Learning Parameters in Discrete Naive Bayes models by Computing Fibers of the Parametrization Map

Vincent Auvray and Louis Wehenkel

EE & CS Dept. and GIGA-R, University of Liège, Belgium

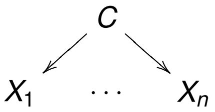
NIPS - AML 2008 - Whistler

Naive Bayesian networks

A discrete naive Bayesian network (or latent class model) with m classes is a distribution p over discrete variables $X_1, \ldots, X_n$ such that

$$p(X_1 = x_1, \ldots, X_n = x_n) = \sum_{t=1}^{m} p(C = t) \prod_{i=1}^{n} p(X_i = x_i | C = t).$$

Graphically, the independencies between $C, X_1, \ldots, X_n$ are encoded by



Problem statement

Given a naive Bayesian network p, compute the parameters $p(C = t), p(X_i = x_i | C = t)$ for $i = 1, \ldots, n$, $t = 1, \ldots, m$, and $x_i \in \mathcal{X}_i$ mapped to p.

Why?

- better understanding of the model
- estimation of parameters
- model selection
- study of parameter identifiability

Outline

Some notation

Given parameters of a naive Bayesian distribution, we define new parameters

$$\begin{aligned} w_t &= p(C = t), \\ A^t_{x_i} &= p(X_i = x_i | C = t) - p(X_i = x_i). \end{aligned}$$

Given a distribution p, let

$$\begin{aligned} q(x_{i_1}, \ldots, x_{i_k}) &= p(x_{i_1}, \ldots, x_{i_k}) \\ &- \sum_{\{X_{j_1},\ldots,X_{j_l}\} \subsetneq \{X_{i_1},\ldots,X_{i_k}\}} q(x_{j_1}, \ldots, x_{j_l}) \\ &\qquad \prod_{X_k \in \{X_{i_1},\ldots,X_{i_k}\} \setminus \{X_{j_1},\ldots,X_{j_l}\}} p(x_k). \end{aligned}$$

Some notation

For example, we have

$$\begin{aligned} q(x_i) =& 0, \\ q(x_i, x_j) =& p(x_i, x_j) - p(x_i)p(x_j), \\ q(x_i, x_j, x_k) =& p(x_i, x_j, x_k) - p(x_i)p(x_j, x_k) - p(x_j)p(x_i, x_k) \\ & - p(x_k)p(x_i, x_j) + 2p(x_i)p(x_j)p(x_k). \end{aligned}$$

With this notation, one can see that

$$q(x_{i_1}, \ldots, x_{i_k}) = \sum_{t=1}^{m} w_t \prod_{j=1}^{k} A^t_{x_{i_j}},$$

$$w^T A_{x_i} = 0,$$

where $w = \begin{pmatrix} w_1 & \ldots & w_m \end{pmatrix}^T$ and $A_{x_i} = \begin{pmatrix} A^1_{x_i} & \ldots & A^m_{x_i} \end{pmatrix}^T$.

w is normal to the hyperplane spanned by the A_{x_i}

Consider the parameters of a naive Bayesian distribution. Given vectors $A_{u_1}, \ldots, A_{u_{m-1}}$, we have

$$(-1)^t \det \begin{pmatrix} A_{u_1} & \ldots & A_{u_{m-1}} \end{pmatrix}^{\hat{t}} = w_t \det \begin{pmatrix} 1 & A_{u_1} & \ldots & A_{u_{m-1}} \end{pmatrix},$$

where the superscript $\hat{t}$ denotes the removal of the tth row.

In other words, if

$$\det \begin{pmatrix} 1 & A_{u_1} & \ldots & A_{u_{m-1}} \end{pmatrix} \neq 0,$$

then w is the normal to the hyperplane spanned by $A_{u_1}, \ldots, A_{u_{m-1}}$ and whose components sum to 1.

The components of A_{x_i} are the roots of a degree m polynomial

For $m = 2$, we have

$$s^2 q(u_1, v_1) + s q(x_i, u_1, v_1) - q(x_i, u_1) q(x_i, v_1) = q(u_1, v_1)(s + A_{x_i}^1)(s + A_{x_i}^2).$$

For $m = 3$, we have

$$s^3 \det \begin{pmatrix} q(u_1,v_1) & q(u_1,v_2) \\ q(u_2,v_1) & q(u_2,v_2) \end{pmatrix} + s^2 \left[\det \begin{pmatrix} q(u_1,v_1) & q(u_1,v_2) \\ q(x_i,u_2,v_1) & q(x_i,u_2,v_2) \end{pmatrix} + \det \begin{pmatrix} q(x_i,u_1,v_1) & q(x_i,u_1,v_2) \\ q(u_2,v_1) & q(u_2,v_2) \end{pmatrix} \right] + s \left[-\det \begin{pmatrix} q(u_1,v_1) & q(u_1,v_2) \\ q(x_i,u_2)q(x_i,v_1) & q(x_i,u_2)q(x_i,v_2) \end{pmatrix} - \det \begin{pmatrix} q(x_i,u_1)q(x_i,v_1) & q(x_i,u_1)q(x_i,v_2) \\ q(u_2,v_1) & q(u_2,v_2) \end{pmatrix} + \det \begin{pmatrix} q(x_i,u_1,v_1) & q(x_i,u_1,v_2) \\ q(x_i,u_2,v_1) & q(x_i,u_2,v_2) \end{pmatrix} \right] - \det \begin{pmatrix} q(x_i,u_1,v_1) & q(x_i,u_1,v_2) \\ q(x_i,u_2)q(x_i,v_1) & q(x_i,u_2)q(x_i,v_2) \end{pmatrix} - \det \begin{pmatrix} q(x_i,u_1)q(x_i,v_1) & q(x_i,u_1)q(x_i,v_2) \\ q(x_i,u_2,v_1) & q(x_i,u_2,v_2) \end{pmatrix} = \det \begin{pmatrix} q(u_1,v_1) & q(u_1,v_2) \\ q(u_2,v_1) & q(u_2,v_2) \end{pmatrix} (s + A_{x_i}^1)(s + A_{x_i}^2)(s + A_{x_i}^3).$$

The components of A_{x_i} are the roots of a degree m polynomial

Given $\mathbf{u} = \{u_1, \ldots, u_{m-1}\}$ and $\mathbf{v} = \{v_1, \ldots, v_{m-1}\}$, consider the polynomial of degree m

$$\alpha_{x,\mathbf{u},\mathbf{v}}(s) = s^m \det \begin{pmatrix} q(u_1,v_1) & \cdots & q(u_1,v_{m-1}) \\ \vdots & & \vdots \\ q(u_{m-1},v_1) & \cdots & q(u_{m-1},v_{m-1}) \end{pmatrix} + s^{m-1} \cdots + s \cdots + \ldots$$

whose coefficients are sums of determinants. We have

$$\alpha_{x,\mathbf{u},\mathbf{v}}(s) = \det \begin{pmatrix} q(u_1,v_1) & \cdots & q(u_1,v_{m-1}) \\ \vdots & & \vdots \\ q(u_{m-1},v_1) & \cdots & q(u_{m-1},v_{m-1}) \end{pmatrix} \prod_{t=1}^{m} (s + A_{x_i}^t).$$

The parameters satisfy simple polynomial equations

Consider values $\{x_1, \dots x_k\}$. The following equation holds

$$\det \begin{pmatrix} \prod_{j=1}^{k} A_{x_j}^t & q(x_1, \dots, x_k, v_1) & \dots & q(x_1, \dots, x_k, v_{m-1}) \\ A_{u_1}^t & q(u_1, v_1) & \dots & q(u_1, v_{m-1}) \\ \vdots & \vdots & & \vdots \\ A_{u_{m-1}}^t & q(u_{m-1}, v_1) & \dots & q(u_{m-1}, v_{m-1}) \end{pmatrix}$$

$$= q(x_1, \dots, x_k) \det \begin{pmatrix} q(u_1, v_1) & \cdots & q(u_1, v_{m-1}) \\ \vdots & & \vdots \\ q(u_{m-1}, v_1) & \cdots & q(u_{m-1}, v_{m-1}) \end{pmatrix}.$$

The parameters satisfy simple polynomial equations

For $\{x_1, \dots x_k\} = \{u_0\}$, we have

$$\det \begin{pmatrix} A^t_{u_0} & q(u_0, v_1) & \dots & q(u_0, v_{m-1}) \\ A^t_{u_1} & q(u_1, v_1) & \dots & q(u_1, v_{m-1}) \\ \vdots & \vdots & & \vdots \\ A^t_{u_{m-1}} & q(u_{m-1}, v_1) & \dots & q(u_{m-1}, v_{m-1}) \end{pmatrix} = 0.$$

For $m = 3$ and $\{x_1, \dots x_k\} = \{u_1, u_2\}$, we have

$$\det \begin{pmatrix} A^t_{u_1} A^t_{u_2} & q(u_1, u_2, v_1) & q(u_1, u_2, v_2) \\ A^t_{u_1} & q(u_1, v_1) & q(u_1, v_2) \\ A^t_{u_2} & q(u_2, v_1) & q(u_2, v_2) \end{pmatrix}$$

$$= q(u_1, u_2) \det \begin{pmatrix} q(u_1, v_1) & q(u_1, v_2) \\ q(u_2, v_1) & q(u_2, v_2) \end{pmatrix}.$$

Some determinants have an interpretable decomposition

Consider sets of values $\mathbf{s}_1, \ldots \mathbf{s}_{m-1}$. We have

$$\det \begin{pmatrix} q(\mathbf{s}_1, v_1) & \ldots & q(\mathbf{s}_1, v_{m-1}) \\ \vdots & & \vdots \\ q(\mathbf{s}_{m-1}, v_1) & \ldots & q(\mathbf{s}_{m-1}, v_{m-1}) \end{pmatrix}$$

$$= (\prod_{t=1}^{m} w_t) \det \begin{pmatrix} 1 & A_{v_1} & \ldots & A_{v_{m-1}} \end{pmatrix} \det M,$$

where

$$M = \begin{pmatrix} 1 & \prod_{x \in \mathbf{s}_1} A_x^1 & \ldots & \prod_{x \in \mathbf{s}_{m-1}} A_x^1 \\ \vdots & \vdots & & \vdots \\ 1 & \prod_{x \in \mathbf{s}_1} A_x^m & \ldots & \prod_{x \in \mathbf{s}_{m-1}} A_x^m \end{pmatrix}$$

Simple implicit equations follow

Consider a naive Bayesian distribution with m classes and consider sets of values $\mathbf{s}_1, \ldots, \mathbf{s}_{m'-1}$. If $m' > m$, we have

$$\det \begin{pmatrix} q(\mathbf{s}_1, v_1) & \ldots & q(\mathbf{s}_1, v_{m'-1}) \\ \vdots & & \vdots \\ q(\mathbf{s}_{m'-1}, v_1) & \ldots & q(\mathbf{s}_{m'-1}, v_{m'-1}) \end{pmatrix} = 0.$$

Consider sets of values $\mathbf{s}_1, \ldots, \mathbf{s}_{m-1}$ and $\mathbf{r}_1, \ldots, \mathbf{r}_{m-1}$. We have

$$\det \begin{pmatrix} q(\mathbf{s}_1, v_1) & \ldots & q(\mathbf{s}_1, v_{m-1}) \\ \vdots & & \vdots \\ q(\mathbf{s}_{m-1}, v_1) & \ldots & q(\mathbf{s}_{m-1}, v_{m-1}) \end{pmatrix} \det \begin{pmatrix} q(\mathbf{r}_1, u_1) & \ldots & q(\mathbf{r}_1, u_{m-1}) \\ \vdots & & \vdots \\ q(\mathbf{r}_{m-1}, u_1) & \ldots & q(\mathbf{r}_{m-1}, u_{m-1}) \end{pmatrix}$$
$$= \det \begin{pmatrix} q(\mathbf{s}_1, u_1) & \ldots & q(\mathbf{s}_1, u_{m-1}) \\ \vdots & & \vdots \\ q(\mathbf{s}_{m-1}, u_1) & \ldots & q(\mathbf{s}_{m-1}, u_{m-1}) \end{pmatrix} \det \begin{pmatrix} q(\mathbf{r}_1, v_1) & \ldots & q(\mathbf{r}_1, v_{m-1}) \\ \vdots & & \vdots \\ q(\mathbf{r}_{m-1}, v_1) & \ldots & q(\mathbf{r}_{m-1}, v_{m-1}) \end{pmatrix}$$

Outline

Potential applications of our results

- Compute the set of parameters mapped to a given naive Bayesian distribution
- Estimate parameters from data by applying the previous computation to the distribution of observed frequencies
- Derive sufficient conditions for parameter identifiability and obtain results on the dimensionality of the model
- Building block in the computation of analytic asymptotic approximations to the marginal likelihood of the model
- Building block in model selection and learning of hidden causes

An important hypothesis to compute the parameters

Suppose that we have a distribution p and sets of values

$$\mathbf{t} = \{t_1, \ldots, t_{m-1}\},$$
$$\mathbf{u} = \{u_1, \ldots, u_{m-1}\},$$
$$\mathbf{v} = \{v_1, \ldots, v_{m-1}\}$$

such that

$$\det \begin{pmatrix} q(t_1,u_1) & \ldots & q(t_1,u_{m-1}) \\ \vdots & & \vdots \\ q(t_{m-1},u_1) & \ldots & q(t_{m-1},u_{m-1}) \end{pmatrix} \neq 0,$$
$$\det \begin{pmatrix} q(t_1,v_1) & \ldots & q(t_1,v_{m-1}) \\ \vdots & & \vdots \\ q(t_{m-1},v_1) & \ldots & q(t_{m-1},v_{m-1}) \end{pmatrix} \neq 0,$$
$$\det \begin{pmatrix} q(u_1,v_1) & \ldots & q(u_1,v_{m-1}) \\ \vdots & & \vdots \\ q(u_{m-1},v_1) & \ldots & q(u_{m-1},v_{m-1}) \end{pmatrix} \neq 0.$$

Computation of w from $A_{u_1}, \ldots, A_{u_{m-1}}$

Our hypothesis amounts to

$$\Big(\prod_{i=1}^{m} w_i\Big) \det\begin{pmatrix}1 & A_{t_1} & \ldots & A_{t_{m-1}}\end{pmatrix} \det\begin{pmatrix}1 & A_{u_1} & \ldots & A_{u_{m-1}}\end{pmatrix} \det\begin{pmatrix}1 & A_{v_1} & \ldots & A_{v_{m-1}}\end{pmatrix} \neq 0.$$

Hence, we have

$$w_i = \frac{(-1)^i \det\begin{pmatrix}A_{u_1} & \ldots & A_{u_{m-1}}\end{pmatrix}^{\hat{i}}}{\det\begin{pmatrix}1 & A_{u_1} & \ldots & A_{u_{m-1}}\end{pmatrix}}.$$

Computation of A_x from $A_{u_1}, \ldots, A_{u_{m-1}}$

Since

$$\det \begin{pmatrix} A^t_{u_1} & q(u_1,v_1) & \ldots & q(u_1,v_{m-1}) \\ \vdots & \vdots & & \vdots \\ A^t_{u_{m-1}} & q(u_{m-1},v_1) & \ldots & q(u_{m-1},v_{m-1}) \\ A^t_x & q(x,v_1) & \ldots & q(x,v_{m-1}) \end{pmatrix} = 0,$$

we have, for all values x distinct of $v_1, \ldots, v_{m-1}$,

$$A_x^T = \begin{pmatrix} q(x, v_1) & \ldots & q(x, v_{m-1}) \end{pmatrix} \begin{pmatrix} q(u_1, v_1) & \ldots & q(u_1, v_{m-1}) \\ \vdots & & \vdots \\ q(u_{m-1}, v_1) & \ldots & q(u_{m-1}, v_{m-1}) \end{pmatrix}^{-1} \begin{pmatrix} A_{u_1} & \ldots & A_{u_{m-1}} \end{pmatrix}^T.$$

Computation of $A_{u_1}, \ldots, A_{u_{m-1}}$

Find the roots of the polynomials $\nu_{u_i,\mathbf{t},\mathbf{v}}$ to obtain

$$\{A_{u_1}^1, \ldots, A_{u_1}^m\},$$
$$\vdots$$
$$\{A_{u_{m-1}}^1, \ldots, A_{u_{m-1}}^m\}.$$

Note that these sets are not ordered: we are not able to assign each element to its hidden class.

There is some trivial non-identifiability due to the fact that classes can be permuted freely. To remove this degree of freedom from the analysis, we order the set $\{A_{u_1}^1, \ldots, A_{u_1}^m\}$ arbitrarily.

Computation of $A_{u_1}, \ldots, A_{u_{m-1}}$: a brute force approach

For each ordering of each set $\{A^1_{u_i}, \ldots, A^m_{u_i}\}$ with $i = 2, \ldots, m-1$

1. compute a candidate parameter with the previous procedure
2. test if the candidate satisfies the constraints to be a parameter and if it is mapped to the distribution

However, there are $(m!)^{m-2}$ candidate parameters to test.

Corollary: under our hypothesis, there are at most $(m!)^{m-1}$ parameters mapped to the distribution.

Computation of $A_{u_1}, \ldots, A_{u_{m-1}}$: a second approach

We have

$$\det M\left(\sum_{p=1}^{m}\prod_{j=1}^{k} A^{p}_{x_{i_j}}\right) = q(x_{i_1}, \ldots, x_{i_k}) \det M + \sum_{a=1}^{m-1}\sum_{b=1}^{m-1} (-1)^{a+b} q(x_{i_1}, \ldots, x_{i_k}, t_a, v_b) \det M^{\hat{a}}_{\hat{b}},$$

where

$$M = \begin{pmatrix} q(t_1, v_1) & \ldots & q(t_1, v_{m-1}) \\ \vdots & & \vdots \\ q(t_{m-1}, v_1) & \ldots & q(t_{m-1}, v_{m-1}) \end{pmatrix} \qquad (1)$$

We can constraint the orderings to those satisfying the above equation with $\{x_{i_1}, \ldots, x_{i_k}\} = \{u_1, u_j\}$.

Computation of $A_{u_1}, \ldots, A_{u_{m-1}}$

The previous algorithm do not make use of all our theoretical results. For $m = 3$, recall that we have

$$\det \begin{pmatrix} A_{u_1}^t A_{u_2}^t & q(u_1, u_2, v_1) & q(u_1, u_2, v_2) \\ A_{u_1}^t & q(u_1, v_1) & q(u_1, v_2) \\ A_{u_2}^t & q(u_2, v_1) & q(u_2, v_2) \end{pmatrix} = q(u_1, u_2) \det \begin{pmatrix} q(u_1, v_1) & q(u_1, v_2) \\ q(u_2, v_1) & q(u_2, v_2) \end{pmatrix}.$$

We can derive $A_{u_2}^t$ from $A_{u_1}^t$ by solving the above equation.

We are currently investigating how to make use of all our results in the general case.

The inversion algorithms can be adapted to estimate parameters

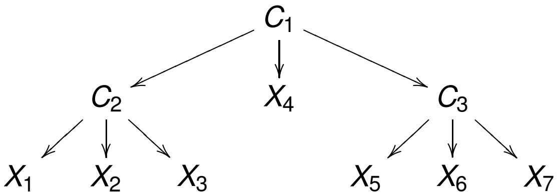
- Basic idea: apply the inversion algorithm to the observed distribution $\hat{p}$.
- Instead of testing whether a candidate parameter is mapped to p, we find the parameter minimizing the relative entropy to $\hat{p}$.
- Suppose that the unknown p is a naive Bayesian distribution with m classes satisfying our inversion assumption. As the sample size increases, $\hat{p}$ converges to p and, by continuity, our estimate converges to a true parameter mapped to p.

Practical issues

The estimation procedure has several issues:

- The computational complexity grows extremely fast with m, but linearly with n.
- The estimates are numerically unstable and require large sample sizes. For smaller sample sizes, there may not even be a single candidate parameter satisfying the parameter constraints.
- There are many degrees of freedom in the choice of $\mathbf{t}, \mathbf{u}$ and $\mathbf{v}$. Asymptotically, any choice is suitable. For small sample size, it is probably important.
- The results are not competitive with the E.M. algorithm.

Extension to hierarchical latent class models



The parameters mapped to a HLC distribution with the above structure can be derived from the parameters mapped to the naive Bayesian distributions over

$$\{X_1, X_2, X_3\}$$
$$\{X_1, X_4, X_5\}$$
$$\{X_5, X_6, X_7\}$$

obtained by marginalization.

Conclusion

We presented some simple and interesting polynomial equations constraining a naive Bayesian distribution and its parameters. These results may be applied to

- compute the parameters mapped to a naive Bayesian distribution,
- estimate parameters from data.

The implicit equation

$$\det \begin{pmatrix} q(u_1, v_1) & \dots & q(u_1, v_{m'-1}) \\ \vdots & & \vdots \\ q(u_{m'-1}, v_1) & \dots & q(u_{m'-1}, v_{m'-1}) \end{pmatrix} = 0$$

holding for $m' > m$ is similar to a tetrad constraint. A future research direction would investigate whether the constraint can indeed be used to learn hidden causes from data.