Maximum Entropy Density Estimation with Incomplete Data

Bert Huang, Ansaf Salleb-Aouissi
Center for Computational Learning Systems
Columbia University, New York, NY 10115
{bert@cs, ansaf@ccls}.columbia.edu

We propose a natural generalization of Regularized Maximum Entropy Density Estimation (maxent) to handle input data with unknown values. While standard approaches to handling missing data usually involve estimating the actual unknown values, then using the estimated, complete data as input, our method avoids the two-step process and handles unknown values directly in the maximum entropy formulation.

The maxent method was recently proposed as an excellent method of presence-only prediction [2, 3]. In a presence-only framework, we are given a set, X, of data in which some of the data are labeled as positive. However, unlike the typical classification framework, the remaining unlabeled instances are not necessarily negative. Instead, they are considered of unknown class. The regularized maxent method treats the positively labeled points as random draws from some hidden distribution over X and attempts to estimate that distribution. Specifically, regularized maxent tries to find a distribution over X with maximum entropy such that the expected values of each feature are close to the observed means of the features with a positive label.

Let F be an $N \times D$ matrix of features such that F_{ij} is the i'th datum's j'th feature. Let vector m be the means of the D features of the labeled positive data. Then the standard regularized maxent optimization is:

$$\max_p -\sum_i p_i \ln p_i \quad s.t. \quad \sum_i p_i = 1, \quad \left| \sum_i p_i F_{ij} - m_j \right| \leq \beta_j, \ \forall j \tag{1}$$

Here, the β terms are regularization parameters. To handle the missing values, we first compute the empirical means $(m_1, \ldots, m_D)$ using only values we actually know. In particular, rather than imputing values for unknown features, we simply omit the unknowns from the averages. Next, we introduce an indicator matrix O such that entry O_{ij} is 1 if we know the value of F_{ij} and it is zero if F_{ij} is missing. In addition, WLOG, we set all unknown F_{ij} to zero. Now we can write term for a normalized expected value of a feature with missing values:

$$E[F_{\bullet j}] = \frac{\sum_i p_i F_{ij}}{\sum_i p_i O_{ij}} \tag{2}$$

As before, we want to keep these expectations close to the empirical means, while maximizing entropy. This gives our proposed optimization:

$$\max_p - \sum_i p_i \ln p_i \quad s.t. \quad \sum_i p_i = 1, \quad \left| \frac{\sum_i p_i F_{ij}}{\sum_i p_i O_{ij}} - m_j \right| \leq \beta_j, \forall j \tag{3}$$

There are various methods to solve this optimization, but as an initial implementation we performed a dual optimization similar to [2] but using a Newton update to optimize each dimension iteratively. Experimentally, this method converged to the maximum for non-degenerate cases.

We compared the performance of our algorithm against running standard regularized maxent with either mean imputation or Gaussian EM imputation [4]. We hand picked four databases from the UCI ML Repository [1] that had real missing features and removed the most complete half of the features to exacerbate the incompleteness. Giving each algorithm a training subset of the positive class, we compare their area under ROC curves with respect to the remaining testing positives. Table 1 lists the results, with the highest AUC in bold.

While our method's performance on these datasets is only slightly better, we find its simplicity and elegance attractive. The algorithm makes no attempt to solve a harder problem than it is given, which seems to follow nicely from the maximum entropy principle itself.

	Proposed	Mean Imputation	EM Imputation
horse-colic	**0.7853± 0.0316**	0.7770±0.0306	0.7748±0.0296
house-votes-84	0.7416± 0.0182	**0.7430 ± 0.0176**	0.7420± 0.0182
echocardiogram	**0.7079± 0.0421**	0.6948± 0.0410	0.6909± 0.0389
hepatitis	**0.8541± 0.0367**	0.8467± 0.0379	0.8496± 0.0381

Table 1: Average and standard deviation of best AUC after cross-validation for β parameters. Averaged over 50 random splits of training and testing positive labeled points.

References

[1] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998.

[2] Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *COLT*, pages 472–486, 2004.

[3] Steven J. Phillips, Miroslav Dudík, and Robert E. Schapire. A maximum entropy approach to species distribution modeling. In *ICML*, 2004.

[4] T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14:853871, 2001.