is organized as follows. We begin by describing two estimation criteria leading to the EM and em algorithms and provide an analysis of the two algorithms in the case of a naive Bayes model and predominantly incomplete data. We then describe our alternative approach based on evolving differential equations. We corroborate the method with experiments in text classification problems.

2 Estimation criteria for incomplete data

Assume two sets of data, one complete $(x_i, y_i)_{1 \le i \le n}$, and the other incomplete $(x'_j)_{1 \le j \le m}$, where $n \ll m$ so that we have many more incomplete than complete observations. Both sets of data are assumed to have been generated i.i.d. from the same unknown generative distribution $P^*(x, y)$ over the joint space $\mathcal{X} \times \mathcal{Y}$. Our goal here is to combine the two types of data sources so as to find the distribution Q from a chosen model family $\mathcal{M}$ that best matches $P^*(x, y)$ in the following KL-divergence sense:

$$Q^* = \arg \min_{Q \in \mathcal{M}} D(P^*_{x,y} \,\|\, Q_{x,y})$$

If we replace P^* with $\hat{P}^l(x, y)$, the empirical distribution from the complete data, we obtain the standard maximum likelihood estimate for Q.

To incorporate any information from the incomplete dataset, we must both restrict the model family $\mathcal{M}$ as well as modify the estimation criterion. Imposing constraints between $Q(x)$ and $Q(y|x)$ is necessary to be able to update $Q(y|x)$ on the basis of the incomplete observations [1] [2]. We consider two related estimation criteria. First, we can try to maximize the likelihood of all the observed data or, more generally, minimize

$$(1 - \lambda) D(\hat{P}^l_{x,y} \,\|\, Q_{x,y}) + \lambda D(\hat{P}^u_x \,\|\, Q_x) \tag{1}$$

where $\hat{P}^l(x, y)$ and $\hat{P}^u(x)$ are the empirical distributions from the two data sources. Note that by setting the source allocation parameter λ equal to $m/(n + m)$, we recover the basic likelihood criterion augmented with unlabeled data. The allocation parameter is introduced here to emphasize the fact that the observed frequencies of the examples need not yield the best combination of the sources. The parameter λ will indeed play a crucial role later in the paper.

Another related criterion is obtained from a geometric perspective. Define $\mathcal{P}_\lambda$ as the set of distributions obtained by completing the incomplete data with any conditional $T(y|x)$:

$$\mathcal{P}_\lambda = \left\{ P : \; P(x, y) = (1 - \lambda)\hat{P}^l(x, y) + \lambda \hat{P}^u(x) T(y|x) \text{ for some } T(y|x) \right\}$$

The estimate Q is now found by minimizing the KL-divergence to the set $\mathcal{P}_\lambda$ or

$$D(\mathcal{P}_\lambda \,\|\, Q_{x,y}) \equiv \min_{P \in \mathcal{P}_\lambda} D(P_{x,y} \,\|\, Q_{x,y}) \tag{2}$$

The first criterion leads to a weighted version of the EM-algorithm and the latter one to the em-algorithm [3]. The two algorithms are often identical but not here.

2.1 Identifiability

The limiting case of infinite unlabeled data with $\lambda = 1$ deserves further consideration. It tells us how much information can be obtained about $Q(y|x)$ knowing only $Q(x)$. The answer is inherently a characteristic of $\mathcal{M}$ as well as the initial choice for Q. For example, in a classification context where y denotes the label, a typical model family $\mathcal{M}$ has at least $|\mathcal{Y}|!$ (permutation of the labels) distinct distributions corresponding to any common achievable marginal $Q(x)$.

A desirable property of the model family in estimation with incomplete data is *identifiability*. We call the family *identifiable with incomplete data* if there are exactly $|\mathcal{Y}|!$ distributions in $\mathcal{M}$ for a given achievable marginal $Q(x)$. If a model family is identifiable with incomplete data, model parameters can be estimated on the basis of unlabeled data alone except for the class permutation, which can be readily inferred from a few complete samples.

The relation of this notion of identifiability to common model families is not yet fully understood. Mixtures of Gaussians are identifiable [4] in this sense, as well as discrete naive Bayes models with binary latent variable and at least three binary features [5], while those with only two features are not. Naive Bayes models with more than two classes and enough features can also be identifiable, though the exact number of features required remains unclear.

We do not assume identifiability with incomplete data in the remainder of the paper but merely emphasize here that this property plays an important role in providing *a priori* guarantees of success. Instead, we focus on finding alternative means of stabilizing the estimation process.

3 The EM and em algorithms

Incomplete-data criteria such as (1) and (2) are typically optimized with iterative procedures like EM [6] and Amari's em [3], respectively. The two algorithms differ only in terms of how they complete incomplete observations in the E/e-steps owing to the different estimation criteria. The M/m-steps of the algorithms employ the same information divergence projection to derive the parameters of the next iterate: find $Q \in \mathcal{M}$ that minimizes $D(P'_{x,y} \,\|\, Q_{x,y})$, where $P'_{x,y}$ embodies the data completion (E/e-step).

We illustrate the two algorithms here on a discrete naive Bayes model family, where the data variable x consists of k features $(x_1, x_2, \ldots, x_k)$ independent given the class label y:

$$Q(x,y) = \left[\prod_{i=1}^{k} Q_i(x_i|y)\right] Q(y) = \left[\prod_{i=1}^{k} Q_i(x_i,y)\right] Q(y)^{1-k} \tag{3}$$

For the naive Bayes model family both algorithms have the following form:

$$\begin{array}{ll} \text{E-step:} & P(x,y) \leftarrow (1-\lambda)\hat{P}^l(x,y) + \lambda \hat{P}^u(x) T(y|x) \\ \text{M-step:} & Q_i(x_i,y) \leftarrow \sum_{x \backslash x_i} P(x,y) \end{array} \tag{4}$$

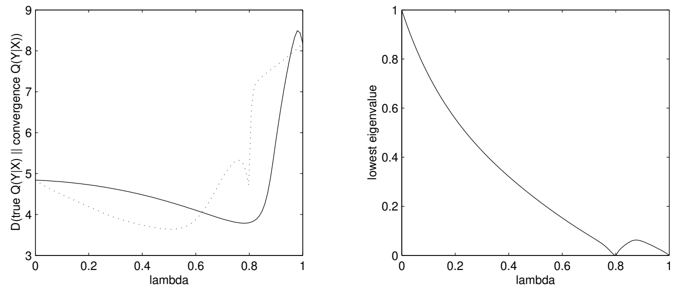

Figure 1: Left: accuracy of the EM (solid) and em (dotted) solutions as a function of λ on 20 complete and 2000 incomplete samples from a known binary naive Bayes model with four features. Right: critical λ detected at discontinuity.

where $T(y|x) = Q(y|x)$ in EM, while in em $T(y|x) = 0$ for $y \in I$ and $T(y|x) = \left(1 + \frac{1-\lambda}{\lambda}\frac{\hat{P}^l(x)}{\hat{P}^u(x)}\right)\frac{Q(y|x)}{\sum_{y' \notin I} Q(y'|x)} - \frac{1-\lambda}{\lambda}\frac{\hat{P}^l(x,y)}{\hat{P}^u(x)}$ for $y \notin I$, where I is minimal such that all $T(y|x)$'s are in $[0,1]$.

Both algorithms find locally optimal solutions of their respective criterion. These solutions are characterized as self-consistent fixed points of the projections. These fixed points are unlikely to correspond to the globally optimal solutions (which can also be degenerate).

When λ approaches 1, the regime where the two estimation criteria are starting to agree, both algorithms may converge to fixed points that are significantly different from the privileged one that remains closest to the complete data (ML) solution. Grounding the solution to the complete data case is desirable since otherwise we have little reason to expect that the resulting model would serve well as a classifier. Figure 1 illustrates this point. We ran a simulated experiment on a known binary naive Bayes model with four features, and plotted the KL-divergence between the true conditional $P(y|x)$ and those coming from the fixed points of the EM and em-algorithms as a function of the source allocation parameter λ. The accuracy of the resulting models diverge rapidly as λ approaches 1.

The geometric interpretation of em as an alternating minimization of divergences in the spirit of Csiszár [7] reveals the problem. em performs e and m projections between the observed data manifold $\mathcal{D}_\lambda = (1-\lambda)\hat{P}^l + \lambda T^{-1}(\hat{P}^u)$ and the model family $\mathcal{M}$, where T is the y-marginalization mapping between distributions on (x, y) and x. Because $\mathcal{D}$ is convex, if $\mathcal{M}$ were convex, the iteration would be well-behaved and guaranteed to converge to the global minimum [7]. However, $\mathcal{M}$ is rarely convex (surely not for naive Bayes), and the convergence behavior is sometimes characterized by large jumps from a local convex region of $\mathcal{M}$ to another when the data manifold $\mathcal{D}_\lambda$ changes as a function of λ (Figure 2). A similar continuity argument applies to EM but due to the weighted allocation of sources, its geometry is governed by a different set of unnormalized measures [1].

4 Controlled selection of fixed points

In what follows we treat the iteration of the optimization algorithm (be it EM, em, or other algorithm) as an operator EM_λ acting on $\mathcal{M}$. The algorithm computes for each λ a fixed point $Q_\lambda = \mathrm{EM}_\lambda(Q_\lambda)$ of the operator. Ideally we want to control stability

[1] Alternating minimization between $\{[(1-\lambda)\hat{P}^l(x,y)\ \lambda\hat{P}^u(x)T(y|x)] \mid \text{for some } T(y|x)\}$ and $\{[(1-\lambda)Q(x,y)\ \lambda Q(x,y)] \mid Q \in \mathcal{M}\}$ where the brackets combine the elements in the product space of unnormalized measures.

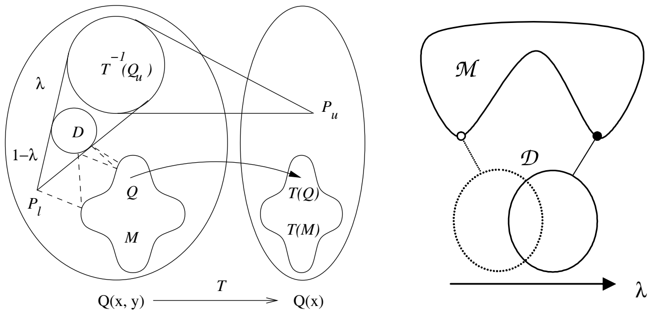

Figure 2: em as an alternating minimization of KL-divergence between the observed family $\mathcal{D}$ and the model family $\mathcal{M}$. Due to the non-convexity of $\mathcal{M}$ as $\mathcal{D}$ moves with λ, uncontrolled jumps in convergence are possible.

and predictability by selecting fixed points that are firmly grounded in the complete data estimate, represented by Q_0. In order to establish the connection between Q_λ and Q_0, we view the intermediate Q_λ's as tracing a continuous curve of fixed points parameterized by λ. Discontinuities may occur, however, and we treat those Q_λ's that cannot be continuously traced back to Q_0 as inappropriate (not grounded).

We can characterize the continuous curve Q_λ by setting up the differential equation that governs the evolution of fixed points with λ:

$$\frac{d}{d\lambda}Q_\lambda = (I - \nabla_Q \mathrm{EM}_\lambda(Q_\lambda))^{-1} \frac{\partial}{\partial\lambda}\mathrm{EM}_\lambda(Q_\lambda) \tag{5}$$

Any initial fixed point can be traced to a unique continuous path until the (transformed) Jacobian $I - \nabla_Q \mathrm{EM}_\lambda(Q_\lambda)$ becomes singular. Such critical λ's may reflect a discontinuity or a bifurcation into separate paths of fixed points. While such paths could be traced on the basis of the eigenvector corresponding to the zero eigenvalue, the choice of the paths would not be supported by the complete data.

For the purpose of solving the equations, the initial condition is the unique fixed point $Q_0 = \hat{P}^l$, the complete-data estimate [2]. We find a continuous path of fixed points connected to Q_0 by evolving the differential equation until we reach the first critical λ. The resulting bifurcation points could be traced further as explained above but discontinuities may be followed only by exploring with EM. However, we do not proceed beyond the critical λ as we believe we no longer have any guarantees about the ensuing estimates. Choosing λ close to the first critical point means that we exploit the unlabeled examples (or more generally incomplete data) to their fullest potential.

4.1 Relationship to fixed points of plain EM

It is worth asking whether the differential equation (call it DIFFEM) also finds different fixed points than the EM-like algorithms initialized with the complete data solution $\hat{P}^l$. Since $\lambda = 0$ is never critical (cf. eigenvalues of the modified Jacobian start at 1) and EM_λ is continuous in λ, the $\lambda = 0$ neighborhood of EM_λ has a unique fixed point. However, as λ increases, parallel fixed-point paths may emerge, even if DIFFEM does not signal a critical λ. Empirically we found that EM converges to the DIFFEM solution for a while, but it can jump to a different fixed-point path even at non-critical λ. From the point of

[2] In practice we smooth $\hat{P}^l$ as EM_λ is undefined on zero probabilities.

view of estimation with incomplete data these are jumps that we want to avoid. Therefore, DIFFEM not only extends $\hat{P}^l$ continuously as much as is reasonable, it also protects against unjustified discontinuities introduced by EM.

4.2 Differential equation for naive Bayes

To illustrate, we detail the differential equation applied to EM and discrete naive Bayes (3). We over-parameterize the model to simplify computation by the vector $(Q(y)\ Q(x_1, y)\ \ldots\ Q(x_k, y))$. EM_λ now acts on an unconstrained vector of $|\mathcal{Y}|(1 + \sum_{i=1}^{k} |\mathcal{X}_k|)$ components, but its fixed points are still valid distributions. Under this parameterization the EM_λ operator is:

$$\begin{array}{l} Q_i(y) \leftarrow (1-\lambda)\hat{P}^l(y) + \lambda \sum_x \hat{P}^u(x) Q(y|x) \\ Q_i(x_i, y) \leftarrow (1-\lambda)\hat{P}^l(x_i, y) + \lambda \sum_{x \setminus x_i} \hat{P}^u(x) Q(y|x) \end{array} \tag{6}$$

Thus to compute $\nabla_Q \mathrm{EM}_\lambda$ we only need to differentiate $Q(y|x)$ obtained from (3) with respect to all components of the parameter vector:

$$\frac{dQ(y')}{dQ(y|x)} = (k-1)\left[\frac{Q(y|x)Q(x|y')}{Q(x)} - (y = y')\frac{Q(x|y')}{Q(x)}\right] \tag{7}$$

$$\frac{dQ_i(x_i, y')}{dQ(y|x)} = -\frac{Q(y|x)Q(x \setminus x_i|y')}{Q(x)} + (y = y')\frac{Q(x \setminus x_i|y')}{Q(x)} \tag{8}$$

if the i'th component of x is x_i, 0 otherwise. Finally from (6) we get $\frac{\partial}{\partial\lambda}\mathrm{EM}_\lambda = \mathrm{EM}_1 - Q_0$.

Regarding computational complexity, the dominant operation is the matrix inversion which is cubic in the number of parameters in the model. In addition, summations in (6) can be carried out linearly in the number of unlabeled samples.

5 Experiments

We illustrate how performance becomes unpredictable beyond the critical source allocation parameter by running experiments on a document classification task from the 20-newsgroups dataset. We have modeled each document by naive Bayes with binary features that indicate the presence or absence of words [3]. To save computational time we performed feature selection by mutual information and reduced the number of features to the most 25 to 40 significant words, and also restricted the number of classes to 3 up to 5. To train the model we have used 10 to 100 labeled, and 1000 to 5000 unlabeled samples. The differential equation was solved by a Runge-Kutta method on 1000 interval subdivisions. Only for the purpose of comparison with EM, we have followed the differential equation even beyond critical λ by running a few EM iterations to stabilize to a new continuous path of fixed points. However, there is more than one direction to pass the discontinuity, and our choice is not necessarily the best (if "best" can be defined at all after critical λ).

You can see in Figure 3 how classification accuracy evolves smoothly before the critical point, while there is a big drop in performance at λ critical. Beyond that λ there is no reason to associate fixed points with complete-data evidence. Note that DIFFEM does not suggest an optimal λ, which might be less than critical, but only imposes an upper bound after which performance is most likely to drop significantly. We found that even without having

[3] The differential equation can also treat multinomial naive Bayes models as in [2]

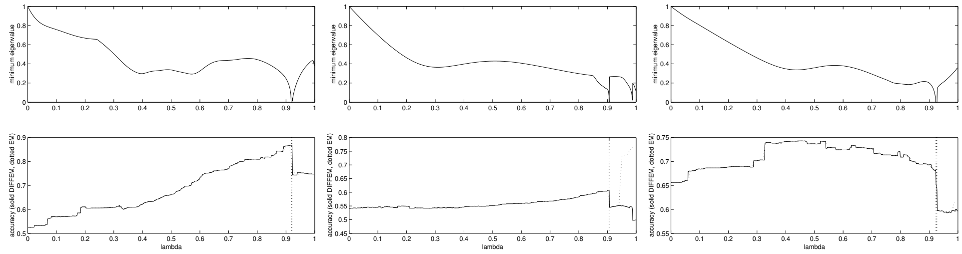

Figure 3: Three runs of DIFFEM versus EM. For each run the upper graph is the minimum eigenvalue of the transformed Jacobian, and the lower graph the classification accuracy of DIFFEM and EM as a function of λ. Most of the time the dotted-line EM coincides with DIFFEM.

a procedure for selecting optimal λ other than setting it to the critical value, DIFFEM achieved better performance than EM most of the time, revealing serious predictability issues with EM.

6 Discussion

We have introduced here a stable alternative to the standard EM (or em) algorithm for estimation with incomplete data. The approach is based on evolving differential equations of fixed points starting with the solution obtained from few available complete observations. The advantage of our approach is that we can explicitly identify any critical points that lead to solutions unsupported by the available complete data. The differential equations can be readily formulated for any incomplete data estimation problem involving graphical models but with a cost of $O(n^3)$, where n is the number of parameters. More efficient approximate methods or methods better exploiting the structure of the problem will need to be developed. Another limitation of our approach is that we cannot yet clearly identify the optimal value for the source allocation parameter λ except for an upper bound. However, in the case of relatively abundant source of incomplete data, a value close to the first critical point is appropriate so as to derive a maximal (controlled) use of the available data.

References

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[2] K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

[3] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[4] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*, chapter 2.7. Wiley, New York, 1997.

[5] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. Technical Report MSR-TR-98-31, Microsoft Research, July 1998.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–22, 1977.

[7] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplement Issue No. 1:205–237, 1984.