Learning naïve Bayes regression models with missing data using mixtures of truncated exponentials

Antonio Fernández
Department of Statistics & Applied Mathematics
University of Almería
04120 Almería, Spain
afalvarez@ual.es

Jens D. Nielsen
Computer Science Department
University of Castilla-La Mancha
02071 Albacete, Spain
dalgaard@dsi.uclm.es

Antonio Salmerón
Department of Statistics & Applied Mathematics
University of Almería
04120 Almería, Spain
antonio.salmeron@ual.es

Abstract

In the last years, mixtures of truncated exponentials (MTEs) have received much attention within the context of probabilistic graphical models, as they provide a framework for hybrid Bayesian networks which is compatible with standard inference algorithms and no restriction on the structure of the network is considered. Recently, MTEs have also been successfully applied to regression problems in which the underlying network structure is a naïve Bayes or a TAN. However, the algorithms described so far in the literature operate over complete databases. In this paper we propose an iterative algorithm for constructing naïve Bayes regression models from incomplete databases. It is based on a variation of the data augmentation method in which the missing values of the explanatory variables are filled by simulating from their posterior distributions, while the missing values of the response variable are generated from its conditional expectation given the explanatory variables. We illustrate through a set of experiments with various databases that the proposed algorithm behaves reasonably well.

1 Introduction

In the last years, mixtures of truncated exponentials (MTEs) (Moral et al., 2001) have received much attention within the context of probabilistic graphical models, as they provide a framework for hybrid Bayesian networks which is compatible with standard inference algorithms and no restriction on the structure of the network is imposed (Cobb and Shenoy, 2006; Rumí and Salmerón, 2007). Recently, MTEs have also been successfully applied to regression problems in which the underlying network structure is a naïve Bayes (Morales et al., 2007) or a tree augmented naïve Bayes (TAN) (Fernández et al., 2007). However, the algorithms described so far in the literature operate over complete databases. In this paper we propose an iterative algorithm for constructing naïve Bayes regression models from incomplete databases. It is based on a variation of the data augmentation method (Tanner and Wong, 1987) in which the missing values of the explanatory variables are filled by simulating from their posterior distributions, while the missing values of the response variable are generated from its conditional expectation given the explanatory variables.

The rest of the paper is organised as follows. The MTE model, which is the basis of our work, is described in Sec. 2. We analyse the background behind existing regression models using

MTEs in Sec. 3, and out of that analysis, we describe our new algorithm that operates over missing values. The behaviour of the algorithm is tested through two experiments in Sec. 4. The paper ends with the concluding remarks in Sec. 5.

2 The MTE model

We denote random variables by capital letters, and their values by lowercase letters. We use boldfaced characters to represent random vectors and their values. The support of the variable $\mathbf{X}$ is denoted by $\Omega_{\mathbf{X}}$. A potential of class MTE (Moral et al., 2001) is defined as follows:

Definition 1. (MTE potential) Let $\mathbf{X}$ be a mixed n-dimensional random vector. Let $\mathbf{W} = (W_1, \ldots, W_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of $\mathbf{X}$, respectively, with $c + d = n$. We say that a function $f : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ is a *Mixture of Truncated Exponentials potential (MTE potential)* if for each fixed value $\mathbf{w} \in \Omega_{\mathbf{W}}$ of the discrete variables $\mathbf{W}$, the potential over the continuous variables $\mathbf{Z}$ is defined as:

$$f(\mathbf{z}) = a_0 + \sum_{i=1}^{m} a_i \exp\left\{\sum_{j=1}^{c} b_i^{(j)} z_j\right\} \quad (1)$$

for all $\mathbf{z} \in \Omega_{\mathbf{Z}}$, where a_i, $i = 0, \ldots, m$ and $b_i^{(j)}$, $i = 1, \ldots, m$, $j = 1, \ldots, c$ are real numbers. We also say that f is an MTE potential if there is a partition $D_1, \ldots, D_k$ of $\Omega_{\mathbf{Z}}$ into hypercubes and in each D_i, f is defined as in Eq. (1).

Definition 2. (MTE density) An MTE potential f is an *MTE density* if

$$\sum_{\mathbf{w} \in \Omega_{\mathbf{W}}} \int_{\Omega_{\mathbf{Z}}} f(\mathbf{w}, \mathbf{z}) d\mathbf{z} = 1 \ .$$

A *conditional MTE density* can be specified by dividing the domain of the conditioning variables and specifying an MTE density for the conditioned variable for each configuration of splits of the conditioning variables (Moral et al., 2001; Moral et al., 2003).

Example 1. Consider two continuous variables X and Y. A possible conditional MTE density for Y given X is the following:

$$f(y|x) = \begin{cases} 1.26 - 1.15e^{0.006y} \\ \quad \text{if } 0.4 \leq x < 5,\ 0 \leq y < 13 \ , \\ 1.18 - 1.16e^{0.0002y} \\ \quad \text{if } 0.4 \leq x < 5,\ 13 \leq y < 43 \ , \\ 0.07 - 0.03e^{-0.4y} + 0.0001e^{0.0004y} \\ \quad \text{if } 5 \leq x < 19,\ 0 \leq y < 5 \ , \\ -0.99 + 1.03e^{0.001y} \\ \quad \text{if } 5 \leq x < 19,\ 5 \leq y < 43 \ . \end{cases} \quad (2)$$

3 Regression using MTEs

Assume we have a set of variables $Y, X_1, \ldots, X_n$, where Y is continuous and the rest are either discrete or continuous. Regression analysis consists of finding a model g that explains the *response* variable Y in terms of the *explanatory* variables $X_1, \ldots, X_n$, so that given a configuration of the explanatory variables, $x_1, \ldots, x_n$, a prediction about Y can be obtained as $\hat{y} = g(x_1, \ldots, x_n)$. Previous work on regression using MTEs (Morales et al., 2007; Fernández et al., 2007) proceeds by representing the joint distribution of $Y, X_1, \ldots, X_n$ as a Bayesian network (naïve Bayes or TAN), and then using the posterior distribution of Y given $X_1, \ldots, X_n$ (more precisely, its expectation or its median) to obtain a prediction for Y.

3.1 Constructing a regression model from incomplete data

In this paper we will concentrate on the use of the expectation to analyse the regression problem with missing data. Therefore, our regression model will be

$$\hat{y} = g(x_1, \ldots, x_n) = E[Y|x_1, \ldots, x_n] = \int_{\Omega_Y} y f(y|x_1, \ldots, x_n) dy \ ,$$

where $f(y|x_1, \ldots, x_n)$ is the conditional density of Y given $x_1, \ldots, x_n$, which we assume to be of class MTE.

A conditional distribution of class MTE can be represented as in Eq. (2), where actually a marginal density is given for each element of the partition of the support of the variables involved. Within the context of regression, the distribution for the response variable Y given an element in a partition of the domain of the explanatory variables $X_1, \ldots, X_n$, can be regarded as an approximation of the true distribution of the actual values of Y for each possible configuration of the explanatory variables in that region of the partition. This fact justifies the selection of $E[Y|x_1, \ldots, x_n]$ as the predicted value for the regression problem, because that value is the one that best represents all the possible values of Y for that region, in the sense that it minimises the *mean squared error* between the actual value of Y and its predictions $\hat{y}$, namely

$$\text{mse} = \int_{\Omega_Y} (y - \hat{y})^2 f(y|x_1, \ldots, x_n) dy \ , \qquad (3)$$

which is known to be minimised for $\hat{y} = E[Y|x_1, \ldots, x_n]$. Therefore, the key point to find a regression model of this kind is to obtain a good estimation of the distribution of Y for each region of values of the explanatory variables. For the complete data case, the problem was studied in (Morales et al., 2007; Fernández et al., 2007), but the estimation of MTE distributions in the presence of missing data has not yet been addressed, but in the more restricted setting of unsupervised data clustering (Gámez et al., 2006). In that case, the only missing values are on the class variable, which is hidden, while the data about the features are complete.

Here we are interested in problems where the missing values can appear in the response variable as well as in the explanatory variables. A first approach to solve this problem could be to apply the EM algorithm (Dempster et al., 1977). However, the application of the methodology is problematic because the likelihood function for the MTE model cannot be optimised in an exact way (Rumí et al., 2006). Also, the aim of the EM algorithm is to find maximum likelihood estimates, which is not our main goal. From the point of view of regression, it is more important that the obtained models provide low values for the mean squared error rather than high likelihood.

Another way of approaching problems with missing values is the so-called *data augmentation* (DA) algorithm (Tanner and Wong, 1987). The advantage with respect to the EM algorithm is that DA does not require to directly optimise the likelihood function. Instead, it is based on imputing the missing values by simulating from the posterior distribution of the missing variables, which is iteratively improved from an initial estimation based on a random imputation. The DA algorithm leads to an approximation of the maximum likelihood estimates of the parameters of the model, as long as the parameters are estimated by maximum likelihood from the complete database in each iteration.

However, as we mentioned before, we are not so interested in the maximum likelihood estimates of the parameters of the model, but rather in reducing the mean squared error for the estimates of Y. With this aim, we show in the next proposition that using the conditional expectation of Y to impute the missing values instead of simulating values for Y (denoted as Y_S in the proposition), reduces the mse even if we simulate from the exact distribution of Y conditional on any configuration on a region of the values of the explanatory variables.

Proposition 1. *Let Y and Y_S be two continuous independent and identically distributed random variables. Then,*

$$E[(Y - Y_S)^2] \geq E[(Y - E[Y])^2] \ . \qquad (4)$$

Proof.

$$\begin{aligned} E[(Y - Y_S)^2] &= E[Y^2 + Y_S^2 - 2YY_S] \\ &= E[Y^2] + E[Y_S^2] - 2E[YY_S] \\ &= E[Y^2] + E[Y_S^2] - 2E[Y]E[Y_S] \\ &= 2E[Y^2] - 2E[Y]^2 \\ &= 2(E[Y^2] - E[Y]^2) = 2\text{Var}(Y) \\ &\geq \text{Var}(Y) = E[(Y - E[Y])^2] \ . \end{aligned}$$

□

In the proof we have used that both variables are independent and identically distributed, and therefore the expectation of the product is the product of the expectations, and the expected value of both variables is the same.

3.2 The algorithm for learning a NB regression model from incomplete data

Our proposal consists of an algorithm which iteratively learns a naïve Bayes regression model by imputing the missing values in each iteration according to the following criterion:

- If the missing value corresponds to the response variable, it is imputed with the conditional expectation of Y given the values of the explanatory variables in the same record of the database, computed from the current NB model.
- Otherwise, the missing cell is imputed by simulating the corresponding variable from its conditional distribution given the values of the other variables in the same record, computed from the current NB model.

As the imputation requires the existence of a model, more precisely a NB in our context, for the construction of the initial model we propose to impute the missing values by simulating from the marginal distribution of each variable computed from the observed values. In preliminary experiments we achieved better results using this alternative rather than pure random initialisation, which is the standard way of proceeding in data augmentation (Tanner and Wong, 1987). Another possibility is to simulate from the conditional distribution of each explanatory variable given the response, but the drawback is that the estimation of the conditional distributions requires more data than the estimation of the marginals, which can be problematic if the amount of missing values is high.

Therefore, the algorithm proceeds by imputing the initial database, learning an initial model and re-imputing the missing cells. Then, a new model is constructed and, if the mean squared error is reduced, the current model is

Algorithm 1: NB regression model from incomplete data

Input: An incomplete database D for variables $Y, X_1, \ldots, X_n$.
Output: A naïve Bayes regression model for response variable Y and explanatory variables $X_1, \ldots, X_n$.

for *each variable* $X \in \{Y, X_1, \ldots, X_n\}$ **do**
  Learn a univariate distribution $f_X(x)$ from its observed values in D.
end
Create a database D' from D by imputing the missing values for each $X \in \{Y, X_1, \ldots, X_n\}$ sampling from $f_X(x)$.
Create a database D_t from D by discarding the records where Y is missing.
Learn a NB regression model M' from D'.
Let $srmse'$ be the sample root mean squared error of M' computed using D_t according to Eq. (5).
$srmse \leftarrow \infty$.
while $srmse' < srmse$ **do**
  $M \leftarrow M'$.
  $srmse \leftarrow srmse'$.
  Create a new database D' from D filling the missing values as follows:
  for *each variable* $X \in \{X_1, \ldots, X_n\}$ **do**
    for *each record* $\mathbf{z}$ *in* D *with missing value for* X **do**
      Obtain $f_X(x|\mathbf{z})$ by probability propagation in model M.
      Impute the missing value for X by simulating from $f_X(x|\mathbf{z})$.
    end
  end
  for *each record* $\mathbf{z}$ *in* D *with missing value for* Y **do**
    Obtain $f_Y(x|\mathbf{z})$ by probability propagation in model M.
    Impute the missing value for Y with $E_{f_Y}[Y|\mathbf{z}]$.
  end
  Re-estimate model M' from D'.
  Let $srmse'$ be the sample root mean squared error of M' computed using D_t.
end
return M

replaced and the process repeated until convergence. As the mse in Eq. (3) requires the knowledge of the exact distribution of Y conditional on each configuration of the explanatory variables, we use as error measure the sample root mean squared error, computed as

$$\text{srmse} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \ , \qquad (5)$$

where m is the sample size, y_i is the true value of Y for record i and $\hat{y}_i$ is its corresponding prediction through the regression model.

The details are given in Alg. 1. Notice that, in steps 6 and 23 the naïve Bayes model is learnt from a complete database, and therefore the existing estimation methods for MTEs can be used (Rumí et al., 2006; Morales et al., 2007).

4 Experimental evaluation

In order to test the performance of the proposed method, we have carried out two experiments over four databases. One database (`mte50`) is synthetic, sampled from an MTE distribution, taken from (Morales et al., 2007). The other three databases are available in the UCI (Blake and Merz, 1998) and StatLib (StatLib, 1999) repositories. A description of the used databases can be found in Tab. 1.

Database	Size	# Cont.	# Disc.
`bodyfat`	251	15	0
`boston`	452	11	2
`cloud`	107	6	2
`mte50`	50	3	1

Table 1: A description of the databases used in the experiments, with their size and number of continuous and discrete variables.

The first experiment was oriented to test whether the model behaves reasonably, in the sense that the error is directly related to the percentage of missing values. With that aim, each database was divided at random into two parts, one for training with a 70% of the records, and one for test, with the remaining records. Then, we randomly inserted missing values in the training databases, ranging from a percentage of 10% to 50%. Previously, for each database, we repeated 100 times the same operation, obtaining the curves displayed in Fig. 1. The points correspond to the average srmse over the same test data by the 100 models learnt, and over each point there is a 95% confidence interval for the mean. The shape of the curves shows the expected behaviour, with the error increasing, in general, as the rate of missing grows.

The graphs in Fig. 2 show the log-likelihood corresponding to the learnt models as a function of the rate of missing values. Even though the goal of our algorithm is not to find highly likely models, the behaviour of the curves is still coherent. As in Fig. 1, the points indicate the average log-likelihood over the test database for the 100 runs of the experiment, and the intervals are 95% confidence intervals computed using the 100 measurements.

The second experiment was oriented to compare the proposed model with the M5' algorithm. The M5' algorithm (Wang and Witten, 1997) is an improved version of the model tree introduced by Quinlan (Quinlan, 1992). The model tree is basically a decision tree where the leaves contain a regression model rather than a single value, and the splitting criterion uses the variance of the values in the database corresponding to each node rather than the information gain. We chose the M5' algorithm because it was the state-of-the-art in graphical models for regression, before the introduction of MTEs for regression in (Morales et al., 2007). We have used the implementation of that method provided by Weka 3.4.11 (Witten and Frank, 2005). Regarding the implementation of the NB model, we have included it in the Elvira software (Elvira Consortium, 2002), which can be downloaded from `http://leo.ugr.es/elvira`.

In this experiment we have used 10-fold cross validation to estimate the srmse. The missing cells in the databases where selected before running the cross validation, therefore, in this case both the training and test databases contain missing cells in each iteration of the cross validation. We discarded from the test set the records for which the value of Y was missing. If the missing cells in the test set correspond to explanatory variables, algorithm M5' imputes

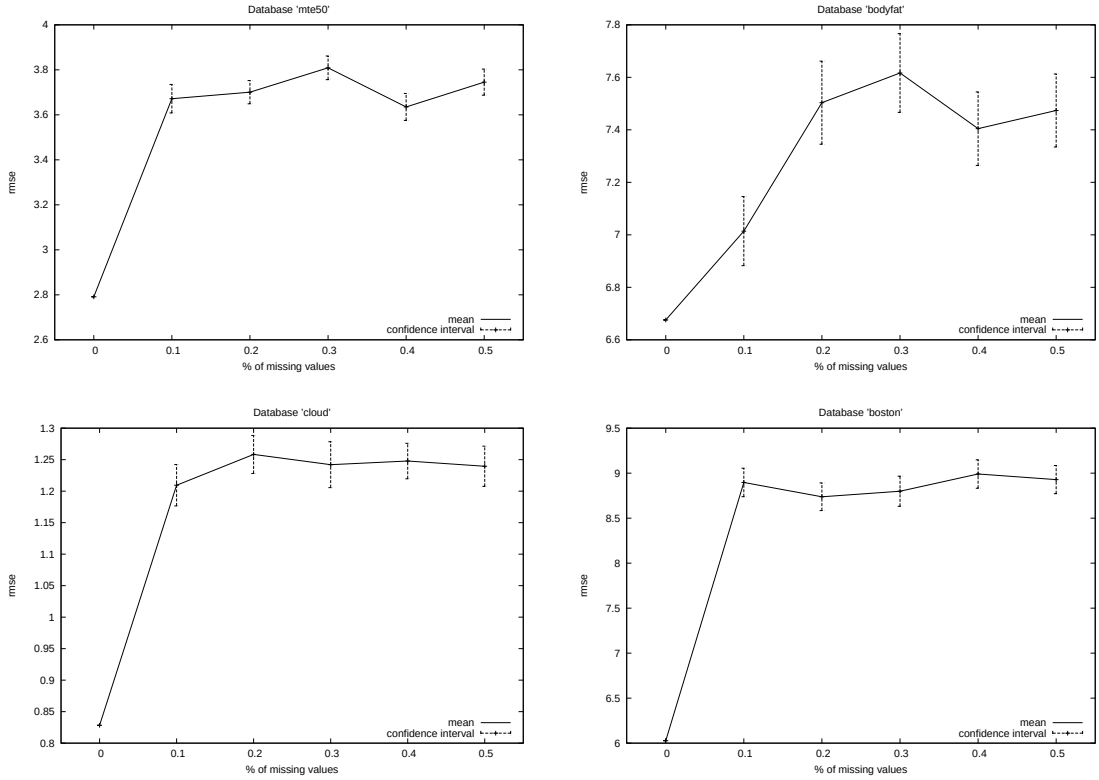

Figure 1: Sample root mean squared error as a function of the percentage of missing values.

them as the column average if the variable is continuous, or the column mode if it is qualitative (Witten and Frank, 2005). The NB model does not require the imputation of the missing explanatory variables in the test set, as the posterior distribution for Y is computed by probability propagation and therefore, the variables which are not observed are marginalised out. The results of the second experiment can be read in Tab. 2. The values displayed correspond to the average srmse computed by 10-fold cross validation. The result of the comparison is a draw, with NB winning for databases `bodyfat` and `mte50` and M5' winning in the other two cases. Friedman's test (Demsar, 2006) reports no statistically significant differences between both methods, with a p-value of 0.6831. This result was to be expected, as it is consistent with the comparison between models when they are learnt from complete datasets. It is surprising the error obtained by M5' for the database `bodyfat` with 50% of missing, which is much better than for lower rates of missing values. We think that this can be due to randomness.

4.1 Results discussion

The experiments carried out suggest that the proposed method behaves in a reasonable way. The graphs corresponding to the first experiment show a tendency of the error to increase along with the rate of missing values, except in some cases where it decreases around the 40% of missing, probably due to overfitting, as mentioned in (Friedman, 1997) for the general case of learning Bayesian networks. Also, we have similar graphs showing how the likelihood of the learnt models decrease as the rate of missing values increases.

Regarding the second experiment, the results are coherent with the ones obtained for the complete data case. However, we believe that our proposal should be superior to M5' in the case of learning from missing data. The reason is that we impute taking into account the condi-

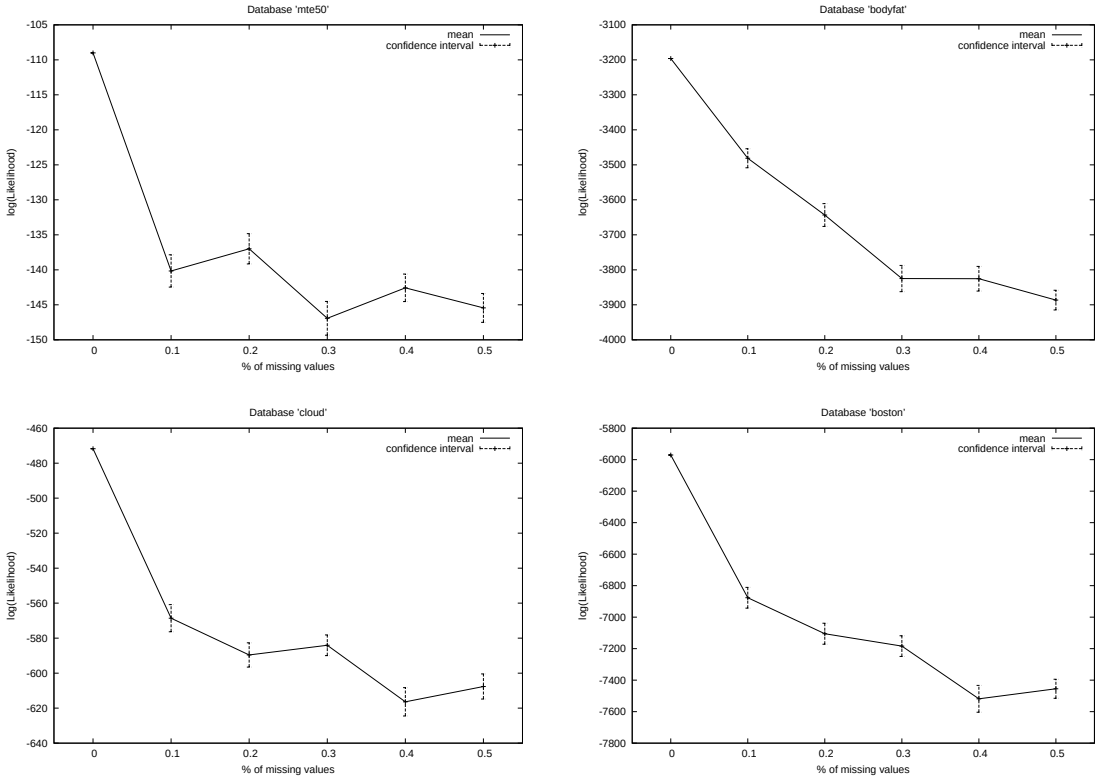

Figure 2: Loglikelihood of the learnt models as a function of the percentage of missing values.

tional distribution of the variable for which the missing value is going to be imputed, whilst M5' uses the marginal distribution. Our impression is that the reason why the results in these experiments are not even better for NB is the limited size of the used databases. Nevertheless, the independence assumptions in the NB model can be a limitation, and therefore more complex structures, like the TAN, might lead to more significant differences.

5 Concluding remarks

In this paper we have described a method for learning regression models from incomplete data based on the MTE distribution over a naïve Bayes network structure. The algorithm is supported by a result on how to minimise the prediction error and the experiments carried out, though somehow limited, show a reasonable performance of the new algorithm, compared to the robust M5' scheme, which is not surprising, as M5' is mainly designed for continuous explanatory variables. The behaviour of the algorithm is also good in terms of likelihood, even though that aspect is not really relevant to the aim of the method, which is to provide low prediction error.

The algorithm presented here can be improved in various ways, as for instance, by considering different manners of imputing the explanatory variables.

We think that the ideas contained in this work can be applied to other regression models like the TAN. However, the application to a broader problem like learning a Bayesian network of general purpose, is not straightforward, since in this case the goal would be to maximise a score based on the likelihood function, which requires maximum likelihood estimates of the parameters of the MTE model.

Acknowledgements

Work supported by the Spanish Ministry of Science and Innovation, projects TIN2007-67418-

		% of missing values					
Model	Database	0	0.1	0.2	0.3	0.4	0.5
NB	bodyfat	6.7095	6.3496	6.4602	6.6235	6.1287	6.9734
M5'	bodyfat	25.21	24.4519	29.0318	28.7724	28.6139	6.0929
NB	boston	6.2088	6.8668	6.4182	6.9748	7.0931	7.3654
M5'	boston	4.1475	5.1185	5.2011	5.6909	5.9646	6.6753
NB	cloud	0.5572	0.4897	0.6282	0.5350	0.7925	0.7137
M5'	cloud	0.3764	0.3237	0.6493	0.4421	0.4925	0.5919
NB	mte50	1.8695	2.0980	2.6392	2.7415	2.8957	3.0541
M5'	mte50	2.4718	2.7489	3.1566	2.6619	3.3681	3.4407

Table 2: Average srmse obtained in the experiment comparing NB vs. M5'.

C03-01 and TIN2007-67418-C03-02, and by Junta de Andalucía, project P05-TIC-00276.

References

C.L. Blake and C.J. Merz. 1998. UCI repository of machine learning databases. www.ics.uci.edu/∼mlearn/MLRepository.html. University of California, Irvine, Dept. of Information and Computer Sciences.

B. Cobb and P.P. Shenoy. 2006. Inference in hybrid Bayesian networks with mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 41:257–286.

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

J. Demsar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1 – 30.

Elvira Consortium. 2002. Elvira: An environment for creating and using probabilistic graphical models. In J.A. Gámez and A. Salmerón, editors, *Proceedings of the First European Workshop on Probabilistic Graphical Models*, pages 222–230.

A. Fernández, M. Morales, and A. Salmerón. 2007. Tree augmented naive Bayes for regression using mixtures of truncated exponentials: Applications to higher education management. *IDA'07. Lecture Notes in Computer Science*, 4723:59–69.

Nir Friedman. 1997. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the ICML-97*.

J.A. Gámez, Rafael Rumí, and A. Salmerón. 2006. Unsupervised naive Bayes for data clustering with mixtures of truncated exponentials. In *Proceedings of the 3rd European Workshop on Probabilistic Graphical Models (PGM'06)*, pages 123–132.

S. Moral, R. Rumí, and A. Salmerón. 2001. Mixtures of truncated exponentials in hybrid Bayesian networks. *ECSQARU'01. Lecture Notes in Artificial Intelligence*, 2143:135–143.

S. Moral, R. Rumí, and A. Salmerón. 2003. Approximating conditional MTE distributions by means of mixed trees. *ECSQARU'03. Lecture Notes in Artificial Intelligence*, 2711:173–183.

M. Morales, C. Rodríguez, and A. Salmerón. 2007. Selective naive Bayes for regression using mixtures of truncated exponentials. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 15:697–716.

J.R. Quinlan. 1992. Learning with continuous classes. In *Procs. of the 5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore. World Scientific.

R. Rumí and A. Salmerón. 2007. Approximate probability propagation with mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 45:191–210.

R. Rumí, A. Salmerón, and S. Moral. 2006. Estimating mixtures of truncated exponentials in hybrid Bayesian networks. *Test*, 15:397–421.

StatLib. 1999. www.statlib.org. Department of Statistics. Carnegie Mellon University.

M.A. Tanner and W.H Wong. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550.

Y. Wang and I.H. Witten. 1997. Induction of model trees for predicting continuous cases. In *Procs. of the Poster Papers of the European Conf. on Machine Learning*, pages 128–137.

I.H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann.