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

Antonio Fernández [1], Jens D. Nielsen [2] and Antonio Salmerón [1]

[1]Department of Statistics & Applied Mathematics, University of Almería (Spain)
[2]Computer Science Department, University of Castilla-La Mancha (Spain)

PGM'08, Hirtshals (Denmark), 19th september 2008

Outline

Outline

Outline

Outline

Learning NB regression models with missing data using MTEs

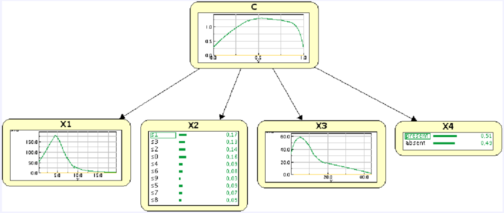
A. Fernández, J. Nielsen and A. Salmerón

Outline

Outline

Motivation

- What is the goal? Construct a NB regression model based on MTEs:

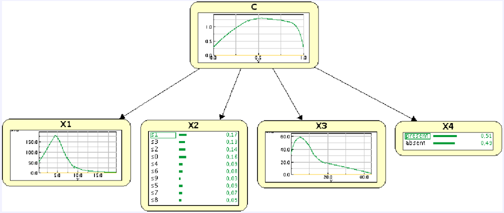


- What is the problem? Incomplete dataset.

X1	X2	X3	X4	C
2.34	?	6.54	6.91	0
?	3	2.77	3.58	?
6.24	?	4.28	?	1
4.55	5	3.64	6.54	?
2.94	?	?	6.54	1

Motivation

- What is the goal? Construct a NB regression model based on MTEs:



- What is the problem? Incomplete dataset.

X1	X2	X3	X4	C
2.34	?	6.54	6.91	0
?	3	2.77	3.58	?
6.24	?	4.28	?	1
4.55	5	3.64	6.54	?
2.94	?	?	6.54	1

Motivation

- MTEs provides a framework for handling hybrid Bayesian networks.
- Regression problems can be solved using Bayesian networks.
- Previous algorithms operates over complete databases.
- We propose an iterative algorithm for constructing NB regression models from incomplete databases.

Previous works

- Mixtures of Truncated Exponentials (MTEs) in Bayesian networks (Moral et. al, 2001)
- MTEs are compatible with standard inference algorithms and there is no restriction on the structure (Cobb and Shenoy, 2006; Rumí and Salmerón, 2007).
- MTEs applied to regression problems (full databases): Naïve Bayes (Morales et. al, 2007) y TAN (Fernández et. al, 2007).
- Unsupervised data clustering (only values of response variable are missing): (Gámez et al., 2006)

The MTE model (Moral et al. 2001)

Definition (MTE potential)

- $\mathbf{X}$: mixed n-dimensional random vector. $\mathbf{Y} = (Y_1, \ldots, Y_d)$, $\mathbf{Z} = (Z_1, \ldots, Z_c)$ its discrete and continuous parts. A function $f : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ is a Mixture of Truncated Exponentials potential (MTE potential) if for each fixed value $\mathbf{y} \in \Omega_{\mathbf{Y}}$ of the discrete variables $\mathbf{Y}$, the potential over the continuous variables $\mathbf{Z}$ is defined as:

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

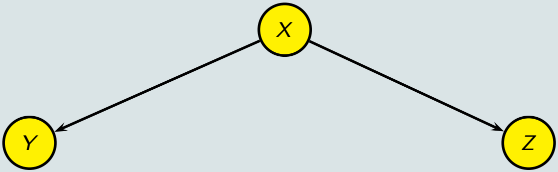
  for all $\mathbf{z} \in \Omega_{\mathbf{Z}}$, where a_i, $b_i^{(j)}$ are real numbers.

- Also, 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 above.

The MTE model

Example

Consider a regression model with continuous class variable X, and with two features Y and Z, where Y is continuous and Z is discrete.



The MTE model

Example

Densities for this regression model:

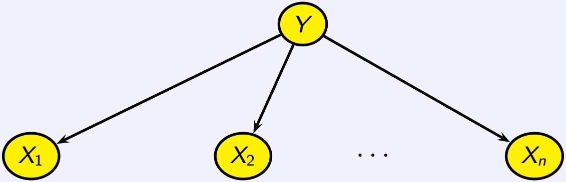
$$f(x) = \begin{cases} 1.16 - 1.12e^{-0.02x} & \text{if } 0.4 \leq x < 4 \ , \\ 0.9e^{-0.35x} & \text{if } 4 \leq x < 19 \ . \end{cases}$$

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

$$f(z|x) = \begin{cases} 0.3 & \text{if } z = 0,\ 0.4 \leq x < 5 \ , \\ 0.7 & \text{if } z = 1,\ 0.4 \leq x < 5 \ , \\ 0.6 & \text{if } z = 0,\ 5 \leq x < 19 \ , \\ 0.4 & \text{if } z = 1,\ 5 \leq x < 19 \ . \end{cases}$$

Bayesian networks for regression

- In general a network with a classifier structure can be used as regression model.
- Naïve Bayes model can be used for regression purposes (Frank et al. 2000, Morales et al. 2007).



- The posterior distribution of Y given $X_1, \ldots, X_n$ can be used to obtain a prediction for Y.
- The expectation or the median can be used.

Regression using MTEs

- Let $Y, X_1, \ldots, X_n$ be, where Y is continuous and the rest are either discrete or continuous.
- Goal: find a model g that explains the **response** variable Y in terms of the **explanatory** variables $X_1, \ldots, X_n$,
- If we have a configuration $x_1, \ldots, x_n$, a prediction of Y is:

$$\hat{y} = g(x_1, \ldots, x_n)$$

Regression using MTEs

Our regression model

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

- $f(y|x_1, \ldots, x_n)$: conditional density of Y given $x_1, \ldots, x_n$, which we assume to be of class MTE.
- The distribution of Y can be regarded as an **approximation** of the true distribution of the actual values of Y.
- This justifies:
  - Selection of $E[Y|x_1, \ldots, x_n]$ as the predicted value.
  - Minimises the mean squared error between the actual value of Y and its prediction $\hat{y}$:

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

Regression model from incomplete data

- Key point: Find a regression model imputing the missing values in the way to obtain the lowest *srmse*.
- Some previous ideas about imputing missing data:
  - EM algoritm (Dempster et al., 1977): Problem: likelihood function cannot be optimised in an exact way (Rumí et al., 2006) and also our goal is to minimise the *mse* rather than high likelihood.
  - Data Augmentation (Tanner and Wong, 1987): Problem: Initial random imputation, maximum likelihood estimates of the parameters.

General algorithm

General steps

1. Initialization: Fill the missing cells in the database and learn an initial model.
2. Iterative process: while *smrse* is improved, create a new database **imputing** the values with the current model and learn a new model for the next step.

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

3. Return **current model**

How do we impute the missing values?

Imputation

- Response variable (Y): ? $\longrightarrow E[Y|x_1, \ldots, x_n]$, because the *mse* is reduced (next proposition).
- Explanatory variables (X_i): ? $\longrightarrow$ simulating from its conditional distribution given the values of the other variables in the same record.

Problem

- We need an initial model (NB). How is the database filled at the beginning?
- For each variable we simulate from a marginal distribution learnt from the present values of the variable.

Imputation of the response variable?

- Why the $E[Y|x_1,\ldots,x_n]$ is the best imputed value?

Proposition

Let Y and $\hat{Y}$ be two continuous independent and identically distributed random variables. Then,

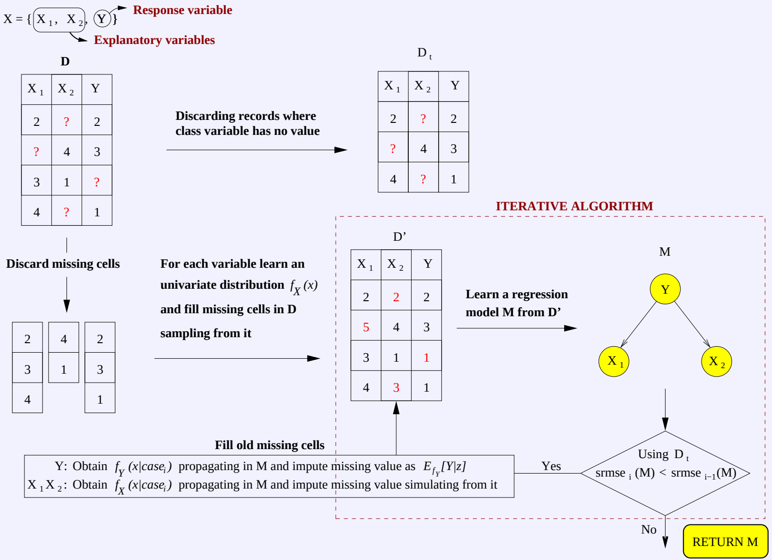
$$E[(Y-\hat{Y})^2] \geq E[(Y-E[Y])^2] \ .$$

Proof.

$$\begin{aligned} E[(Y-\hat{Y})^2] &= E[Y^2+\hat{Y}^2-2Y\hat{Y}] = E[Y^2]+E[\hat{Y}^2]-2E[Y\hat{Y}] \\ &= E[Y^2]+E[\hat{Y}^2]-2E[Y]E[\hat{Y}] = 2E[Y^2]-2E[Y]^2 \\ &= 2(E[Y^2]-E[Y]^2) = 2\,\mathrm{Var}(Y) \geq \mathrm{Var}(Y) = E[(Y-E[Y])^2] \ . \end{aligned}$$

$\square$

Algorithm: Learning model from incomplete data



Experiments: Description of the databases

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

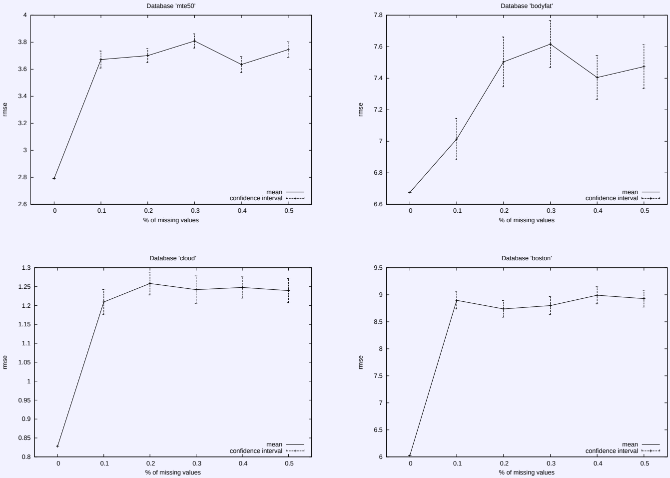
Experiments

- **Experiment 1**: Is the error of the model related to the % of missing values?
- **Experiment 2**: Compare the proposed model with the M5' algorithm (Wang and Witten, 1997)
  - M5' algorithm is a good reference point in graphical models for regression
  - M5' is implemented by Weka (Witten and Frank, 2005)
  - Proposed model has been included in the Elvira software (Elvira Consortium, 2002)

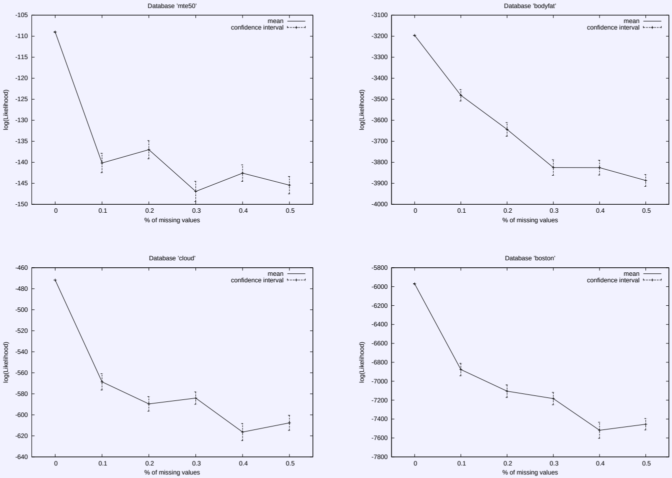
Experiment 1: Error vs. % of missing values

- Database is divided in two parts: training (70 %) and test (30 %).
- Missing values {10%, 20%, 30%, 40%, 50%} are inserted in the training set.
- The experiment has been repeated 100 times. We show the average srmse over the same test data by the 100 models learnt.
- The confidence interval 95 % is shown in the figure.

rmse vs. % missing values



Loglikelihood vs. % missing values



Experiment 2: Comparing NB vs. M5'.

- M5': imputes the missing values in the explanatory variables with the average or the mode.
- NB: doesn't require the imputation of the missing values. It can be marginalised out by propagation.
- 10- fold cross validation
- Friedman test reports no statistically significant differences between both methods (p-value $= 0.6831$)
- It is surprising the error obtained by M5' for `bodyfat`.

srmse

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

Results discussion

1^{st} experiment

- Proposed method behaves in a reasonable way
- Graphs show the error increasing along with the rate of missing values.
- Sometimes the error decreases, due to overfitting (around 40 % of missing)
- In terms of likelihood, we have similar graphs. It decreases when the % missing values is higher.

Results discussion

2^{nd} experiment

- NB should be superior to M5' in the case of missing values. Why?
- NB: Impute the missing values taking into account the conditional distribution for each variable.
- M5': Uses the marginal distribution.
- This is not so clear in the experiments due to: independence assumptions in NB, the size of the databases, MTE learning ...

Some conclusions

Summary

- Bayesian networks (NB) can be applied to solve regression problems.
- MTE distributions have been used for handling hybrid Bayesian networks.
- Use of the conditional density of the response variable to predict the response value.
- Capacity to manage incomplete datasets.
- Experiments behaves in a reasonable way.
- Difficult to apply this metodology for learning Bnets with no restriction in the structure.

Thanks for your attention