Incomplete-Data Classification using Logistic Regression

David Williams DPW@EE.DUKE.EDU
Xuejun Liao XJLIAO@EE.DUKE.EDU
Ya Xue YX10@EE.DUKE.EDU
Lawrence Carin LCARIN@EE.DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA

Abstract

A logistic regression classification algorithm is developed for problems in which the feature vectors may be missing data (features). Single or multiple imputation for the missing data is avoided by performing analytic integration with an estimated conditional density function (conditioned on the non-missing data). Conditional density functions are estimated using a Gaussian mixture model (GMM), with parameter estimation performed using both expectation maximization (EM) and Variational Bayesian EM (VB-EM). Using widely available real data, we demonstrate the general advantage of the VB-EM GMM estimation for handling incomplete data, vis-à-vis the EM algorithm. Moreover, it is demonstrated that the approach proposed here is generally superior to standard imputation procedures.

1. Introduction

The incomplete-data problem, in which certain features are missing from particular feature vectors, exists in a wide range of fields, including social sciences, computer vision, biological systems, and remote sensing, among others. For example, partial responses in surveys are common in the social sciences, leading to incomplete data sets with arbitrary patterns of missing data. In remote sensing applications, incomplete data can result when only a subset of sensors (*e.g.,* radar, infrared, acoustic) are deployed at certain regions. Increasing focus in the future on using (and fusing data from) multiple sensors or information sources will make such incomplete-data problems increasingly common (see (Tsuda, Akaho & Asai, 2003; Lanckriet et al., 2004)). This work assumes the data are either missing completely at random (MCAR) or missing at random (MAR), meaning that the missing data is independent of its value (see (Ghahramani & Jordan, 1994; Rässler, 2004) for more details).

Incomplete-data problems are often circumvented in the initial stage of analysis—before specific algorithms become involved—via imputation (*i.e.,* by "completing" the missing data by filling in specific values). Common imputation schemes include "completing" missing data with zeros, the unconditional mean, or the conditional mean (if one has an estimate for the distribution of missing features given the observed features, $P(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i})$).

Semidefinite programming has been used to complete kernel matrices that have a *limited* number of missing elements (Graepel, 2002). The *em* algorithm (Tsuda, Akaho & Asai, 2003) is applicable when both an incomplete auxiliary kernel matrix and a complete primary kernel matrix exist, but not when the patterns of missing data are completely arbitrary. Both of these methods can be viewed as single imputation schemes, since missing data are completed with single values before standard classification algorithms are applied. Since single imputation treats the missing data as fixed known data, the uncertainty of the missing data is ignored (Rässler, 2004).

The method of multiple imputation (Rubin, 1987), which has flourished in the statistics community for dealing with incomplete data, goes beyond these single, deterministic completions. In multiple imputation, $n > 1$ samples are generated according to the (estimated) distribution $P(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i})$ for every missing feature, and n imputed data sets are formed with the missing data completed by these samples. Standard complete-data analysis (*e.g.,* learning and classification) is then performed on each of these completed data sets. The results of each imputed data set are

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

then combined (*e.g.,* averaged) to obtain a single set of results. Theoretical work (Rubin, 1987) has shown the proximity between an estimate's uncertainty resulting from a small number of imputations and an infinite number of imputations. It should be noted that the imputation (sampling) is performed only because the desired posterior distribution of a parameter involves an intractable integral.

The intractable integral is avoided in (Ibrahim, 1990) by requiring the data to be discrete. This discrete assumption allows missing data to be summed over, leading to a "weighted EM" algorithm from which maximum likelihood parameter estimates (*e.g.,* classifier weights) can be obtained. However, the method, developed for generalized linear models with incomplete data, does not extend to the continuous case.

In this paper we tackle the incomplete (continuous) data problem for logistic regression classification in a principled manner, avoiding explicit imputation. When calculating the posterior distribution of a parameter, it is proper to integrate out missing data (Duda, Hart & Stork, 2000):

$$P\left(y_i|\mathbf{x}_i^{o_i}\right) = \int P\left(y_i|\mathbf{x}_i^{m_i},\mathbf{x}_i^{o_i}\right) P\left(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i}\right) d\mathbf{x}_i^{m_i}. \quad (1)$$

This is the aforementioned integral that is intractable in general. However, in the case of logistic regression (with y_i the class label), this integral can be solved analytically using two minor assumptions. The first assumption is that $P(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i})$ is a Gaussian mixture model (GMM). This assumption is mild, since it is well-known that a mixture of Gaussians can approximate any distribution. The second (highly accurate) assumption is that the sigmoid function can be approximated as the cumulative distribution function (cdf) of a Gaussian. Since the integral in (1) can be solved analytically, the likelihood can be maximized in a manner analogous to the complete-data case, to obtain classifier weights. Once the weights are obtained, the classification algorithm can be applied to classify incomplete testing data.

The key idea is that the integral can be computed analytically in the special case of logistic regression, which allows us to avoid using data completion (imputation) methods. Since the GMM plays an integral role in the proposed algorithm, we show two different methods to accurately perform this GMM density estimation in the presence of missing data. The first method uses the Expectation-Maximization (EM) algorithm (Dempster, Laird & Rubin, 1977), while the second, more robust method uses the Variational Bayesian EM (VB-EM) algorithm (Beal & Ghahramani, 2003).

The remainder of the paper is organized as follows. In Section 2 we derive the logistic regression algorithm for incomplete data. In Section 3 we show the equations for the EM and VB-EM algorithms for estimating GMMs from incomplete data. Experimental classification results are shown in Section 4, before concluding remarks are made in Section 5.

2. Logistic Regression for Incomplete Data

Assume we have an incomplete labeled data set

$$\mathcal{D}_l = \{(\mathbf{x}_i, y_i, m_i) : \mathbf{x}_i \in \mathbb{R}^d, x_{ij} \,\text{missing}\, \forall j \in m_i\}_{i=1}^N \quad (2)$$

where $\mathbf{x}_i$ is the i-th data point, labeled as $y_i \in \{-1, 1\}$; the features in $\mathbf{x}_i$ indexed by m_i (*i.e.,* $x_{ij}, j \in m_i$) are missing. Each $\mathbf{x}_i$ has its own (possibly unique) set of missing features, m_i. One special case is when a subset of data share common missing features, as with multisensor data where the common missing features result from a sensor that has not collected data.

In logistic regression, the probability of label y_i given $\mathbf{x}_i$ is $P\left(y_i|\mathbf{x}_i\right) = \sigma(y_i\mathbf{w}^T\mathbf{x}_i)$, where $\sigma(\nu) = (1 + \exp(-\nu))^{-1}$ and $\mathbf{w}$ constitutes a classifier. We partition $\mathbf{x}_i$ into its observed and missing parts, $\mathbf{x}_i = [\mathbf{x}_i^{o_i}; \mathbf{x}_i^{m_i}]$ where $\mathbf{x}_i^{o_i} = [x_{ij}, j \in o_i]^T$, $\mathbf{x}_i^{m_i} = [x_{ij}, j \in m_i]^T$, and $o_i = \{1, \cdots, d\} \setminus m_i$ is the set of observed features in $\mathbf{x}_i$. We apply the same partition to $\mathbf{w}$ to obtain $\mathbf{w} = [\mathbf{w}_{o_i}; \mathbf{w}_{m_i}]$, yielding

$$P\left(y_i|\mathbf{x}_i\right) = \sigma(y_i(\mathbf{w}_{o_i}^T\mathbf{x}_i^{o_i} + \nu_i)) \quad (3)$$

where $\nu_i = \mathbf{w}_{m_i}^T\mathbf{x}_i^{m_i}$. Because $\mathbf{x}_i^{m_i}$ (and hence ν_i) is missing, (3) cannot be evaluated. Instead the needed probability of y_i given the observed features $\mathbf{x}_i^{o_i}$ can be written as

$$\begin{aligned} P\left(y_i|\mathbf{x}_i^{o_i}\right) &= \int P\left(y_i|\mathbf{x}_i^{m_i},\mathbf{x}_i^{o_i}\right) P\left(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i}\right) d\mathbf{x}_i^{m_i} \\ &= \int \sigma(y_i(\mathbf{w}_{o_i}^T\mathbf{x}_i^{o_i} + \nu_i)) P\left(\nu_i|\mathbf{x}_i^{o_i}\right) d\nu_i \quad (4) \end{aligned}$$

To perform the integration in (4), $P\left(\nu_i|\mathbf{x}_i^{o_i}\right)$ must be known. We assume that $P\left(\mathbf{x}_i\right)$ is a Gaussian mixture model (GMM):

$$P(\mathbf{x}_i) = \sum_{k=1}^K \pi_k \mathcal{N}\left(\left[\begin{array}{c} \mathbf{x}_i^{o_i} \\ \mathbf{x}_i^{m_i} \end{array}\right]; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \quad (5)$$

where $\pi_k \geq 0$, $\sum_{k=1}^K \pi_k = 1$, and

$$\boldsymbol{\mu}_k = \left[\begin{array}{c} \boldsymbol{\mu}_k^{o_i} \\ \boldsymbol{\mu}_k^{m_i} \end{array}\right], \qquad \boldsymbol{\Sigma}_k = \left[\begin{array}{cc} \boldsymbol{\Sigma}_k^{o_i o_i} & \boldsymbol{\Sigma}_k^{m_i o_i\, T} \\ \boldsymbol{\Sigma}_k^{m_i o_i} & \boldsymbol{\Sigma}_k^{m_i m_i} \end{array}\right]. \quad (6)$$

Because of the linear relation $\nu_i = \mathbf{w}_{m_i}^T \mathbf{x}_i^{m_i}$, $P\left(\nu_i|\mathbf{x}_i^{o_i}\right)$ is also a GMM,

$$P\left(\nu_i|\mathbf{x}_i^{o_i}\right) = \sum_{k=1}^{K} \delta_k^i G\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right), \tag{7}$$

with the parameters

$$\delta_k^i = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_i^{o_i}; \boldsymbol{\mu}_k^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i}\right)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\left(\mathbf{x}_i^{o_i}; \boldsymbol{\mu}_\ell^{o_i}, \boldsymbol{\Sigma}_\ell^{o_i o_i}\right)} \tag{8}$$

$$\zeta_k^i = \mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^i \tag{9}$$

$$\alpha_k^i = \sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i}} \tag{10}$$

$$\boldsymbol{\xi}_k^i = \boldsymbol{\mu}_k^{m_i} + \boldsymbol{\Sigma}_k^{m_i o_i} \boldsymbol{\Sigma}_k^{o_i o_i\ -1} (\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i}) \tag{11}$$

$$\boldsymbol{\Omega}_k^i = \boldsymbol{\Sigma}_k^{m_i m_i} - \boldsymbol{\Sigma}_k^{m_i o_i} \boldsymbol{\Sigma}_k^{o_i o_i\ -1} \boldsymbol{\Sigma}_k^{m_i o_i\ T} \tag{12}$$

where

$$G(\nu_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\nu_i^2}{2}\right) \tag{13}$$

is a standard univariate Gaussian density function with zero mean and unit variance.

We approximate the sigmoid function as the cdf of a Gaussian (*i.e.,* a probit function)

$$\sigma(\nu) = \int_{-\infty}^{\nu} G\left(\frac{z}{\beta}\right) dz \tag{14}$$

where $\beta = \frac{\pi}{\sqrt{3}}$. Substituting (7) and (14) into (4), we obtain

$$\begin{aligned}
P\left(y_i|\mathbf{x}_i^{o_i}\right) &= \int\int_{-\infty}^{y_i(\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \nu_i)} G\left(\frac{z}{\beta}\right) dz \\
&\quad \sum_{k=1}^{K} \delta_k^i G\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right) d\nu_i \\
&\overset{a}{=} \int\int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} G\left(\frac{z' + y_i \nu_i}{\beta}\right) dz' \\
&\quad \sum_{k=1}^{K} \delta_k^i G\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right) d\nu_i \\
&\overset{b}{=} \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \int G\left(\frac{z' + y_i \nu_i}{\beta}\right) \\
&\quad G\left(\frac{y_i \nu_i - y_i \zeta_k^i}{y_i \alpha_k^i}\right) d\nu_i\, dz' \\
&\overset{c}{=} \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} G\left(\frac{z' + y_i \zeta_k^i}{\sqrt{(y_i \alpha_k^i)^2 + \beta^2}}\right) dz' \\
&\overset{d}{=} \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{\frac{y_i(\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \zeta_k^i)\beta}{\sqrt{(\alpha_k^i)^2 + \beta^2}}} G\left(\frac{z}{\beta}\right) dz \\
&\overset{e}{=} \sum_{k=1}^{K} \delta_k^i \sigma\left(\frac{y_i \beta (\zeta_k^i + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{(\alpha_k^i)^2 + \beta^2}}\right)
\end{aligned} \tag{15}$$

where equation a results from the change of variable $z' = z - y_i \nu_i$, equation b is due to exchanging the order of integrals and summation, equation c results because the convolution of two Gaussians is a Gaussian, equation d results from the change of variable $z = \frac{(z' + y_i \zeta_k^i)\beta}{\sqrt{(\alpha_k^i)^2 + \beta^2}}$, and equation e is obtained by reverting to sigmoid representation. Thus we have expressed $P\left(y_i|\mathbf{x}_i^{o_i}\right)$ as a mixture of sigmoids. Substituting (9) and (10) into (15), we obtain the probability of y_i given only the observed portion of $\mathbf{x}_i$:

$$P\left(y_i|\mathbf{x}_i^{o_i}\right) = \sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i \beta (\mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^i + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i} + \beta^2}}\right). \tag{16}$$

For the data set in (2), assuming the data points are independent of each other, we have the log-likelihood function

$$\begin{aligned}
\ell(\mathbf{w}) &= \ln P\left(\{y_i\}_{i=1}^N | \{\mathbf{x}_i^{o_i}\}_{i=1}^N\right) \\
&= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i \beta (\mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^i + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i} + \beta^2}}\right).
\end{aligned} \tag{17}$$

Since the objective function (17) to be maximized is not concave, the solution may be trapped in local maxima. A good initialization is important, so we initialize $\mathbf{w}$ as follows. We "complete" the data set by replacing the missing features $\mathbf{x}_i^{m_i}$ with the conditional mean $E(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i}) = \sum_{k=1}^{K} \delta_k^i \boldsymbol{\xi}_k^i$, where δ_k^i and $\boldsymbol{\xi}_k^i$ are defined in (8) and (11), respectively. This "completed" data set is then submitted to the standard logistic regression to obtain $\mathbf{w}_0$, which is the maximizer of

$$\sum_{i=1}^{N} \ln \sigma\left(y_i \mathbf{w}_{m_i}^T \sum_{k=1}^{K} \delta_k^i \boldsymbol{\xi}_k^i + y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}\right).$$

We then use $\mathbf{w}_0$ as the initialization of $\mathbf{w}$ in maximizing (17) by gradient ascent.

We reiterate that with only two assumptions—that $P(\mathbf{x}_i)$ is a GMM and that the sigmoid function can be approximated as the cdf of a Gaussian—all requisite integrals have been computed analytically. As a result, the log-likelihood can be easily maximized to find the logistic regression classifier $\mathbf{w}$ in the presence of missing data. Thereafter, the class predictions of an unlabeled testing data point with incomplete (missing) features can also be computed trivially using (16).

3. Estimating GMM from Incomplete Data

3.1. Expectation Maximization (EM)

In (Ghahramani & Jordan, 1994), the algorithm for estimating a GMM from incomplete data was given. However, that derivation admittedly assumed equal priors for the Gaussians. To conserve space, we shall show only the update equations for the general case.

The complete-data likelihood function for $\mathbf{x}_i$ is

$$\begin{aligned} &P\left(\mathbf{x}_i^{o_i}, \mathbf{x}_i^{m_i}, \gamma_i = k|\Theta\right) \\ &= P\left(\gamma_i = k|\Theta\right) P\left(\mathbf{x}_i^{o_i}, \mathbf{x}_i^{m_i}|\gamma_i = k, \Theta\right) \\ &= \pi_k \frac{1}{\sqrt{(2\pi)^d \left|\boldsymbol{\Sigma}_k\right|}} \exp\left(-\frac{1}{2}\left[\begin{array}{c} \mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i} \\ \mathbf{x}_i^{m_i} - \boldsymbol{\mu}_k^{m_i} \end{array}\right]^T \right. \\ &\left. \times \left[\begin{array}{cc} \boldsymbol{\Sigma}_k^{-1,o_i o_i} & \boldsymbol{\Sigma}_k^{-1,o_i m_i} \\ \boldsymbol{\Sigma}_k^{-1,m_i o_i} & \boldsymbol{\Sigma}_k^{-1,m_i m_i} \end{array}\right] \left[\begin{array}{c} \mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i} \\ \mathbf{x}_i^{m_i} - \boldsymbol{\mu}_k^{m_i} \end{array}\right]\right) \quad (18) \end{aligned}$$

where $\gamma_i = k$ denotes that $\mathbf{x}_i$ is generated by the k-th Gaussian of the GMM, and $\Theta = \{\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$.

For the data points $\{\mathbf{x}_i\}_{i=1}^N$, which are independent, the update equations for a K-component GMM are

$$\pi_k = \frac{1}{N} \sum_{i=1}^N \delta_k^i \quad (19)$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^N \delta_k^i \left[\begin{array}{c} \mathbf{x}_i^{o_i} \\ \widehat{\boldsymbol{\xi}}_k^i \end{array}\right]}{\sum_{i=1}^N \delta_k^i} \quad (20)$$

$$\begin{aligned} \boldsymbol{\Sigma}_k = \frac{1}{\sum_{i=1}^N \delta_k^i} \sum_{i=1}^N \delta_k^i \left\{ \left(\left[\begin{array}{c} \mathbf{x}_i^{o_i} \\ \widehat{\boldsymbol{\xi}}_k^i \end{array}\right] - \boldsymbol{\mu}_k \right) \right. \\ \left. \times \left(\left[\begin{array}{c} \mathbf{x}_i^{o_i} \\ \widehat{\boldsymbol{\xi}}_k^i \end{array}\right] - \boldsymbol{\mu}_k \right)^T + \left[\begin{array}{cc} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{\Omega}}_k^i \end{array}\right] \right\} \quad (21) \end{aligned}$$

where parameters from the previous iteration are denoted with hats and

$$\delta_k^i = \frac{\mathcal{N}\left(\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\mu}}_k^{o_i}, \widehat{\boldsymbol{\Sigma}}_k^{o_i o_i}\right) \widehat{\pi}_k}{\sum_{\ell=1}^K \mathcal{N}\left(\mathbf{x}_i^{o_i}; \widehat{\boldsymbol{\mu}}_\ell^{o_i}, \widehat{\boldsymbol{\Sigma}}_\ell^{o_i o_i}\right) \widehat{\pi}_\ell} \quad (22)$$

$$\widehat{\boldsymbol{\xi}}_k^i = \widehat{\boldsymbol{\mu}}_k^{m_i} + \widehat{\boldsymbol{\Sigma}}_k^{m_i o_i} \widehat{\boldsymbol{\Sigma}}_k^{o_i o_i\ -1} \left(\mathbf{x}_i^{o_i} - \widehat{\boldsymbol{\mu}}_k^{o_i}\right) \quad (23)$$

$$\widehat{\boldsymbol{\Omega}}_k^i = \widehat{\boldsymbol{\Sigma}}_k^{m_i m_i} - \widehat{\boldsymbol{\Sigma}}_k^{m_i o_i} \widehat{\boldsymbol{\Sigma}}_k^{o_i o_i\ -1} \widehat{\boldsymbol{\Sigma}}_k^{m_i o_i\ T}. \quad (24)$$

3.2. Variational Bayesian EM

Variational methods provide a lower bound on the log marginal likelihood. In our GMM problem with incomplete data, $\mathbf{x}_i^{o_i}$ are the observed variables, $\Phi = \{\mathbf{x}_i^{m_i}, \gamma_i\}$ is the set of hidden variables, and $\Theta = \{\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is the set of parameters. The log marginal likelihood of $\mathbf{x}_i^{o_i}$ can be lower bounded by writing (Beal & Ghahramani, 2003)

$$\begin{aligned} \ln P(\mathbf{x}_i^{o_i}) &= \ln \int P(\mathbf{x}_i^{o_i}, \Phi, \Theta) d\Phi d\Theta \\ &= \ln \int q(\Phi, \Theta) \frac{P(\mathbf{x}_i^{o_i}, \Phi, \Theta)}{q(\Phi, \Theta)} d\Phi d\Theta \\ &\geq \int q(\Phi, \Theta) \ln \frac{P(\mathbf{x}_i^{o_i}, \Phi, \Theta)}{q(\Phi, \Theta)} d\Phi d\Theta \quad (25) \\ &= \int q(\Phi) q(\Theta) \ln \frac{P(\mathbf{x}_i^{o_i}, \Phi, \Theta)}{q(\Phi) q(\Theta)} d\Phi d\Theta \quad (26) \end{aligned}$$

where (25) follows from Jensen's inequality, and (26) is the result of making the factorized approximation $q(\Phi, \Theta) \approx q(\Phi) q(\Theta)$. The variational Bayesian algorithm maximizes (26) with respect to the distributions $q(\Phi)$ and $q(\Theta)$. Since these two distributions are coupled, functional derivatives with respect to each distribution are iteratively taken while the opposite distribution is held fixed. The resulting Variational Bayesian Expectation (VB-E) and Maximization (VB-M) steps are respectively

$$q(\Phi) \propto \exp\left\{\int \ln P(\mathbf{x}_i^{o_i}, \Phi|\Theta) q(\Theta) d\Theta\right\} \quad (27)$$

$$q(\Theta) \propto P(\Theta) \exp\left\{\int \ln P(\mathbf{x}_i^{o_i}, \Phi|\Theta) q(\Phi) d\Phi\right\}. \quad (28)$$

The algorithm for estimating a GMM from complete data using the VB-EM algorithm has been done previously (*e.g.,* (Nasios & Bors, 2003)), but the incomplete-data version has not. We have derived the algorithm for this new case, dealing with incomplete data. To conserve space, we shall give only the relevant update equations, and other equations necessary for their interpretation.

To establish notation, the complete-data likelihood function for $\mathbf{x}_i$ is given in (18).

For the GMM, we choose conjugate-exponential priors for tractability (Nasios & Bors, 2003); that is, we choose a Dirichlet distribution on the mixing coefficients (π), normal distributions on the means ($\boldsymbol{\mu}_k$), and Wishart distributions on the precisions (inverse covariances, $\boldsymbol{\Sigma}_k^{-1}$). The prior distribution of the GMM parameters is therefore

$$P(\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = D(\pi) \prod_{k=1}^K N(\boldsymbol{\mu}_k) W\left(\boldsymbol{\Sigma}_k^{-1}\right) \quad (29)$$

where

$$D(\pi) = Z_{\pi}^{-1} \prod_{k=1}^{K} \pi_k^{\lambda_k^0 - 1} \tag{30}$$

$$N(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) = Z_{\boldsymbol{\mu}_k}^{-1} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\mu}_k - \mathbf{m}_k^0\right)^T \times \beta_k^{0,-1} \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{\mu}_k - \mathbf{m}_k^0\right) \right\} \tag{31}$$

$$W\left(\boldsymbol{\Sigma}_k^{-1}\right) = Z_{\boldsymbol{\Sigma}_k}^{-1} \left|\boldsymbol{\Sigma}_k^{-1}\right|^{\left(\alpha_k^0 - d - 1\right)/2} \times \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\mathbf{S}_k^{0,-1} \boldsymbol{\Sigma}_k^{-1}\right) \right\} \tag{32}$$

with normalization constants

$$Z_{\pi} = \frac{\prod_{k=1}^{K} \Gamma\left(\lambda_k^0\right)}{\Gamma\left(\sum_{k=1}^{K} \lambda_k^0\right)} \tag{33}$$

$$Z_{\boldsymbol{\mu}_k} = (2\pi)^{d/2} \left|\beta_k^0 \boldsymbol{\Sigma}_k\right|^{1/2} \tag{34}$$

$$Z_{\boldsymbol{\Sigma}_k} = 2^{\alpha_k^0 d/2} \pi^{d(d-1)/4} |\mathbf{S}_k^0|^{\alpha_k^0/2} \prod_{j=1}^{d} \Gamma\left(\frac{\alpha_k^0 + 1 - j}{2}\right). \tag{35}$$

In (30)–(35), λ_k^0, $\mathbf{m}_k^0$, β_k^0, α_k^0, and $\mathbf{S}_k^0$ are parameters of the priors, and Γ is the gamma function.

The update formulas of the incomplete-data GMM posterior parameters are

$$\lambda_k^{new} = \lambda_k^0 + \sum_{i=1}^{N} \widetilde{\delta}_k^i \tag{36}$$

$$\mathbf{m}_k^{new} = \frac{\beta_k^{0,-1} \mathbf{m}_k^0 + \sum_{i=1}^{N} \widetilde{\delta}_k^i \widetilde{\mathbf{x}}_i^k}{\beta_k^{0,-1} + \sum_{i=1}^{N} \widetilde{\delta}_k^i} \tag{37}$$

$$\beta_k^{new,-1} = \beta_k^{0,-1} + \sum_{i=1}^{N} \widetilde{\delta}_k^i \tag{38}$$

$$\alpha_k^{new} = \alpha_k^0 + \sum_{i=1}^{N} \widetilde{\delta}_k^i \tag{39}$$

$$\mathbf{S}_k^{new,-1} = \mathbf{S}_k^{0,-1} + \sum_{i=1}^{N} \widetilde{\delta}_k^i \widetilde{\mathbf{S}}_k^i + \beta_k^{0,-1} \mathbf{m}_k^0 \mathbf{m}_k^{0\,T} + \sum_{i=1}^{N} \widetilde{\delta}_k^i \widetilde{\mathbf{x}}_i^k \widetilde{\mathbf{x}}_i^{k\,T} - \left(\beta_k^{0,-1} + \sum_{i=1}^{N} \widetilde{\delta}_k^i\right) \overline{\mathbf{x}}_i^k \overline{\mathbf{x}}_i^{k\,T} \tag{40}$$

where

$$\overline{\mathbf{x}}_i^k = \frac{\beta_k^{0,-1} \mathbf{m}_k^0 + \sum_{i=1}^{N} \widetilde{\delta}_k^i \widetilde{\mathbf{x}}_i^k}{\beta_k^{0,-1} + \sum_{i=1}^{N} \widetilde{\delta}_k^i} \tag{41}$$

$$\widetilde{\mathbf{x}}_i^k = \begin{bmatrix} \mathbf{x}_i^{o_i} \\ \mathbf{m}_k^{m_i|o_i} \end{bmatrix} \tag{42}$$

$$\widetilde{\mathbf{S}}_k^i = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_k^{m_i|o_i} \end{bmatrix} \tag{43}$$

$$\widetilde{\delta}_k^i = \frac{A_k \mathcal{N}\left(\mathbf{x}_i^{o_i}; \mathbf{m}_k^{o_i}, \alpha_k^{-1} \mathbf{S}_k^{-1,o_i o_i}\right)}{\sum_{\ell=1}^{K} A_\ell \mathcal{N}\left(\mathbf{x}_i^{o_i}; \mathbf{m}_\ell^{o_i}, \alpha_\ell^{-1} \mathbf{S}_\ell^{-1,o_i o_i}\right)} \tag{44}$$

$$\mathbf{m}_k^{m_i|o_i} = \mathbf{m}_k^{m_i} + \mathbf{S}_k^{-1,m_i o_i} \mathbf{S}_k^{-1,o_i o_i\,-1} \left(\mathbf{x}_i^{o_i} - \mathbf{m}_k^{o_i}\right) \tag{45}$$

$$\mathbf{S}_k^{m_i|o_i} = \alpha_k^{-1} \left(\mathbf{S}_k^{-1,m_i m_i} - \mathbf{S}_k^{-1,m_i o_i} \mathbf{S}_k^{-1,o_i o_i\,-1} \mathbf{S}_k^{-1,m_i o_i\,T}\right) \tag{46}$$

$$A_k = \exp\left(\psi\left(\lambda_k\right) - \psi\left(\sum_{k=1}^{K} \lambda_k\right) + \frac{1}{2} \sum_{j=1}^{d} \psi\left(\frac{\alpha_k + 1 - j}{2}\right) + \frac{1}{2} d \ln 2 - \frac{1}{2} \ln \alpha_k - \mathrm{tr}\left(\beta_k \mathbf{I}_d\right)\right) \tag{47}$$

and where ψ is the digamma function.

4. Results

The main goal of this work is to develop a principled means of extending the logistic regression classifier for the case of missing data. Since the GMM density estimation plays a major role in the algorithm, an auxiliary goal was to compare the performance of the VB-EM and EM algorithms in estimating the GMM. To accomplish this secondary goal we created a 2-D toy data set, defined by a mixture of four Gaussians. We randomly removed 40% of the features, and then built GMMs using the VB-EM and EM algorithms. For each number of samples used to train the GMM, fifty trials were run. Each trial consisted of different data generated from the true GMM and different patterns of missing features.

An approximation to the Kullback-Leibler (KL) divergence between two Gaussian mixture models can be computed analytically using the unscented transform (Goldberger, Greenspan & Gordon, 2003). The smaller the KL divergence, the closer the estimated distribution is to the true distribution. The difference between the VB-EM and EM algorithms is most pronounced when a small amount of data is available to build the GMMs, in which case the VB-EM GMM is superior (see Figure 1).

The area under a receiver operating characteristic (ROC) curve is given by the Wilcoxon statistic (Hanley & McNeil, 1982)

$$A = (MN)^{-1} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{1}_{x_m > y_n} \tag{48}$$

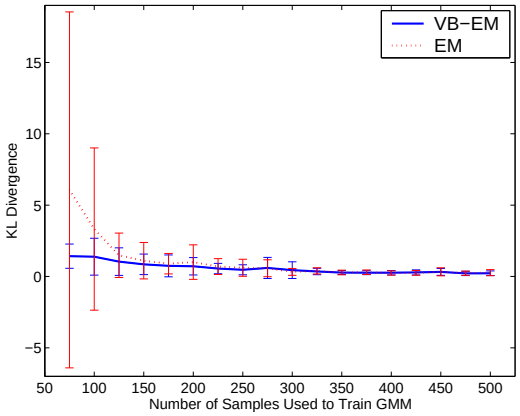

Figure 1. Approximate KL divergence between a toy (known) GMM and the estimated GMMs using VB-EM and EM.

where $x_1, \ldots, x_M$ are the classifier decisions of data belonging to class 1, $y_1, \ldots, y_N$ are the classifier decisions of data belonging to class -1, and $\mathbf{1}$ is an indicator function.

To examine our main goal, the proposed missing-data logistic regression algorithm was applied to the Ionosphere and Wisconsin Diagnostic Breast Cancer (WDBC) data sets (from the UCI Machine Learning Repository). The Ionosphere data set has 351 data points and 34 features, while the WDBC data set has 569 data points and 30 features. Experimental results are shown in Figures 2 and 3 in terms of the area under the ROC curves, computed from (48). To allow one to observe the performance of the methods as a function of data set size, the GMMs are trained using only training (labeled) data. Since the GMMs do not require labels, in practice all available data (labeled and unlabeled) can be used to build the GMMs.

Each point on every curve in Figure 2 is an average over ten trials. Every trial consists of a random partition of training and testing data, and a random pattern of missing features, the amounts of which are determined by the given parameters. Because both the training sets as well as the patterns of missing features in every trial are unique, performance can vary widely between trials. The relative differences between two methods over all trials vary less. That is, the methods have a consistent relative difference in performance, even though the absolute difference in performance may vary widely from trial to trial. Therefore we report in Table 1 the standard deviation of the *difference* between the proposed method using VB-EM for the GMM and each of the other methods considered. In the table (abbreviated to conserve space), positive differences indicate that the proposed method using VB-EM performed better.

Supplementary results in Figure 3 were obtained by following the same experimental setup as that used for the results in Figure 2.

From Figure 2, it can be observed that the proposed method using VB-EM for the GMM estimation consistently performed better than the same method using EM for the GMM estimation. In particular, this difference was most significant when a small number of data points were available to train the GMM (*cf.* Figure 1 also). We also observed that both of these versions of the proposed method were superior to the three single imputation schemes considered. For the proposed method using VB-EM, having fewer training data points with a higher fraction of features present appears to be more important (in terms of performance) than having more training data points with a lower fraction of features present (*e.g.,* when the fraction of training data points is 0.2, 0.3, and 0.6 in Figures 2(a), 2(b), and 2(c), respectively, the training set has the same total number of present features).

The incomplete-data problem, and in particular our proposed approach using GMMs, raises several interesting questions. For instance, the number of data points required to accurately estimate the GMM will increase as the square of the feature dimension because the covariance matrix is modeled. In contrast, the number of parameters in the standard logistic regression is equal to the feature dimension. Despite this ostensibly increased data set size requirement, our proposed algorithm using the VB-EM GMM still performs better than single imputation schemes when the number of training data points is small. For example in Figure 2, when the fraction of training data points is 0.1 (corresponding to only 35 training data points, each of which have 34 features), our proposed algorithm still outperforms the single imputation methods. This result suggests that the benefits of our algorithm outweigh the added parameter estimation burden.

Another question the incomplete-data problem raises is whether ignoring data with missing features is better than using an incomplete-data method (either our proposed method or even a simple imputation scheme). It is of course displeasing to discard data (information), but can doing so improve performance? There is a major problem with simply ignoring data with missing features. It is true that ignoring data with missing features in the training stage will eliminate incomplete-data training issues. However, in the testing stage,

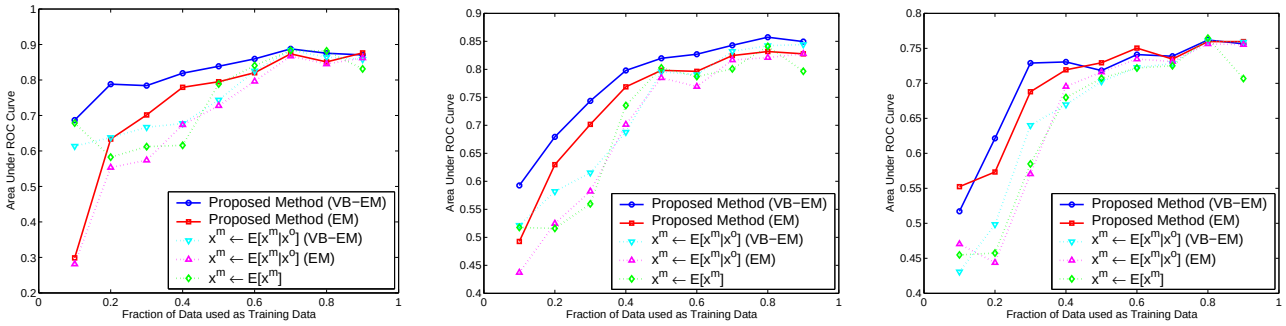

Figure 2. Experimental results on the IONOSPHERE data set. The proposed methods use the new logistic regression method (no imputation), with the requisite GMMs trained using the VB-EM or EM algorithm. The other three methods complete the missing data via imputation using the conditional mean (obtained via the VB-EM or EM GMMs) or the unconditional mean. The results are for the cases when (a) 25%, (b) 50%, and (c) 75% of the features are missing.

Table 1. Abbreviated summary of variance (error bars) for Figure 2 results. Values in the table are the mean ± one standard deviation of the following difference: the area under the ROC curve of each listed method subtracted from the area under the ROC curve of the proposed method using VB-EM for the GMM. A(VB) and A(EM) are the proposed methods using the VB-EM and EM GMMs, respectively; A(μ_V^C), A(μ_E^C), and A(μ^U) complete the missing data via imputation using the conditional mean (obtained via the VB-EM or EM GMMs) or the unconditional mean, respectively.

PERCENTAGE OF MISSING FEATURES	FRACTION OF DATA USED TO TRAIN	A(VB) – A(EM)	A(VB) – A(μ_V^C)	A(VB) – A(μ_E^C)	A(VB) – A(μ^U)
25	0.1	0.3881 ± 0.1504	0.0735 ± 0.0667	0.4056 ± 0.1706	0.0082 ± 0.1493
25	0.3	0.0826 ± 0.0447	0.1172 ± 0.0591	0.2103 ± 0.0521	0.1720 ± 0.0991
25	0.7	0.0141 ± 0.0448	0.0041 ± 0.0171	0.0206 ± 0.0459	0.0045 ± 0.0400
25	0.9	-0.0058 ± 0.0863	0.0126 ± 0.0167	0.0085 ± 0.0878	0.0394 ± 0.0708
50	0.1	0.1000 ± 0.0772	0.0711 ± 0.0279	0.1555 ± 0.1031	0.0750 ± 0.1302
50	0.3	0.0419 ± 0.0494	0.1284 ± 0.0736	0.1616 ± 0.1369	0.1840 ± 0.0925
50	0.7	0.0183 ± 0.0401	0.0109 ± 0.0121	0.0263 ± 0.0445	0.0421 ± 0.0368
50	0.9	0.0218 ± 0.1149	0.0053 ± 0.0089	0.0218 ± 0.1104	0.0530 ± 0.0948
75	0.1	-0.0352 ± 0.1718	0.0865 ± 0.0621	0.0467 ± 0.1647	0.0623 ± 0.1814
75	0.3	0.0410 ± 0.0532	0.0890 ± 0.0533	0.1584 ± 0.0734	0.1441 ± 0.0982
75	0.7	0.0041 ± 0.0319	0.0105 ± 0.0103	0.0076 ± 0.0336	0.0136 ± 0.0348
75	0.9	-0.0031 ± 0.0610	-0.0023 ± 0.0241	0.0012 ± 0.0751	0.0496 ± 0.0543

one cannot simply ignore a data point to be classified because it is missing some features. One would still be forced to resort to ad hoc procedures such as filling in zeros or the unconditional mean for the missing features of such incomplete testing data. In contrast, our proposed method is completely principled, and does not rely on any ad hoc methods in either the training or testing stage.

5. Conclusion

Our main contribution has been the derivation of a missing-data logistic regression classification algorithm. By making two mild assumptions, the algorithm solves the incomplete-data problem in a principled manner, avoiding imputation heuristics. Experimental results have shown the proposed classifier to be superior to commonly used single-imputation methods. The proposed algorithm has also been successful even when a high percentage of features are missing. Moreover, despite the additional parameters to be estimated, the proposed algorithm has been successful when the training set size is small. The update equations for building a GMM with incomplete data via the EM and VB-EM algorithms have also been given. The extension of the VB-EM GMM algorithm to the case of incomplete data is also a new contribution. Experimental evidence has shown that the VB-EM algorithm is markedly superior in terms of density estimation when data is scarce.

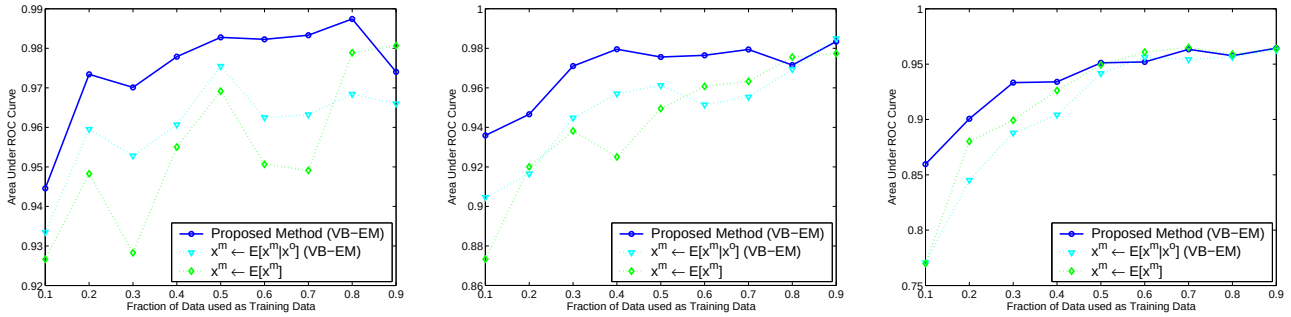

Figure 3. Experimental results on the WDBC data set. Refer to Figure 2 caption for legend details. The results are for the cases when (a) 25%, (b) 50%, and (c) 75% of the features are missing.

Future work will examine the possibility that a classifier constructed from incomplete data can outperform a classifier constructed from complete data. If features that decrease or confuse class separation are missing, this hypothesis may be true. This view also suggests potential links to feature selection. Additional work will also investigate the use of Dirichlet processes to address choosing the number of components for the GMM.

References

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Doctoral dissertation, Gatsby Computational Neuroscience Unit, University College London.

Beal, M. & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: application to scoring graphical model structures. *Bayesian Statistics 7*, 453–464.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B 39*, 1–38.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification*. New York: Wiley.

Ghahramani, Z. & Jordan, M. (1994). Supervised learning from incomplete data via the EM approach. In J. Cowan and G. Tesauro and J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6*. San Mateo, CA: Morgan Kaufmann.

Goldberger, J., Greenspan, H., & Gordon, S. (2003). An efficient similarity measure based on approximations of KL-divergence between two Gaussian mixtures. *International Conference on Computer Vision (ICCV)*.

Graepel, T. (2002). Kernel matrix completion by semidefinite programming. *Proceedings of the International Conference on Artificial Neural Networks* (pp. 694–699).

Hanley, J. & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*, 29–36.

Ibrahim, J. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association 85*, 765–769.

Lanckriet, G., Deng, M., Cristianini, N., Jordan, M., & Noble, W. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. *Proceedings of the Pacific Symposium on Biocomputing 9* (pp. 300–311).

Nasios, N. & Bors, A. (2003). Variational expectation-maximization training for Gaussian networks. *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing* (pp. 339–348).

Rässler, S. (2004). *The impact of multiple imputation for DACSEIS* (DACSEIS Research Paper Series 5). University of Erlangen-Nürnberg, Nürnberg, Germany.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Tsuda, K., Akaho, S., & Asai, K. (2003). The *em* algorithm for kernel matrix completion with auxiliary data. *Journal of Machine Learning Research 4*, 67–81.