Building Bayes Networks:

Parameter Learning

Learning Naive Bayes Classifier

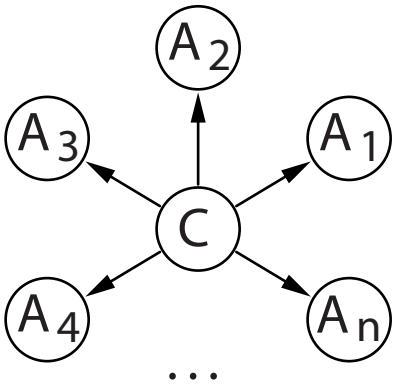
Given: A database of samples from domain of interest.
The graph underlying a graphical model for the domain.
Desired: Good values for the numeric parameters of the model.

Example: Naive Bayes Classifiers

A naive Bayes classifier is a Bayesian network with star-like structure.

The class attribute is the only unconditional attribute.

All other attributes are conditioned on the class only



The structure of a naive Bayes classifier is fixed once the attributes have been selected. The only remaining task is to estimate the parameters of the needed probability distributions.

Probabilistic Classification

A classifier is an algorithm that assigns a class from a predefined set to a case or object, based on the values of descriptive attributes.

An optimal classifier maximizes the probability of a correct class assignment.

- Let C be a class attribute with $\mathrm{dom}(C) = \{c_1, \ldots, c_{n_C}\}$, which occur with probabilities p_i, $1 \leq i \leq n_C$.
- Let q_i be the probability with which a classifier assigns class c_i. ($q_i \in \{0, 1\}$ for a deterministic classifier)
- The probability of a correct assignment is

$$P(\text{correct assignment}) = \sum_{i=1}^{n_C} p_i q_i.$$

- Therefore the best choice for the q_i is

$$q_i = \begin{cases} 1, & \text{if } p_i = \max_{k=1}^{n_C} p_k, \\ 0, & \text{otherwise.} \end{cases}$$

Probabilistic Classification

Consequence: An optimal classifier should assign the **most probable class**.

This argument does not change if we take descriptive attributes into account.

- Let $U = \{A_1, \ldots, A_m\}$ be a set of descriptive attributes with domains $\mathrm{dom}(A_k)$, $1 \leq k \leq m$.
- Let $A_1 = a_1, \ldots, A_m = a_m$ be an instantiation of the descriptive attributes.
- An optimal classifier should assign the class c_i for which

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) =$$
$$\max_{j=1}^{n_C} P(C = c_j \mid A_1 = a_1, \ldots, A_m = a_m)$$

Problem: We cannot store a class (or the class probabilities) for every possible instantiation $A_1 = a_1, \ldots, A_m = a_m$ of the descriptive attributes. (The table size grows exponentially with the number of attributes.)

Therefore: **Simplifying assumptions are necessary.**

Bayes' Rule and Bayes' Classifiers

Bayes' classifiers: Compute the class probabilities as

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) =$$
$$\frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_m = a_m)}.$$

Looks unreasonable at first sight: Even more probabilities to store.

Naive Assumption:
The descriptive attributes are conditionally independent given the class.

Bayes' Rule:

$$P(C = c_i \mid \omega) = \frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_m = a_m)}$$

$\leftarrow p_0$
abbrev. for the
normalizing constant

Chain Rule of Probability:

$$P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid A_1 = a_1, \ldots, A_{k-1} = a_{k-1}, C = c_i)$$

Conditional Independence Assumption:

$$P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid C = c_i)$$

Naive Bayes Classifiers (continued)

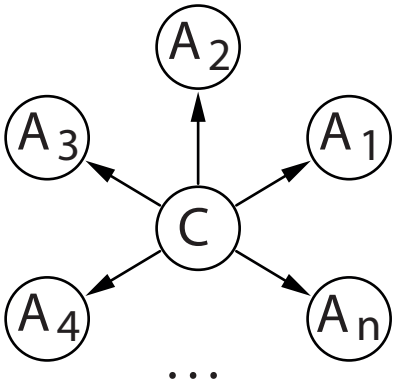
Consequence: Manageable amount of data to store.
Store distributions $P(C = c_i)$ and $\forall 1 \leq k \leq m : P(A_k = a_k \mid C = c_i)$.

Classification: Compute for all classes c_i

$$P(C = c_i | A_1 = a_1, \ldots, A_m = a_m) \cdot p_0 = P(C = c_i) \cdot \prod_{j=1}^{n} P(A_j = a_j | C = c_i)$$

and predict the class c_i for which this value is largest.

Relation to Bayesian Networks:



Decomposition formula:

$$P(C = c_i, A1 = a_1, \ldots, A_n = a_n)$$
$$= P(C = c_i) \cdot \prod_{j=1}^{n} P(A_j = a_j | C = c_i)$$

Estimation of Probabilities:

Nominal/Categorical Attributes

$$\hat{P}(A_k = a_k \mid C = c_i) = \frac{\#(A_k = a_k, C = c_i) + \gamma}{\#(C = c_i) + n_{A_k}\gamma}$$

$\#(\varphi)$ is the number of example cases that satisfy the condition φ

n_{A_j} is the number of values of the attribute A_j.

γ is called **Laplace correction**

$\gamma = 0$: Maximum likelihood estimation.

Common choices: $\gamma = 1$ or $\gamma = \frac{1}{2}$.

Laplace correction help to avoid problems with attribute values that do not occur with some class in the given data.

It also introduces a bias towards a uniform distribution.

Estimation of Probabilities:

Metric/Numeric Attributes: Assume a normal distribution.

$$P(A_k = a_k \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_k(c_i)} \exp\left(-\frac{(a_k - \mu_k(c_i))^2}{2\sigma_k^2(c_i)}\right)$$

Estimate of mean value

$$\hat{\mu}_k(c_i) = \frac{1}{\#(C = c_i)} \sum_{j=1}^{\#(C=c_i)} a_k(j)$$

Estimate of variance

$$\hat{\sigma}_k^2(c_i) = \frac{1}{\xi} \sum_{j=1}^{\#(C=c_i)} (a_k(j) - \hat{\mu}_k(c_i))^2$$

$\xi = \#(C = c_i)$: Maximum likelihood estimation
$\xi = \#(C = c_i) - 1$: Unbiased estimation

Naive Bayes Classifiers: Simple Example 1

No	Sex	Age	Blood pr.	Drug
1	male	20	normal	A
2	female	73	normal	B
3	female	37	high	A
4	male	33	low	B
5	female	48	high	A
6	male	29	normal	A
7	female	52	normal	B
8	male	42	low	B
9	male	61	normal	B
10	female	30	normal	A
11	female	26	low	B
12	male	54	high	A

$P(\text{Drug})$	A	B
	0.5	0.5

$P(\text{Sex} \mid \text{Drug})$	A	B
male	0.5	0.5
female	0.5	0.5

$P(\text{Age} \mid \text{Drug})$	A	B
μ	36.3	47.8
σ^2	161.9	311.0

$P(\text{Blood Pr.} \mid \text{Drug})$	A	B
low	0	0.5
normal	0.5	0.5
high	0.5	0

A simple database and estimated (conditional) probability distributions.

Naive Bayes Classifiers: Simple Example 1

$$
\begin{aligned}
&P(\text{Drug A} \mid \text{male, 61, normal}) \\
&\quad = \quad c_1 \cdot P(\text{Drug A}) \cdot P(\text{male} \mid \text{Drug A}) \cdot P(61 \mid \text{Drug A}) \cdot P(\text{normal} \mid \text{Drug A}) \\
&\quad \approx \quad c_1 \cdot 0.5 \cdot 0.5 \cdot 0.004787 \cdot 0.5 \quad = \quad c_1 \cdot 5.984 \cdot 10^{-4} \quad = \quad 0.219 \\
&P(\text{Drug B} \mid \text{male, 61, normal}) \\
&\quad = \quad c_1 \cdot P(\text{Drug B}) \cdot P(\text{male} \mid \text{Drug B}) \cdot P(61 \mid \text{Drug B}) \cdot P(\text{normal} \mid \text{Drug B}) \\
&\quad \approx \quad c_1 \cdot 0.5 \cdot 0.5 \cdot 0.017120 \cdot 0.5 \quad = \quad c_1 \cdot 2.140 \cdot 10^{-3} \quad = \quad 0.781 \\
&P(\text{Drug A} \mid \text{female, 30, normal}) \\
&\quad = \quad c_2 \cdot P(\text{Drug A}) \cdot P(\text{female} \mid \text{Drug A}) \cdot P(30 \mid \text{Drug A}) \cdot P(\text{normal} \mid \text{Drug A}) \\
&\quad \approx \quad c_2 \cdot 0.5 \cdot 0.5 \cdot 0.027703 \cdot 0.5 \quad = \quad c_2 \cdot 3.471 \cdot 10^{-3} \quad = \quad 0.671 \\
&P(\text{Drug B} \mid \text{female, 30, normal}) \\
&\quad = \quad c_2 \cdot P(\text{Drug B}) \cdot P(\text{female} \mid \text{Drug B}) \cdot P(30 \mid \text{Drug B}) \cdot P(\text{normal} \mid \text{Drug B}) \\
&\quad \approx \quad c_2 \cdot 0.5 \cdot 0.5 \cdot 0.013567 \cdot 0.5 \quad = \quad c_2 \cdot 1.696 \cdot 10^{-3} \quad = \quad 0.329
\end{aligned}
$$

Naive Bayes Classifiers: Simple Example 2

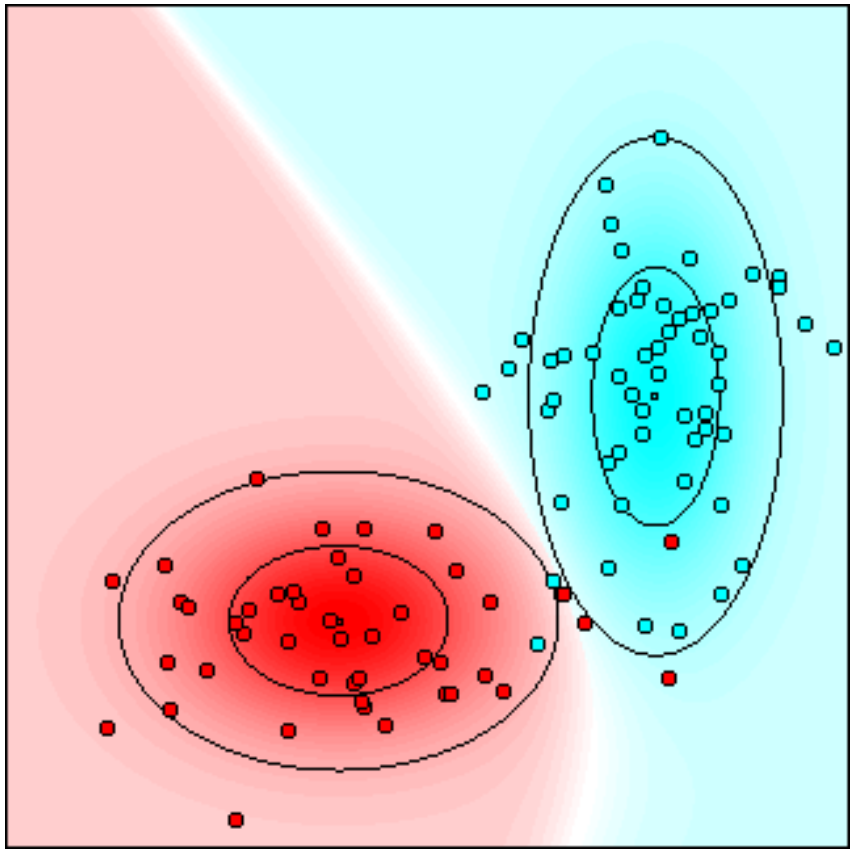
100 data points, 2 classes

Small squares: mean values

Inner ellipses:
one standard deviation

Outer ellipses:
two standard deviations

Classes overlap:
classification is not perfect



Naive Bayes Classifier

Naive Bayes Classifiers: Simple Example 3

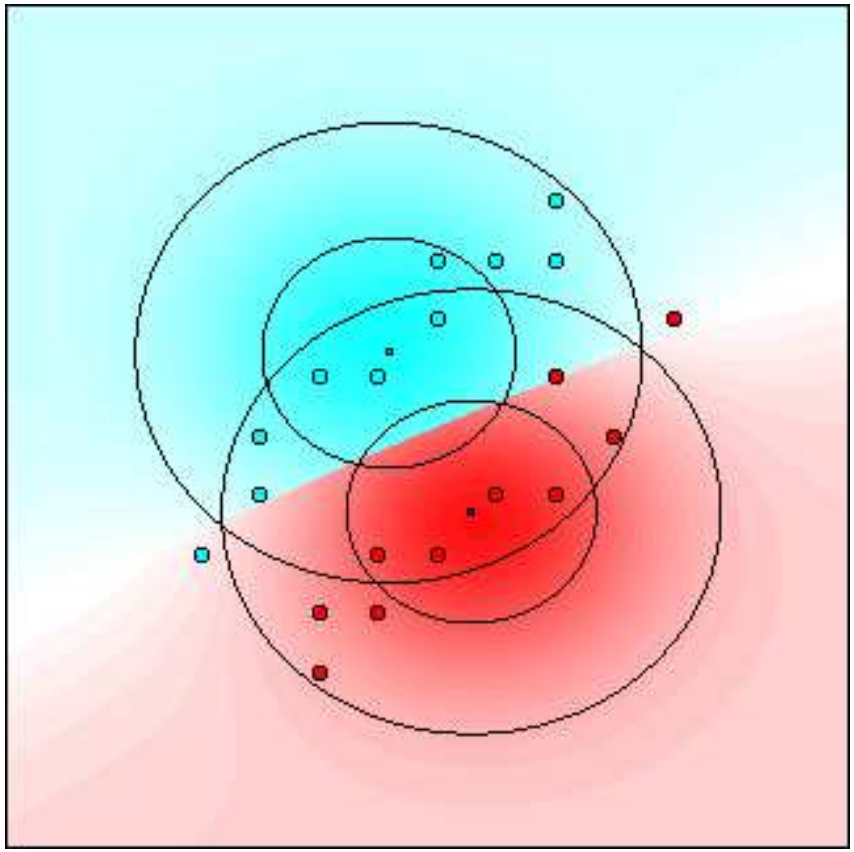
20 data points, 2 classes

Small squares: mean values

Inner ellipses:
one standard deviation

Outer ellipses:
two standard deviations

Attributes are not conditionally independent given the class



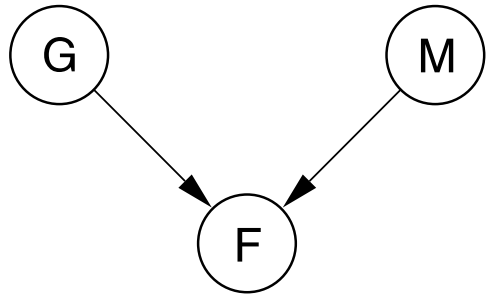
Naive Bayes Classifier

Learning the parameters of a Graphical Model

Probability values can be estimated using methods of inductive statistics.

	P(G)
$a_{11} = \mathsf{g}$	
$a_{12} = \overline{\mathsf{g}}$	

	P(M)
$a_{12} = \mathsf{m}$	
$a_{22} = \overline{\mathsf{m}}$	



P(F|G,M)	g, m	$\mathsf{g}, \overline{\mathsf{m}}$	$\overline{\mathsf{g}}, \mathsf{m}$	$\overline{\mathsf{g}}, \overline{\mathsf{m}}$
$a_{31} = \mathsf{f}$				
$a_{32} = \overline{\mathsf{f}}$				

$$V = \{\mathsf{G}, \mathsf{M}, \mathsf{F}\}$$
$$\mathrm{dom}(\mathsf{G}) = \{\mathsf{g}, \overline{\mathsf{g}}\}$$
$$\mathrm{dom}(\mathsf{M}) = \{\mathsf{m}, \overline{\mathsf{m}}\}$$
$$\mathrm{dom}(\mathsf{F}) = \{\mathsf{f}, \overline{\mathsf{f}}\}$$

Learning the parameters of a Graphical Model

Flu G Malaria M Fever F	$\overline{\mathsf{g}}$ $\overline{\mathsf{m}}$ $\overline{\mathsf{f}}$	$\overline{\mathsf{g}}$ $\overline{\mathsf{m}}$ f	$\overline{\mathsf{g}}$ m $\overline{\mathsf{f}}$	$\overline{\mathsf{g}}$ m f	g $\overline{\mathsf{m}}$ $\overline{\mathsf{f}}$	g $\overline{\mathsf{m}}$ f	g m $\overline{\mathsf{f}}$	g m f
#	34	6	2	8	16	24	0	10

Database D with 100 entries for 3 attributes.

As the structure given by the graph of the previous slide suggests, the probability of $P(g, m, f)$ can be computed by:

$$P(\mathsf{g}, \mathsf{m}, \mathsf{f}) = P(\mathsf{g})P(\mathsf{m})P(\mathsf{f} \mid \mathsf{g}, \mathsf{m})$$

Estimates for these probabilities can be calculated, e.g. using the database

$$\hat{P}(\mathsf{f} \mid \mathsf{g}, \mathsf{m}) = \frac{\hat{P}(\mathsf{f}, \mathsf{g}, \mathsf{m})}{\hat{P}(\mathsf{g}, \mathsf{m})} = \frac{\frac{\#(\mathsf{g},\mathsf{m},\mathsf{f})}{|D|}}{\frac{\#(\mathsf{g},\mathsf{m})}{|D|}} = \frac{\#(\mathsf{g}, \mathsf{m}, \mathsf{f})}{\#(\mathsf{g}, \mathsf{m})} = \frac{10}{10} = 1.00$$

$$\hat{P}(\mathsf{f} \mid \overline{\mathsf{g}}, \overline{\mathsf{m}}) = \frac{\hat{P}(\mathsf{f}, \overline{\mathsf{g}}, \overline{\mathsf{m}})}{\hat{P}(\overline{\mathsf{g}}, \overline{\mathsf{m}})} = \frac{\frac{\#(\overline{\mathsf{g}},\overline{\mathsf{m}},\mathsf{f})}{|D|}}{\frac{\#(\overline{\mathsf{g}},\overline{\mathsf{m}})}{|D|}} = \frac{\#(\overline{\mathsf{g}}, \overline{\mathsf{m}}, \mathsf{f})}{\#(\overline{\mathsf{g}}, \overline{\mathsf{m}})} = \frac{6}{40} = 0.15$$

Likelihood of a Database

Let B_P be the description of the parameters, B_S be the given structure and D the data.
The likelihood of the calculated probabilities $P(D \mid B_S, B_P)$ can be computed under presence of three assumptions:

1. The data generation process can be described exactly by a Bayesian network (B_S, B_P)
2. The single tuples of the dataset are independent of each other.
3. All tuples are complete, therefore no missing values hinder the probability inference

The first assumption legitimates the search of an appropriate bayesian network.

The second assumption is required for an unbiased observation of dataset tuples.

Assumption three ensures the inference of B_P using D and B_S as shown on the previous slides.

Likelihood of a Database

Flu G	$\overline{\mathsf{g}}$	$\overline{\mathsf{g}}$	$\overline{\mathsf{g}}$	$\overline{\mathsf{g}}$	g	g	g	g
Malaria M	$\overline{\mathsf{m}}$	$\overline{\mathsf{m}}$	m	m	$\overline{\mathsf{m}}$	$\overline{\mathsf{m}}$	m	m
Fever F	$\overline{\mathsf{f}}$	f	$\overline{\mathsf{f}}$	f	$\overline{\mathsf{f}}$	f	$\overline{\mathsf{f}}$	f
#	34	6	2	8	16	24	0	10

Database D with 100 entries c_h for 3 attributes.

$$P(D \mid B_S, B_P) \quad = \quad \prod_{h=1}^{100} P(c_h \mid B_S, B_P)$$

$$= \overbrace{P(\mathsf{g},\mathsf{m},\mathsf{f}) \cdot \ldots \cdot P(\mathsf{g},\mathsf{m},\mathsf{f})}^{\text{Case 1}\ \ldots\ \text{Case 10}} \quad \cdots \quad \overbrace{P(\overline{\mathsf{g}},\mathsf{m},\mathsf{f}) \cdot \ldots \cdot P(\overline{\mathsf{g}},\mathsf{m},\mathsf{f})}^{\text{Case 51}\ \ldots\ \text{Case 58}} \quad \cdots \quad \overbrace{P(\overline{\mathsf{g}},\overline{\mathsf{m}},\overline{\mathsf{f}}) \cdot \ldots \cdot P(\overline{\mathsf{g}},\overline{\mathsf{m}},\overline{\mathsf{f}})}^{\text{Case 67}\ \ldots\ \text{Case 100}}$$

$$= \overbrace{P(\mathsf{g},\mathsf{m},\mathsf{f})^{10}}^{\text{10 times}} \quad \cdots \quad \overbrace{P(\overline{\mathsf{g}},\mathsf{m},\mathsf{f})^{8}}^{\text{8 times}} \quad \cdots \quad \overbrace{P(\overline{\mathsf{g}},\overline{\mathsf{m}},\overline{\mathsf{f}})^{34}}^{\text{34 times}}$$

$$= P(\mathsf{f} \mid \mathsf{g},\mathsf{m})^{10} P(\mathsf{g})^{10} P(\mathsf{m})^{10} \quad \cdots \quad P(\mathsf{f} \mid \overline{\mathsf{g}},\mathsf{m})^{8} P(\overline{\mathsf{g}})^{8} P(\mathsf{m})^{8} \quad \cdots \quad P(\overline{\mathsf{f}} \mid \overline{\mathsf{g}},\overline{\mathsf{m}})^{34} P(\overline{\mathsf{g}})^{34} P(\overline{\mathsf{m}})^{34}$$

Likelihood of a Database

$$P(D \mid B_S, B_P) = \prod_{h=1}^{100} P(c_h \mid B_S, B_P)$$

$$= P(\mathsf{f} \mid \mathsf{g}, \mathsf{m})^{10} P(\overline{\mathsf{f}} \mid \mathsf{g}, \mathsf{m})^{0} P(\mathsf{f} \mid \mathsf{g}, \overline{\mathsf{m}})^{24} P(\overline{\mathsf{f}} \mid \mathsf{g}, \overline{\mathsf{m}})^{16}$$

$$\cdot P(\mathsf{f} \mid \overline{\mathsf{g}}, \mathsf{m})^{8} P(\overline{\mathsf{f}} \mid \overline{\mathsf{g}}, \mathsf{m})^{2} P(\mathsf{f} \mid \overline{\mathsf{g}}, \overline{\mathsf{m}})^{6} P(\overline{\mathsf{f}} \mid \overline{\mathsf{g}}, \overline{\mathsf{m}})^{34}$$

$$\cdot P(\mathsf{g})^{50} P(\overline{\mathsf{g}})^{50} P(\mathsf{m})^{20} P(\overline{\mathsf{m}})^{80}$$

The last equation shows the principle of reordering the factors:

- First, we sort by attributes (here: F, G then M).
- Within the same attributes, factors are grouped by the parent attributes' values combinations (here: for F: (g, m), $(\mathsf{g}, \overline{\mathsf{m}})$, $(\overline{\mathsf{g}}, \mathsf{m})$ and $(\overline{\mathsf{g}}, \overline{\mathsf{m}})$).
- Finally, it is sorted by attribute values (here: for F: first f, then $\overline{\mathsf{f}}$).
- Bayes Theorem gives the likelihood $P(B_P \mid D, B_S)$.
- Maximum likelihood approach gives a good estimate for $\hat{B}_P$.

Likelihood of a Database

General likelihood of a database D given a known Bayesian network structure B_S and the parameters B_P:

$$P(D \mid B_S, B_P) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}}$$

General potential table:

A_i	Q_{i1}	$\cdots$	Q_{ij}	$\cdots$	Q_{iq_i}
a_{i1}	θ_{i11}	$\cdots$	θ_{ij1}	$\cdots$	θ_{iq_i1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_{ik}	θ_{i1k}	$\cdots$	θ_{ijk}	$\cdots$	θ_{iq_ik}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_{ir_i}	θ_{i1r_i}	$\cdots$	θ_{ijr_i}	$\cdots$	$\theta_{iq_ir_i}$

$$P(A_i = a_{ik} \mid \text{parents}(A_i) = Q_{ij}) = \theta_{ijk}$$

$$\sum_{k=1}^{r_i} \theta_{ijk} = 1$$