

Full Bayesian Model Averaging of Naïve Bayes for Clustering

Guzmán Santafé
Jose Antonio Lozano
Pedro Larrañaga

FBMA of Naïve Bayes for Clustering

- Full Bayesian Model Averaging (FBMA):

$$p(c, \mathbf{x}|D) = \sum_S p(S|D) \int p(c, \mathbf{x}|S, \boldsymbol{\theta})p(\boldsymbol{\theta}|S, D)d\boldsymbol{\theta}$$

- Complete Data:

Dash and Cooper (2002) obtain a closed form to calculate a FBMA with supervised naïve Bayes classifiers

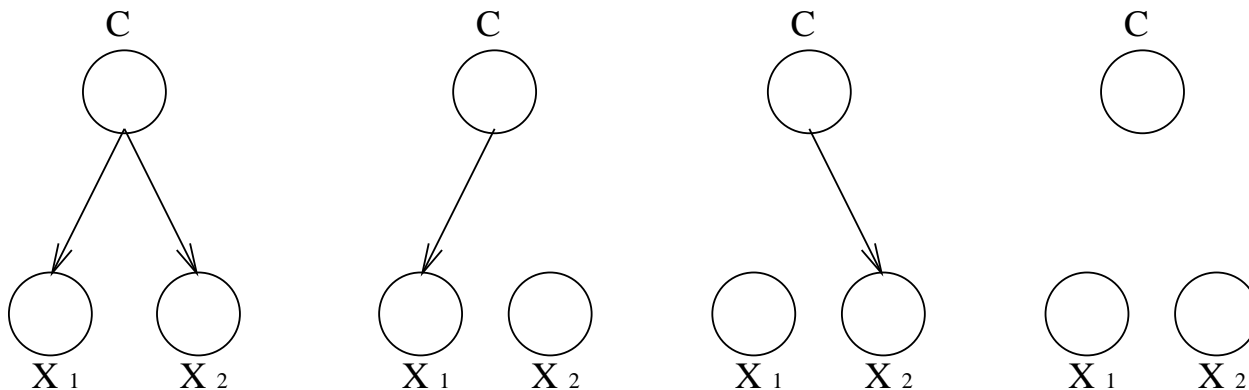
- Incomplete Data:

No exact computation of FBMA is possible

We propose an approximation of FBMA with naïve Bayes for clustering

FBMA of Naïve Bayes for Clustering

- **Naïve Bayes:** every predictive variable is conditional independent given C
- **Selective naïve Bayes:** every predictive variable can be independent or dependent on C



All selective naïve Bayes structures with two predictive variables

FBMA of Naïve Bayes for Clustering

- FBMA \approx averaging over MAP parameter configurations for all selective naïve Bayes structures
- Approximation to FBMA computed in the same time complexity required to learn MAP parameters:
 - Expectation and Model Averaging (EMA) algorithm allows to deal with missing data
 - Dash and Cooper (2002) formula is extended to clustering
- A unique naïve Bayes for clustering is obtained

Assumptions

1. Multinomial variables: $X_i = \{x_i^1, \dots, x_i^{r_i}\}$ with $i = 1, \dots, n$ and $C = \{c^1, \dots, c^{r_C}\}$
2. Complete dataset except for the latent cluster variable (C)
3. Dirichlet priors over parameters

$$\theta_{ijk} \sim \mathcal{D}(\alpha_{ijk})$$

4. Parameter independence when the dataset is complete

$$p(\boldsymbol{\theta}|S) = \boldsymbol{\theta}_C \prod_{i=1}^n \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij}|S)$$

5. Structure modularity

$$p(S) \propto p_S(C) \prod_{i=1}^n p_S(X_i, Pa_i)$$

EMA algorithm

- Adaptation of the EM algorithm (Dempster et al. 1979)
- Random initialization of the parameters
- Solution calculated iteratively in two steps:
 - Expectation
 - Model Averaging
- EMA is a greedy algorithm

EMA algorithm. E step

- The same E step as the one from EM algorithm
 - The dataset, D , is ‘completed’ obtaining D^{C_E}

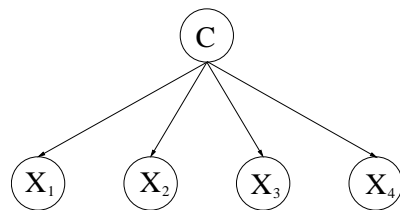
$$E(N_{ijk} | \boldsymbol{\theta}_S, S) = \sum_{l=1}^N p(c^j | \mathbf{x}^{(l)}, x_i^k, \boldsymbol{\theta}_S, S)$$

- $E(N_{ijk} | \boldsymbol{\theta}_S, S)$ is considered as actual N_{ijk}

FBMA of Naïve Bayes for Clustering

EMA algorithm. Example of E step

- Model



$$\begin{aligned} \theta_{C-1} &= 0,4 \\ \theta_{100} &= 0,27 & \theta_{110} &= 0,80 \\ \theta_{200} &= 0,70 & \theta_{210} &= 0,10 \\ \theta_{300} &= 0,95 & \theta_{310} &= 0,20 \\ \theta_{400} &= 0,30 & \theta_{410} &= 0,70 \end{aligned}$$

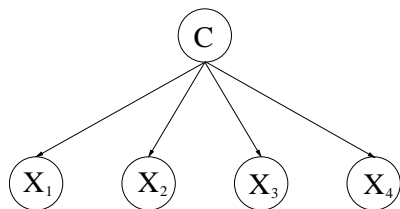
- Data Set

	X_1	X_2	X_3	X_4	C
1	1	1	0	1	?
2	0	1	0	0	?
3	1	0	1	0	?
4	1	1	1	0	?
5	0	0	1	0	?

FBMA of Naïve Bayes for Clustering

EMA algorithm. Example of E step

- Model



$$\begin{aligned} \theta_{C-1} &= 0,4 \\ \theta_{100} &= 0,27 & \theta_{110} &= 0,80 \\ \theta_{200} &= 0,70 & \theta_{210} &= 0,10 \\ \theta_{300} &= 0,95 & \theta_{310} &= 0,20 \\ \theta_{400} &= 0,30 & \theta_{410} &= 0,70 \end{aligned}$$

- Data Set

	X_1	X_2	X_3	X_4	C	
					C_0	C_1
1	1	1	0	1	0,85	0,15
2	0	1	0	0	0,13	0,87
3	1	0	1	0	0,33	0,67
4	1	1	1	0	0,02	0,95
5	0	0	1	0	0,04	0,96

$$\begin{aligned} N_{C-0} &= 1,37 & N_{C-1} &= 3,63 \\ N_{100} &= 0,17 & N_{101} &= 1,2 \\ N_{110} &= 1,83 & N_{111} &= 1,8 \\ & & \dots & \\ N_{410} &= 3,48 & N_{411} &= 0,15 \end{aligned}$$

FBMA of Naïve Bayes for Clustering

EMA algorithm. MA step

- MAP approximation for averaging over parameters (Heckerman, 1995)

$$\int p(c^j, \mathbf{x}|S, \boldsymbol{\theta})p(\boldsymbol{\theta}|S, D^{C_E})d\boldsymbol{\theta} \approx \tilde{\theta}_{C^j}^S \prod_{i=1}^n \tilde{\theta}_{ijk}^S$$

- FBMA approximation at each iteration:

$$\begin{aligned} p(c^j, \mathbf{x}|D^{C_E}) &= \sum_S \int p(c^j, \mathbf{x}|S, \boldsymbol{\theta})p(\boldsymbol{\theta}|S, D^{C_E})d\boldsymbol{\theta}p(S|D^{C_E}) \\ &\approx \kappa \sum_S \tilde{\theta}_{C^j}^S \prod_{i=1}^n \tilde{\theta}_{ijk}^S p(D^{C_E}|S)p(S) \end{aligned}$$

- Given assumptions (3) and (4) we can compute $p(D^{C_E}|S)$ (Cooper and Heskovits, 1992):

$$p(D^{C_E}|S) \approx \prod_{i=0}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + E(N_{ij}|\boldsymbol{\theta}, S))} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + E(N_{ijk}|\boldsymbol{\theta}, S))}{\Gamma(\alpha_{ijk})}$$

being, in this case, X_0 the cluster variable, C

FBMA of Naïve Bayes for Clustering

- Given assumption (5):

$$p(c^j, \mathbf{x} | D^{CE}) \approx \kappa \sum_S \rho_{C-j}^S \prod_{i=1}^n \rho_{ijk}^S$$

$$\rho_{ijk}^S = \tilde{\theta}_{ijk}^S p_S(X_i, Pa_i) \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + E(N_{ij} | \boldsymbol{\theta}^{(t)}, S))} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + E(N_{ijk} | \boldsymbol{\theta}^{(t)}, S))}{\Gamma(\alpha_{ijk})}$$

$$\rho_{C-j}^S = \tilde{\theta}_{C-j}^S p_S(C) \frac{\Gamma(\alpha_C)}{\Gamma(\alpha_C + E(N_C | \boldsymbol{\theta}^{(t)}, S))} \prod_{j=1}^{r_C} \frac{\Gamma(\alpha_{C-j} + E(N_{C-j} | \boldsymbol{\theta}^{(t)}, S))}{\Gamma(\alpha_{C-j})}$$

- If we expand the summation of structures:

$$p(c^j, \mathbf{x} | D^{CE}) \approx \kappa \left(\begin{array}{l} \rho_{C-j} \rho_{1-k} \rho_{2-k} \cdots \rho_{n-k} \\ + \rho_{C-j} \rho_{1jk} \rho_{2-k} \cdots \rho_{n-k} \\ \vdots \\ + \rho_{C-j} \rho_{1jk} \rho_{2jk} \cdots \rho_{njk} \end{array} \right) \left. \vphantom{\begin{array}{l} \rho_{C-j} \rho_{1-k} \rho_{2-k} \cdots \rho_{n-k} \\ + \rho_{C-j} \rho_{1jk} \rho_{2-k} \cdots \rho_{n-k} \\ \vdots \\ + \rho_{C-j} \rho_{1jk} \rho_{2jk} \cdots \rho_{njk} \end{array}} \right\} 2^n \text{ terms}$$

FBMA of Naïve Bayes for Clustering

- Σ_m^{jk} denotes the sum of the product up to $m - th$ variable:

$$\begin{aligned} \Sigma_m^{jk} &\equiv \rho_{C-j} \rho_{1-k} \rho_{2-k} \dots \rho_{m-k} \\ &+ \rho_{C-j} \rho_{1jk} \rho_{2-k} \dots \rho_{m-k} \\ &\vdots \\ &+ \rho_{C-j} \rho_{1jk} \rho_{2jk} \dots \rho_{mjk} \end{aligned}$$

- Thus, Σ_i^{jk} can be written as a recurrence relationship:

$$\Sigma_i^{jk} = \Sigma_{i-1}^{jk} (\rho_{i-k} + \rho_{ijk}), \quad \Sigma_0^{jk} = \rho_{C-j}$$

- Then, $p(c^j, \mathbf{x} | D^{CE})$ can be approximated by a closed form

$$p(c^j, \mathbf{x} | D^{CE}) \approx \kappa \rho_{cj} \prod_{i=1}^n (\rho_{i-k} + \rho_{ijk})$$

- Parameters for the model in current iteration of the EMA:

$$\theta_{ijk}^* \propto (\rho_{i-k} + \rho_{ijk})$$

Empirical Testing

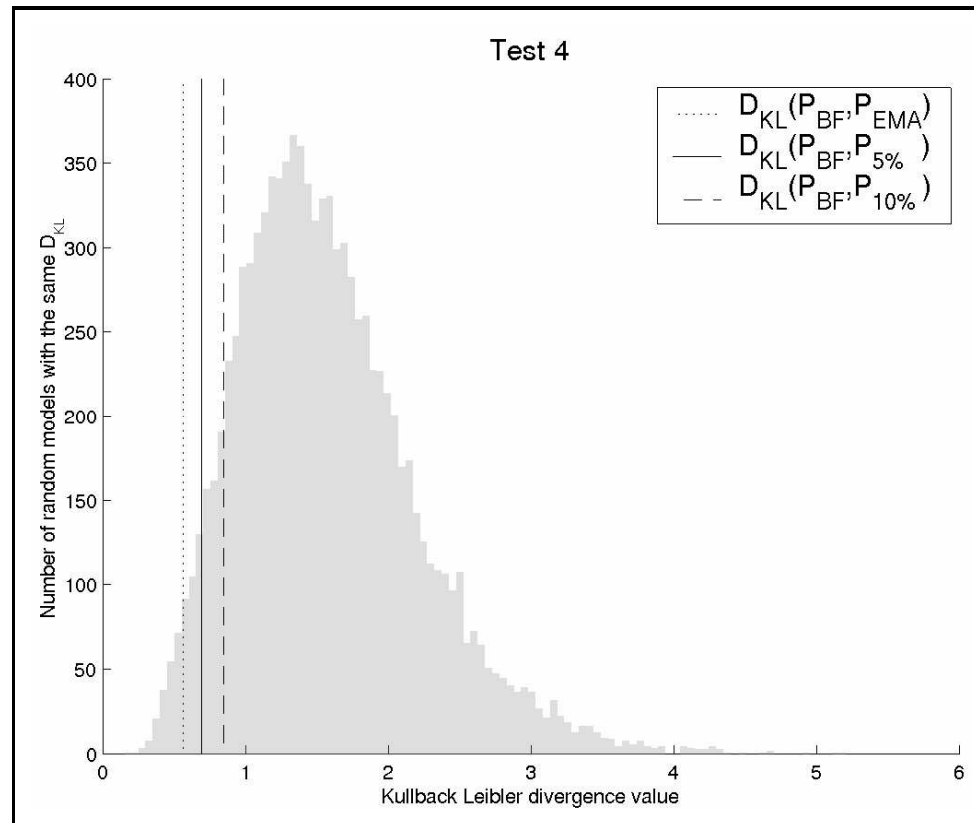
- Empirical evidence of the approximation is close to FBMA
- Model learned with EMA algorithm, P_{EMA}
- Brute force model, P_{BF} :

$$p(c, \mathbf{x}|D) = \sum_S \int p(c, \mathbf{x}|S, \boldsymbol{\theta})p(\boldsymbol{\theta}|S, D)d\boldsymbol{\theta} p(D|S)P(S)$$

- Averaging over parameters \approx MAP (EM)
- $P(D|S) \approx$ Candidate method (Chickering et al., 1997)
- Averaging over structures: Brute force
- Distance between models: $D_{KL}(P_{BF}, P_{EMA})$
- Distances between P_{BF} and 10000 random naïve Bayes models

FBMA of Naïve Bayes for Clustering

Empirical Testing



Test for a model with 6 predictive variables

Empirical Testing

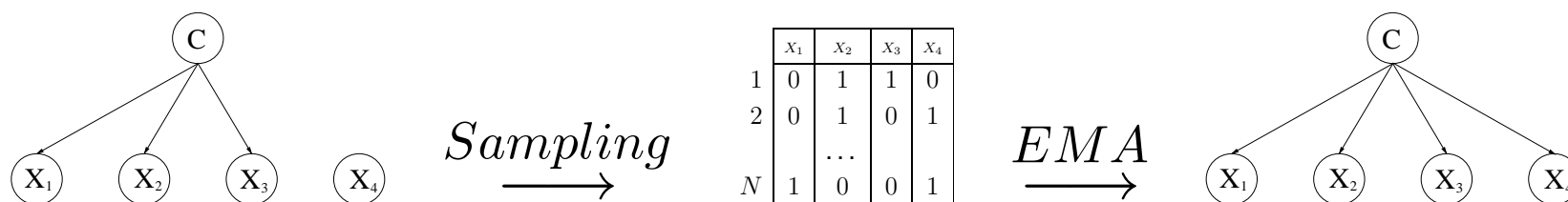
Test	$D_{KL}(P_{BF}, P_{EMA})$	$D_{KL}(P_{BF}, P_{5\%})$
1	1,15495	0,97264
2	0,55039	0,98286
3	0,41811	0,67852
4	0,56426	0,69513
5	0,33207	0,90145
6	0,42465	0,88862
7	0,40876	1,07019
8	0,45795	1,06288
9	0,24308	0,76765
10	0,13184	1,74327

10 independent tests for models with 6 predictive variables

FBMA of Naïve Bayes for Clustering

Model Detection

- Test if the EMA model detects the real model structure
- Set a selective naïve Bayes model
- Sample a data set
- Learn an EMA model from the data set

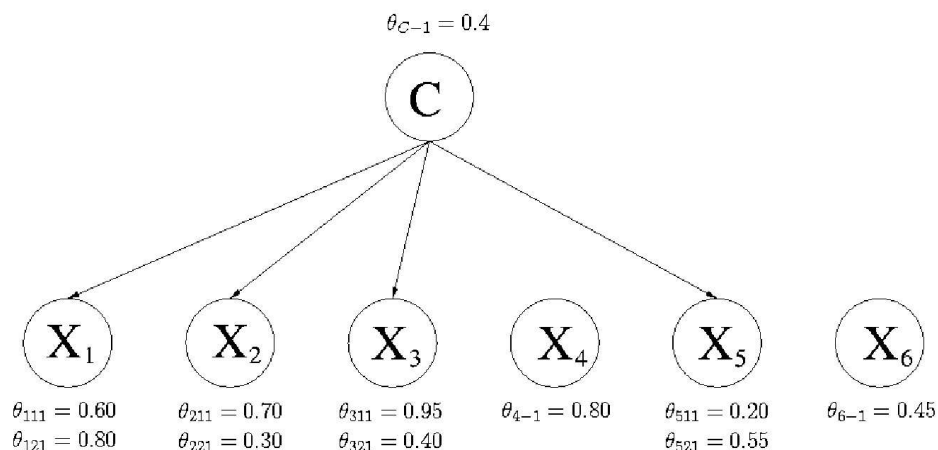


FBMA of Naïve Bayes for Clustering

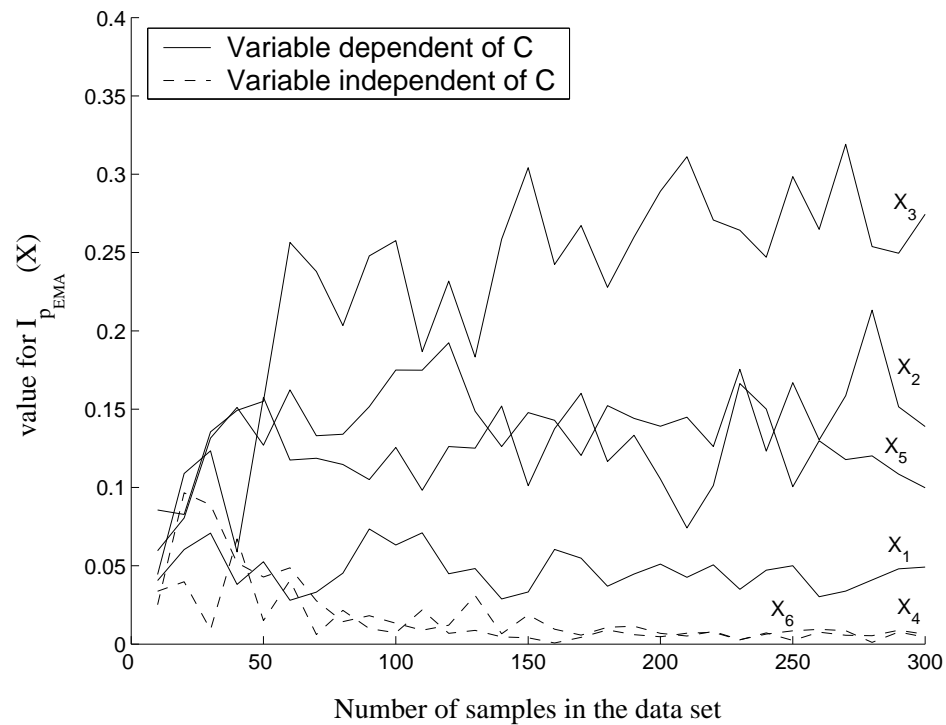
- Measure of independence for X_i in the EMA model

$$I_P(X_i) = \frac{\sum_{j=1}^{r_C} D_{KL}(p(X_i), p(X_i|c^j))}{r_C}$$

- The bigger is $N \rightarrow \begin{cases} \text{the better is the MAP approx.} \\ \text{the better is the EMA approx.} \end{cases}$
- Example of model detection



FBMA of Naïve Bayes for Clustering



Test for model detection with 6 predictive variables (X_4 and X_6 independent of C)

Conclusions

- Empirical test $\Rightarrow P_{EMA}$ good approximation to FBMA
- Approximate a FBMA with EMA is not much expensive than a classical MAP approach with EM
- EMA is able to detect independencies between variables:
EMA can be used for FSS
- EMA can be extended in order to deal with incomplete data.
- EMA can be extended to more complicated model (TAN)

FBMA of Naïve Bayes for Clustering

