Building Bayes Networks:

Parameter Learning

Learning Naive Bayes Classifier

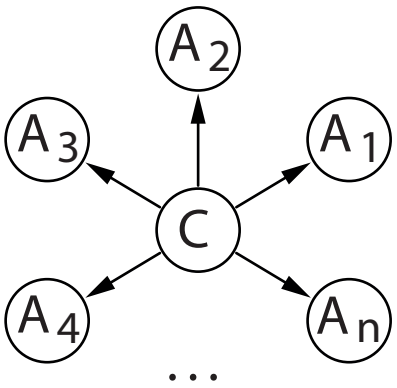
Given: A database of samples from domain of interest.
The graph underlying a graphical model for the domain.
Desired: Good values for the numeric parameters of the model.

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:

$$P(C = c_i \mid \omega) = \frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_m = a_m)} \leftarrow p_0$$

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)

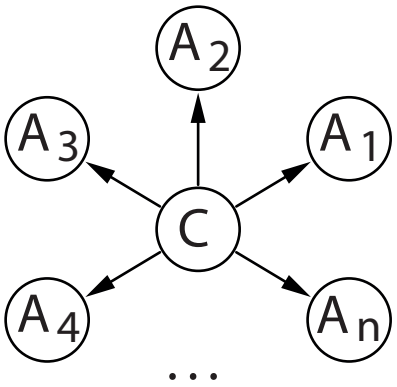
Consequence: Manageable amount of data to store.
Store distributions $P(C = c_i)$ and $\forall 1 \leq k \leq m : P(A_k = a_k \mid C = c_i)$.

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:



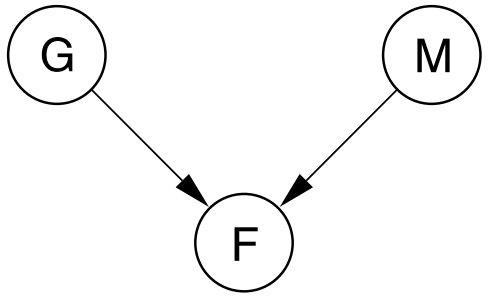
Decomposition formula:

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

Learning the parameters of a Graphical Model

$A_1 = \mathsf{G}$	$Q_{11} = \phi$
$a_{11} = \mathsf{g}$	
$a_{12} = \overline{\mathsf{g}}$	

$A_2 = \mathsf{M}$	$Q_{21} = \phi$
$a_{12} = \mathsf{m}$	
$a_{22} = \overline{\mathsf{m}}$	



$A_3 = \mathsf{F}$	$Q_{31} = (\mathsf{g}, \mathsf{m})$	$Q_{32} = (\mathsf{g}, \overline{\mathsf{m}})$	$Q_{33} = (\overline{\mathsf{g}}, \mathsf{m})$	$Q_{34} = (\overline{\mathsf{g}}, \overline{\mathsf{m}})$
$a_{31} = \mathsf{f}$				
$a_{32} = \overline{\mathsf{f}}$				

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

The potential tables' layout is determined by the graph structure.

The parameters (i. e. the table entries) can be easily estimated from the database, e. g.:

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

Likelihood of a Database

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:

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

Necessary conditional probabilities can be calculated using the Bayes-Theorem:

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

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

Likelihood of a Database (2)

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 (3)

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 for 3 attributes.

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

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

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

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

Likelihood of a Database (4)

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

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

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

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

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 $\overline{\mathsf{f}}$).

Likelihood of a Database (5)

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}	$\cdots$	$\theta_{iq_ir_i}$

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

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