Incomplete-Data Classification using Logistic Regression

David Williams
dpw@ee.duke.edu
Xuejun Liao
xjliao@ee.duke.edu
Department of Electrical and Computer Engineering,
Duke University

The incomplete-data problem exists in a wide range of fields like social sciences, computer vision, remote sensing, etc. For e.g. social science data collection, certain people may refrain from giving information, or during sensing, certain sensors might not work correctly and this may lead to incomplete data. These incomplete-data problems are often solved by imputations i.e. by completing missing data by filling in specific values. Some common imputation schemes include filling missing data with zeroes, or unconditional/conditional mean.

We show in this paper how a logistic regression classification algorithm can be implemented for filling out the missing values. We perform analytic integration with an estimated conditional density function. The conditional density functions are estimated using Gaussian Matrix Model (GMM) with parameter estimation performed using two different approaches:

1) Expectation Maximization(EM)

2) Variational Bayesian EM(VB-EM)

The paper; using available real data also demonstrates how using the (VB-EM) method is better than (EM) method for handling the incomplete data. Also the paper demonstrated how the proposed approach is better than standard imputation procedures.

It is important to integrate out missing data for calculating the posterior distribution of a parameter.

$$P(y_i|x_i^{oi}) = \int P(y_i|x_i^{mi}, x_i^{oi}) P(x_i^{mi}|x_i^{oi}) dx_i mi \tag{1}$$

$$P(x_i^m i|x_i^o i) \tag{2}$$

The Eqn(2) represents the estimate of missing features given the observed features.

The integral in Eqn(1) is difficult to solve in general, but in case of logistic regression the integral can be solved analytically using two assumptions:

1) Eqn(1) is a Gaussian Mixture Model (GMM)

2) The sigmoid function can be approximated as the cumulative distribution function (cdf) of a Gaussian. It is well known that a mixture of Gaussians can approximate any distribution. since we can sole the integral analytically, the likelihood can be maximized in a manner similar to that of a complete data set.

Since we know GMM plays an important role; we use two different methods to accurately perform GMM density estimation in presence of missing data.These two methods are the one mentioned above (EM) and (VB-EM).

Logistic Regression for incomplete data:

$$P(y_i|x_i^{oi}) = \int \sigma(y_i|w_{oi}^T x_i^{oi} + v_i) P(v_i|x_i^{oi}) dv_i \tag{3}$$

$$P(x_i) = \sum_{k=1}^{K} \pi_k N(\begin{bmatrix} x_i^{oi} \\ x_i^{mi} \end{bmatrix} \mu_k; \sigma_k) \tag{4}$$

$$v_i = w_{mi}^T x_i^{mi}, P(v_i|x_i^{oi}) \tag{5}$$

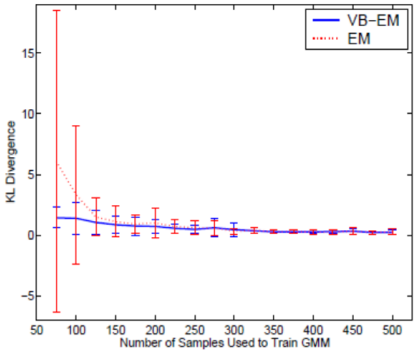
$$\sigma(v) = \int_{-\infty}^{v} G(z/\beta) dz \tag{6}$$

$$l(w) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_k^i \sigma(\frac{y_i \beta (w_{mi}^T \xi_k^i + w_{oi}^T x_i^{oi})}{\sqrt{w_{mi}^T \Omega_k^i w_{mi} + \beta^2}}) \tag{7}$$

In Eqn.(3); We have re-written Eqn.(2) by using the concepts of logistic Regression. Since we need to perform integration, therefore we assume Eqn.(4) as Gaussian Mixture Model (GMM) Also Eqn.(5) is GMM because of the linear relation. In Eqn.(6) we approximate the sigmoid function as the cdf of a Gaussian(i.e., a prohibit function). Finally We substitute all the values and assuming data points are independent of each other, we find the log likelihood function in Eqn.(7)

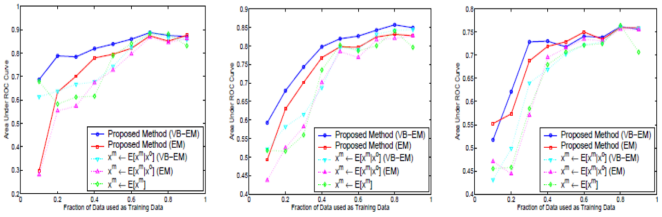
In the later half of the paper we perform parameter estimation of GMM using EM and VB-EM. In EM we perform maximum likelihood estimation in the presence of latent variables. We do this by first finding values for latent variables and then optimizing the model. We repeat these two steps until convergence. VB-EM is not much different, it just provides a lower bound on log marginal likelihood.

Since GMM density estimation plays an important role, another goal of the paper was also to compare EM and VB-EM algorithms in estimating GMM. An approximation to Kullback-Leibler(KL) divergence between the two GMMs is computed analytically. Smaller the KL divergence, closer is the estimated distribution to the true distribution. Also the difference between two models is much pronounced when a small amount of data is available to build GMMs; in such case VB-EM performs superior than EM. Refer the below diagram for results of the experiment.



Our main goal was to examine the model proposed in the paper against standard imputation techniques used to fill the missing data. Refer to the below figure for the results of the experiment. Here, every point on the curve is an average of over ten trials. In the figure it is visible that VB-EM consistently performed better than EM. Also both the proposed methods performed better than the tree imputation methods. Further the positive values for standard deviation of VB-EM method in the below suggests that VB-EM performed better.

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



1 References:

[1] Rubin, D. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987

[2] Duda, R., Hart, P., Stork, D. Pattern classification. New York: Wiley, 2000