Mixture of Truncated Exponentials in Supervised Classification: case study for the Naive Bayes and Averaged One-Dependence Estimators Classifiers

Presentation at ISDA
on 23/11/2011

M. Julia Flores[1], José A. Gámez[1], Ana M. Martínez[1] and Antonio Salmerón[2]

[1]Intelligent Systems and Data Mining's group (SIMD)
UCLM - Albacete - Spain
[2]Department of Statistics and Applied Mathematics
UAL - Almería - Spain

UNIVERSIDAD DE ALMERÍA
UCLM UNIVERSIDAD DE CASTILLA~LA MANCHA

Outline

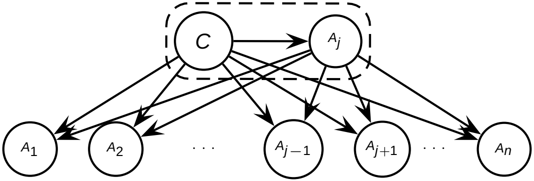
1. **Motivation**
2. **Previous proposals based on CGNs**
   - GAODE
   - HAODE
3. **Mixture of Truncated Exponentials (MTEs)**
   - Learning MTEs from data
4. **The MTE-AODE classifier**
5. **Experimental Methodology and Results**
   - Selection of the number of cutpoints in MTEs
6. **Conclusions**

Outline

1. **Motivation**
2. **Previous proposals based on CGNs**
   - GAODE
   - HAODE
3. **Mixture of Truncated Exponentials (MTEs)**
   - Learning MTEs from data
4. **The MTE-AODE classifier**
5. **Experimental Methodology and Results**
   - Selection of the number of cutpoints in MTEs
6. **Conclusions**

AODE classifier I

- **AODE** is significantly better in terms of error reduction compared to the rest of semi-naive techniques.



- **MAP hypothesis:**

$$argmax_{c\in\Omega_C}\left(\sum_{j=1,N(a_j)>m}^{n} p(c,a_j)\prod_{i=1,i\neq j}^{n} p(a_i|c,a_j)\right)$$

- Quadratic in learning and classification time.

× : **Only discrete variables**.

Motivation

- A BN assume **all** the variables are **discrete**.
- **Large amount of methods** developed to solve problems with **discrete variables**.
- It is common the **coexistence** of **discrete** and **continuous** variables in the same problem.
- Direct solution $\rightarrow$ **discretization**.
  - ✗ : Unavoidable lost of precision.
  - ✗ : Which discretization method should we choose?.
- Some alternative solutions:
  1. **Conditional Gaussian Networks (CGNs)**: GAODE and HAODE.
  2. **Kernel-based distributions**:high costs, problems in inference.
  3. **Mixtures of truncated exponentials**: exact frame for working with hybrid nets.

Outline

Previous proposals based on CGNs

- Proposals based on the assumption that for each configuration of the categorical variables, all the numerical attributes are **sampled from** a Gaussian density distribution.

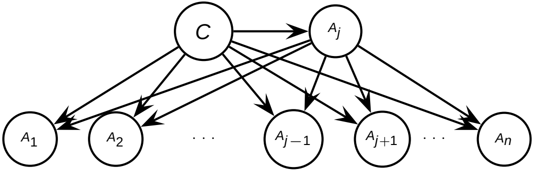
- Despite this strong assumption, Gaussian distributions usually provide reasonably **good approximations** to many real-world distributions.

- GAODE and HAODE inherit AODE's graphical structure, but differ in the way the superparent attribute is managed.

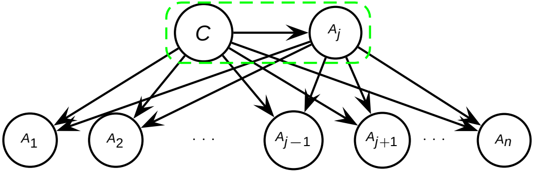
Outline

1. Motivation
2. **Previous proposals based on CGNs**
   - GADOE
   - HAODE
3. Mixture of Truncated Exponentials (MTEs)
   - Learning MTEs from data
4. The MTE-AODE classifier
5. Experimental Methodology and Results
   - Selection of the number of cutpoints in MTEs
6. Conclusions

GAODE classifier

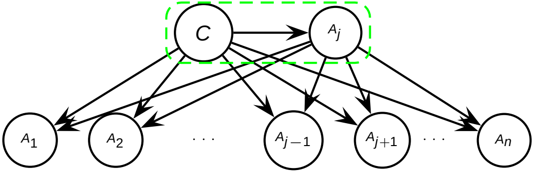


GAODE classifier



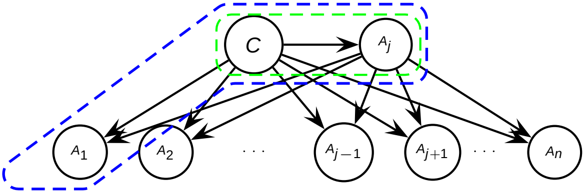
GAODE classifier

- Univariate Gaussian Distribution -



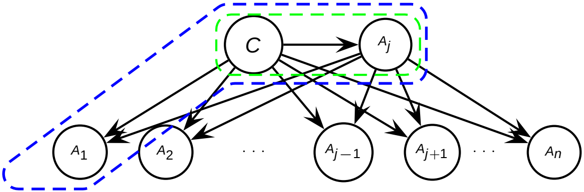
GAODE classifier

- Univariate Gaussian Distribution -



GAODE classifier

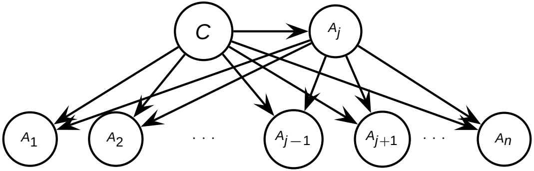
- Univariate Gaussian Distribution -



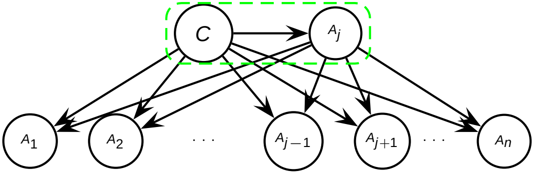
- Conditional Gaussian Distribution -

Outline

HAODE classifier

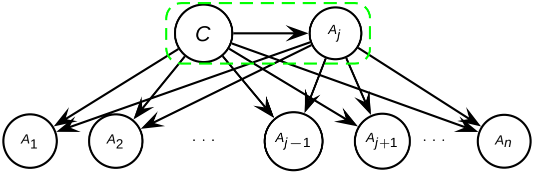


HAODE classifier



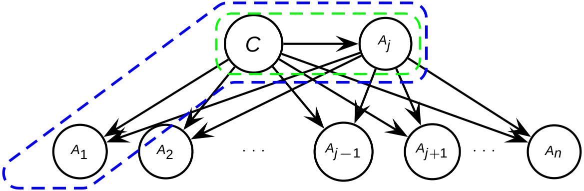
HAODE classifier

- Multinomial distribution -



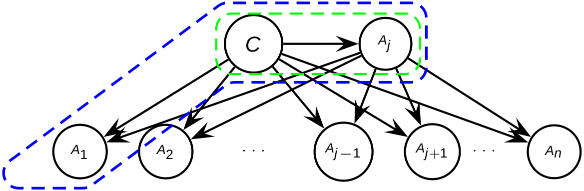
HADOE classifier

- Multinomial distribution -



HAODE classifier

- Multinomial distribution -

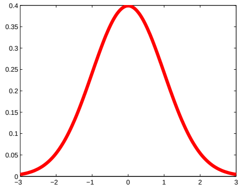


- Univariate Gaussian distribution -

Performance of GAODE and HAODE

- According to previous experiments carried out [2]:
  - In numerical domains:
    - There is no statistical difference between GAODE and AODE in numerical domains.
    - HAODE is better than both of them.
  - In hybrid domains:
    - No statistical difference was found between the performance of AODE and HAODE.
    - Reconsider Gaussian asumption.

CGs: illegal configuration



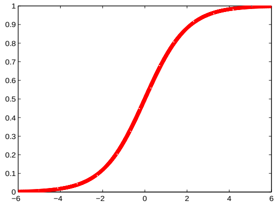
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$$

Z

$\downarrow$

Y

$$P(Y = 1)?$$

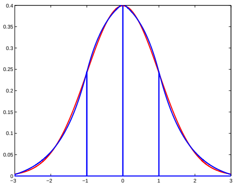


$$P(Y = 1|z) = \frac{1}{1+\exp(-z)}$$

Outline

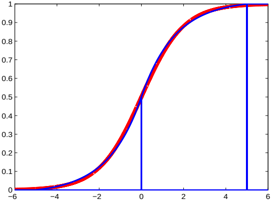
1. Motivation
2. Previous proposals based on CGNs
   - GAODE
   - HAODE
3. **Mixture of Truncated Exponentials (MTEs)**
   - Learning MTEs from data
4. The MTE-AODE classifier
5. Experimental Methodology and Results
   - Selection of the number of cutpoints in MTEs
6. Conclusions

MTEs



Z

Y



$$f(z) = \begin{cases} -0{,}0172 + 0{,}931e^{1{,}27z} & \text{if } -3 \leq z < -1 \\ 0{,}442 - 0{,}0385e^{-1{,}64z} & \text{if } -1 \leq z < 0 \\ 0{,}442 - 0{,}0385e^{1{,}64z} & \text{if } 0 \leq z < 1 \\ -0{,}0172 + 0{,}9314e^{-1{,}27z} & \text{if } 1 \leq z < 3 \end{cases}$$

Calculate $P(Y = 1)$ with MTEs: $P(Y = 1) \approx 0{,}4996851$

$$P(Y = 1|z) = \begin{cases} 0 & \text{si } z < -5 \\ -0{,}0217 + 0{,}522e^{0{,}635z} & \text{if } -5 \leq z < 0 \\ 1{,}0217 - 0{,}522e^{-0{,}635z} & \text{if } 0 \leq z \leq 5 \\ 1 & \text{si } z > 5 \end{cases}$$

The MTE model (Moral et al., 01)

Definition (MTE potential)

- $\mathbf{X}$: mixed n-dimensional random vector. $\mathbf{Y} = (Y_1, \dots, Y_d)$, $\mathbf{Z} = (Z_1, \dots, Z_c)$ 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:

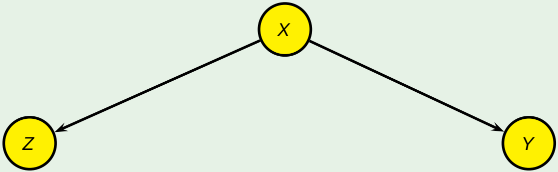
$$f(\mathbf{z}) = a_0 + \sum_{i=1}^{m} a_i \exp\left\{ \sum_{j=1}^{c} b_i^{(j)} z_j \right\}$$

  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 $D_1, \dots, D_k$ of $\Omega_{\mathbf{Z}}$ into hypercubes and in each D_i, f is defined as above.

The MTE model

Example

Consider a Bayesian network with continuous variables X, and Z and discrete variable Y.



The MTE model

Example

Densities for this BN:

$$f(x) = \begin{cases} 1{,}16 - 1{,}12e^{-0{,}02x} & \text{if } 0{,}4 \leq x < 4 \ , \\ 0{,}9e^{-0{,}35x} & \text{if } 4 \leq x < 19 \ . \end{cases}$$

$$f(z|x) = \begin{cases} 1{,}26 - 1{,}15e^{0{,}006z} & \text{if } 0{,}4 \leq x < 5,\ 0 \leq z < 13 \ , \\ 1{,}18 - 1{,}16e^{0{,}0002z} & \text{if } 0{,}4 \leq x < 5,\ 13 \leq z < 43 \ , \\ 0{,}07 - 0{,}03e^{-0{,}4z} + 0{,}0001e^{0{,}0004z} & \text{if } 5 \leq x < 19,\ 0 \leq z < 5 \ , \\ -0{,}99 + 1{,}03e^{0{,}001z} & \text{if } 5 \leq x < 19,\ 5 \leq z < 43 \ . \end{cases}$$

$$f(y|x) = \begin{cases} 0{,}3 & \text{if } y = 0,\ 0{,}4 \leq x < 5 \ , \\ 0{,}7 & \text{if } y = 1,\ 0{,}4 \leq x < 5 \ , \\ 0{,}6 & \text{if } y = 0,\ 5 \leq x < 19 \ , \\ 0{,}4 & \text{if } y = 1,\ 5 \leq x < 19 \ . \end{cases}$$

Outline

1. Motivation
2. Previous proposals based on CGNs
   - GAODE
   - HAODE
3. **Mixture of Truncated Exponentials (MTEs)**
   - Learning MTEs from data
4. The MTE-AODE classifier
5. Experimental Methodology and Results
   - Selection of the number of cutpoints in MTEs
6. Conclusions

Learning MTEs from data:

Univariate case

Learning a univariate MTE density involves three basic steps:

Learning MTEs from data:

Univariate case

Learning a univariate MTE density involves three basic steps:

- Determination of the splits into which Ω_X will be partitioned.

Learning MTEs from data:

Univariate case

Learning a univariate MTE density involves three basic steps:

- Determination of the splits into which Ω_X will be partitioned.
- Determination of the number of exponential terms in the mixture for each split.

Learning MTEs from data:

Univariate case

Learning a univariate MTE density involves three basic steps:

- Determination of the splits into which Ω_X will be partitioned.
- Determination of the number of exponential terms in the mixture for each split.
- Estimation of the parameters.

Learning MTEs from data:

Univariate case

Learning a univariate MTE density involves three basic steps:

- Determination of the splits into which Ω_X will be partitioned.
- Determination of the number of exponential terms in the mixture for each split.
- Estimation of the parameters.

Conditional case

Learning MTEs from data:

Univariate case

Learning a univariate MTE density involves three basic steps:

- Determination of the splits into which Ω_X will be partitioned.
- Determination of the number of exponential terms in the mixture for each split.
- Estimation of the parameters.

Conditional case

- Mixed tree: efficient data structure for dealing with conditional MTE densities.

Learning MTEs from data:

Univariate case

Learning a univariate MTE density involves three basic steps:

- Determination of the splits into which Ω_X will be partitioned.
- Determination of the number of exponential terms in the mixture for each split.
- Estimation of the parameters.

Conditional case

- Mixed tree: efficient data structure for dealing with conditional MTE densities.

Alternatives explored in the literature

- Least squared estimation (LS).
- Maximum likelihood estimation (ML).

Determination of the **structure**: it is usually fixed in advanced.

Outline

1. Motivation
2. Previous proposals based on CGNs
   - GAODE
   - HAODE
3. Mixture of Truncated Exponentials (MTEs)
   - Learning MTEs from data
4. **The MTE-AODE classifier**
5. Experimental Methodology and Results
   - Selection of the number of cutpoints in MTEs
6. Conclusions

MTE-AODE classifier

Definition

An MTE-AODE classifier classifies an individual as in AODE, where all the probability functions involved are of class MTE.

- **Advantages**:
  - (over GAODE) the underlying distribution is not assumed to be of Gaussian type, as MTEs are able to accurately represent the most common distributions.
  - (over HAODE) there is no need to discretize the super-parent nodes, as the MTE mode allows discrete variables with continuous parents.
- **Disadvantages**: the learning phase is slower, as the parameters are adjusted iteratively following the algorithm described in [5].
- We can trivially extend this definition to the MTE-NB classifier as a particular case of MTE-AODE.

Outline

Previous considerations

- Study the application of MTEs over a **large group of datasets** without individual parametrization.
  - 16 hybrid datasets.
- Can a general recommendation on how to best handle continuous attributes can be given?
  - So far, only studies on specific datasets.
- The number of intervals into which the domain for a continuous variable is split in an MTE potential.
  - In MTEs, EF discretization is applied only to parent variables in a conditional distribution represented by a mixed tree.
  - The domain of the variable that actually appears in the functional definition of the MTE potential in each leaf of the mixed tree is split taking into account the inflection and extreme points of the sample density.
  - Based on empirical results: trade-off between low complexity and high fitting power.
  - Equal frequency division with 5 and 10 intervals ((**EF5** and **EF10**) as in [6] and Fayyad & Irani (**F&I**).
- 5x2 cross validation.
- Filters to replace missing values and remove useless attributes.

Accuracy comparisons

Table: Accuracy comparison between NB and AODE with MTE and the other algorithms. The bullet indicates significant performance applying the Wilcoxon test.

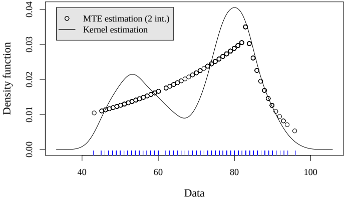
	Naive Bayes		AODE	
	MTE vs G	MTE vs D	MTE vs H	MTE vs D
EF5	11-5	7-9	9-7	7-9
EF10	10-6	5-11	8-8	•5-11
F&I	10-6	9-7	8-8	8-8

Outline

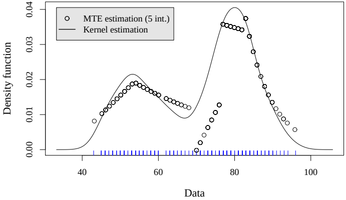
Table: Accuracy results obtained for MTE-NB and MTE-AODE when using 5EF, 10EF and F&I to create the intervals.

	Naive Bayes			AODE		
Id	5EF	10EF	F&I	5EF	10EF	F&I
1	82,5612	•82,6726	81,2695	•82,2717	82,0935	81,4477
2	90,2227	•90,3341	89,6659	•93,2294	92,9621	92,9621
3	59,2109	•59,8934	59,6031	62,0350	63,0164	•63,5028
4	68,6957	69,1304	•69,7826	69,5652	69,4565	•70,7065
5	80,4891	•80,7609	80,1087	•81,7935	81,6848	81,6304
6	•84,4348	84,2029	83,7101	•85,3333	84,9275	84,8986
7	•74,5600	74,4200	74,4400	74,7400	•75,1000	75,0800
8	•82,6965	82,5645	81,4434	82,5000	82,5658	•83,0281
9	•84,1497	84,0136	83,8776	•84,2177	83,6054	•84,2177
10	•85,0316	•85,0316	83,4898	84,2524	84,2541	•84,9018
11	94,1569	94,1729	•94,2312	93,8388	93,7964	•93,8600
12	•90,1847	•90,1847	88,4236	•91,2562	•91,2562	90,1847
13	83,2432	83,2432	•83,3784	•83,7838	•83,7838	83,6486
14	•94,3584	94,3054	94,2948	94,1676	•94,5387	93,6638
15	55,1313	•55,5758	51,7980	70,7879	•71,3333	69,5152
16	93,6627	93,6627	93,6627	93,8588	93,8588	93,8588
Av	81,4244	81,5105	80,8237	82,9770	83,0146	82,9442

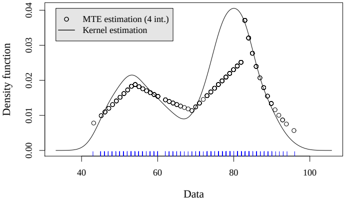
Estimation of MTEs when selecting different number of cutpoints



(a) 2 intervals



(b) 5 intervals



(c) 4 intervals

Outline

1. Motivation
2. Previous proposals based on CGNs
   - GAODE
   - HAODE
3. Mixture of Truncated Exponentials (MTEs)
   - Learning MTEs from data
4. The MTE-AODE classifier
5. Experimental Methodology and Results
   - Selection of the number of cutpoints in MTEs
6. **Conclusions**

Conclusions

- Generalization of AODE to all kind of datasets through MTEs.
    - Extensible to NB and other semi-naïve classifiers.

1. The selection of the number of intervals is not trivial:
    - It is advisable to find the optimum for every data set, but also,
    - To find the optimum for every domain to partition in each leaf of the mixed tree independently, requiring a more sophisticated supervised “discretization” technique oriented to the estimation of MTEs.
2. Time complexity:
    - The estimation of MTEs requires more than EF discretization.
    - This time can be controlled: limiting the number of exponential terms and playing with the maximum number of intervals into which divide the domain.

- Good alternative to CGs and discretization for some datsets.

Thank you

References I

- Zheng, F., Webb, G.I.: A Comparative Study of Semi-naive Bayes Methods in Classification Learning. In: 4th Australasian Data Mining Conference (AusDM05), Simoff, S.J, Williams, G.J, Galloway, J., Kolyshkina, I. editors, pp. 141–156. University of Technology, Sydney (2005)
- Flores, M. J., Gámez, J. A., Martínez, A. M., Puerta J. M.: GAODE and HAODE: Two Proposals based on AODE to Deal with Continuous Variables. In: 26th International Conference on Machine Learning: 40. Montreal, Canada (2009)
- Webb, G. I., Boughton, J. R., Wang, Z.: Not So Naive Bayes: Aggregating One-Dependence Estimators. J. Mach. Learn. 58 (1), 5-24 (2005)
- Yang Y., Webb G. Non-Disjoint Discretization for Naive-Bayes Classifiers. Proceedings of the 19th International Conference on Machine Learning (ICML-2002), 666-673. (2002).

References II

- R. Rumí, A. Salmerón, and S. Moral.: Estimating mixtures of truncated exponentials in hybrid Bayesian network. Test. 15, 397-421 (2006).
- Flores, M. J., Gámez, J. A., Martínez, A. M., Puerta J. M., Handling numeric attributes when comparing Bayesian network classifiers: does the discretization method matter?, APPLIED INTELLIGENCE, 2011