Business Analytics

Naïve Bayes and Bayes Networks

Prof. Bichler

Decision Sciences & Systems

Department of Informatics

Technische Universität München

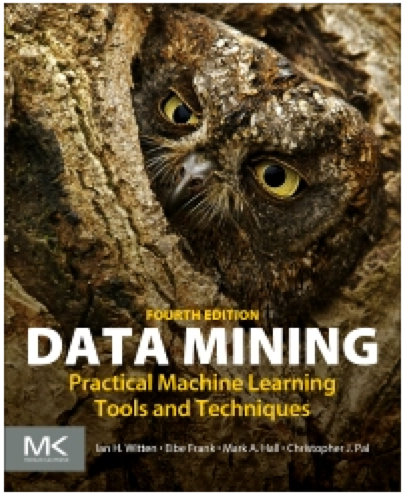
Course Content

- Introduction
- Regression Analysis
- Regression Diagnostics
- Logistic and Poisson Regression
- **Naive Bayes and Bayesian Networks**
- Decision Tree Classifiers
- Data Preparation and Causal Inference
- Model Selection and Learning Theory
- Ensemble Methods and Clustering
- High-Dimensional Problems
- Association Rules and Recommenders
- Neural Networks

Recommended Literature

- **Data Mining: Practical Machine Learning Tools and Techniques**
  - Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher Pal
  - http://www.cs.waikato.ac.nz/ml/weka/book.html
  - Section: 4.1, 4.2, 9.1, 9.2

- **Alternative literature**
  - Machine Learning
    - Tom M. Mitchell, 1997
  - Data Mining: Introductory and Advanced Topics
    - Margaret H. Dunham, 2003


FOURTH EDITION
DATA MINING
Practical Machine Learning
Tools and Techniques
MK
Ian H. Witten • Eibe Frank • Mark A. Hall • Christopher J. Pal

Formal Definition of Classification

- **Classification**: Given a database $D = \{x_1, x_2, \dots, x_n\}$ of tuples (items, records) and a set of classes $C = \{C_1, C_2, \dots, C_m\}$, the classification problem is to define a mapping $f: D \to C$ where each x_i is assigned to one class. A class, C_j, contains precisely those tuples mapped to it; that is, $C_j = \{x_i \mid f(x_i) = C_j, 1 \leq i \leq n,$ and $x_i \in D\}$.
  - The *logistic regression* is used for classification.

- **Prediction** is similar, but usually implies a mapping to numeric values instead of a class C_j

Example Applications

- Determine if a bank customer for a loan is a low, medium, or high risk customer
- Churn prediction (typically a classification task)
- Determine if a sequence of credit card purchases indicates questionable behavior
- Identify customers that may be willing to purchase particular insurance policies
- Identify patterns of gene expression that indicate the patient has cancer
- Identify spam mail
- …

Algorithms for Classification

- Logistic Regression
- Statistical Modeling (e.g., Naïve Bayes)
- Decision Trees: Divide and Conquer
- Classification Rules (e.g. PRISM)
- Instance-Based Learning (e.g. kNN)
- Support Vector Machines
- …

Naïve Bayes Classifier

- Naive Bayes classifier takes all attributes into account
- Assumptions
  - All attributes equally important
  - All attributes independent
    - This means that knowledge about the value of a particular attribute doesn’t tell us anything about the value of another attribute
- Although based on assumptions that are almost never correct, this scheme works often well in practice!

Bayes Theorem: Some Notation

- Let $\Pr[e]$ represent the prior or unconditional probability that proposition e is true.
  - Example: Let e represent that a customer is a high credit risk.
    $\Pr[e] = 0.1$ means that there is a 10% chance a customer is a high credit risk.

- Probabilities of events change when we know something about the world
  - The notation $\Pr[e|h]$ represents the conditional or posterior probability of e
  - Read "the probability of e given that all we know is h."
    - $\Pr[e = high\, risk|\, h = unemployed] = 0.60$

- The notation $\Pr[E]$ is used to represent the probability distribution of all possible values of a random variable E
  - E.g.: $\Pr[Risk] = < 0.7,0.2,0.1 >$

Conditional Probability

- Imagine that 5% of people of a given population own at least one TV. 2% of people own at least one TV and at least one computer. What is the probability that someone will own a computer, given that they also have a TV?

- Let a = “TV owner”, b = “computer owner”, then:
  - $\Pr[a] = 0.05; \Pr[a \cap b] = 0.02$
  - $\Pr[b \mid a] = \Pr[a \cap b] / \Pr[a] = 0.4$

- If events a and b do not influence each other, then
  - $\Pr[a \mid b] = \Pr[a]$ and $\Pr[a \cap b] = \Pr[a] P[b]$
  - For example, throw a coin two times

Conditional Probability and Bayes Rule

- The product rule for conditional probabilities
  - $\Pr[e|h] = \Pr[e \cap h]/\Pr[h]$
  - $\Pr[e \cap h] = \Pr[e|h]\Pr[h] = \Pr[h|e]\Pr[e]$ (product rule)
  - $\Pr[e \cap h] = \Pr[e]\Pr[h]$ (for independent random variables)

- Bayes' rule relates conditional probabilities
  - $\Pr[e \cap h] = \Pr[e|h]\Pr[h]$
  - $\Pr[e \cap h] = \Pr[h|e]\Pr[e]$

$$\boxed{\Pr[h|e] = \frac{\Pr[e|h]\Pr[h]}{\Pr[e]}}$$

Bayes‘ Theorem: An Example

Number of occurences	Beard: B	No beard: ¬B	sum
Astigmatic: A	2 ☺☺	3 ☺☺☺	5
Not astigmatic: ¬A	6 ☺☺☺☺☺☺	9 ☺☺☺☺☺☺☺☺☺	15
sum	8	12	20

Pr[A] and Pr[B] are known:

$$\Pr[\mathrm{A}|\mathrm{B}] = \Pr[A \cap B]/\Pr[B] = \frac{2}{8} = \frac{1}{4}$$

$$\Pr[B|\mathrm{A}] = \frac{2}{5} \qquad \Pr[B|\neg \mathrm{A}] = \frac{6}{15}$$

$$\Pr[A|\mathrm{B}] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} = \frac{\frac{2}{5}\cdot\frac{5}{20}}{\frac{8}{20}} = \frac{1}{4}$$

If Pr[B] is unknown, then use the law of total probability:

$$\Pr[\neg A|\mathrm{B}] = \frac{\Pr[B|\neg A]\Pr[\neg A]}{\Pr[B]} = \frac{\frac{6}{15}\cdot\frac{15}{20}}{\frac{8}{20}} = \frac{6/20}{8/20} = \frac{3}{4}$$

$$\Pr[A|\mathrm{B}] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B|A]\Pr[A] + \Pr[B|\neg A]\Pr[\neg A]} = \frac{\frac{2}{5}\cdot\frac{5}{20}}{\frac{2}{5}\cdot\frac{5}{20} + \frac{6}{15}\cdot\frac{15}{20}} = \frac{1}{4}$$

Source: https://en.wikipedia.org/wiki/Bayes_theorem

Does a patient have Corona or not after a positive test?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, 0.8% of the entire population currently have Corona. The prior probability of a positive test is not given.

Does a patient have Corona or not after a positive test?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, 0.8% of the entire population currently have Corona. The prior probability of a positive test is not given.

Given:

$$\Pr[\text{corona}] = .008, \qquad \Pr[\neg\text{corona}] = .992, \qquad \Pr[+|\text{corona}] = .98, \qquad \Pr[-|\text{corona}] = .02$$

Therefore:

$$\Pr[+|\neg\text{corona}] = .03, \qquad \Pr[-|\neg\text{corona}] = .97$$

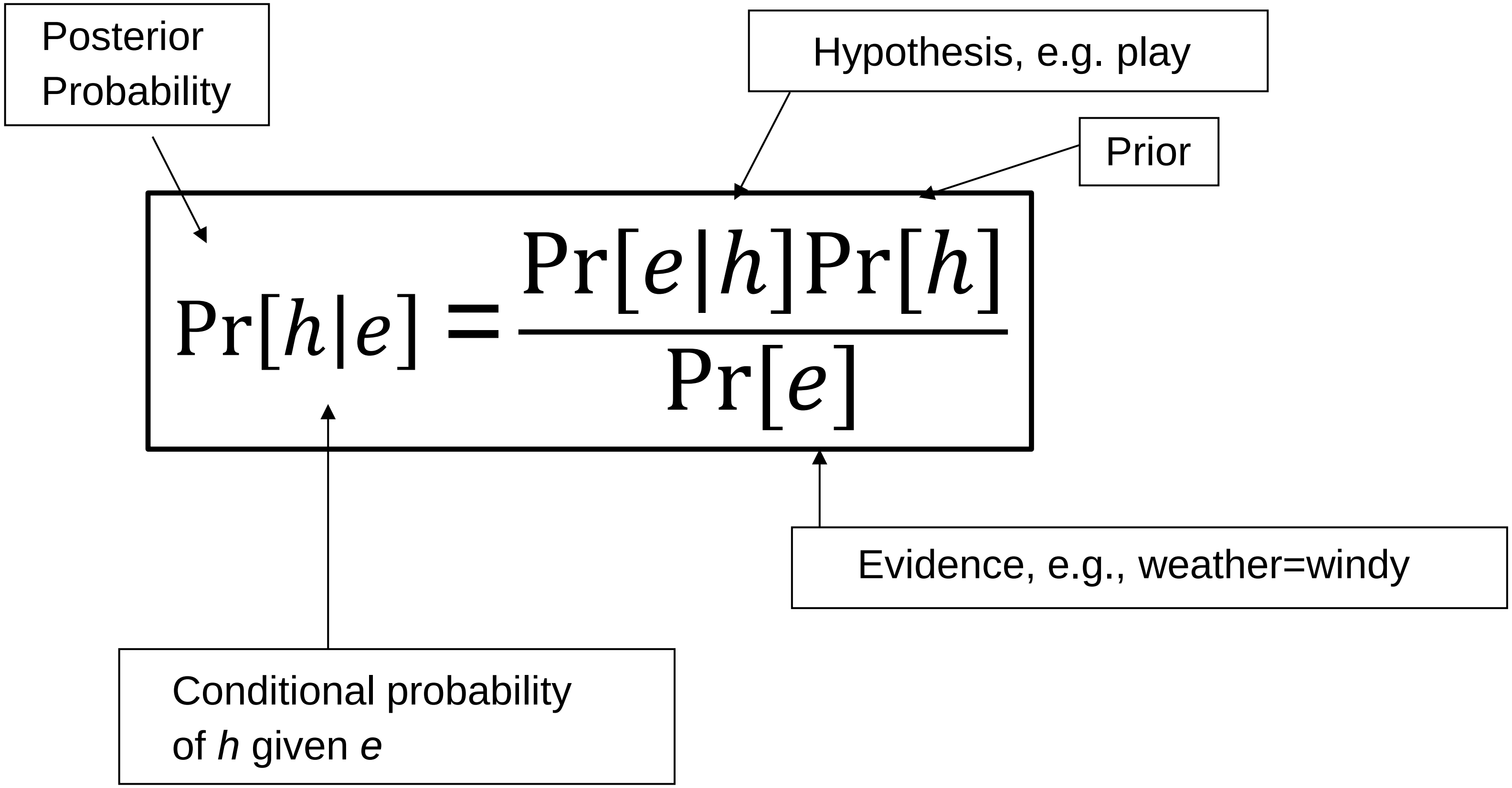
$$\Pr[+|\text{corona}]\Pr[\text{corona}] = 0.98 \cdot 0.008 = 0.0078$$

$$\Pr[+|\neg\text{corona}]\Pr[\neg\text{corona}] = 0.03 \cdot 0.992 = 0.0298$$

$$\Pr[+] = 0.0078 + 0.0298 = 0.0376$$

$$\Pr[\neg\text{corona}|+] = \frac{\Pr[+|\neg\text{corona}]\Pr[\neg\text{corona}]}{\Pr[+]} = 0.79$$

Bayes Theorem



Posterior Probability

Hypothesis, e.g. play

Prior

$$\boxed{\Pr[h|e] = \frac{\Pr[e|h]\Pr[h]}{\Pr[e]}}$$

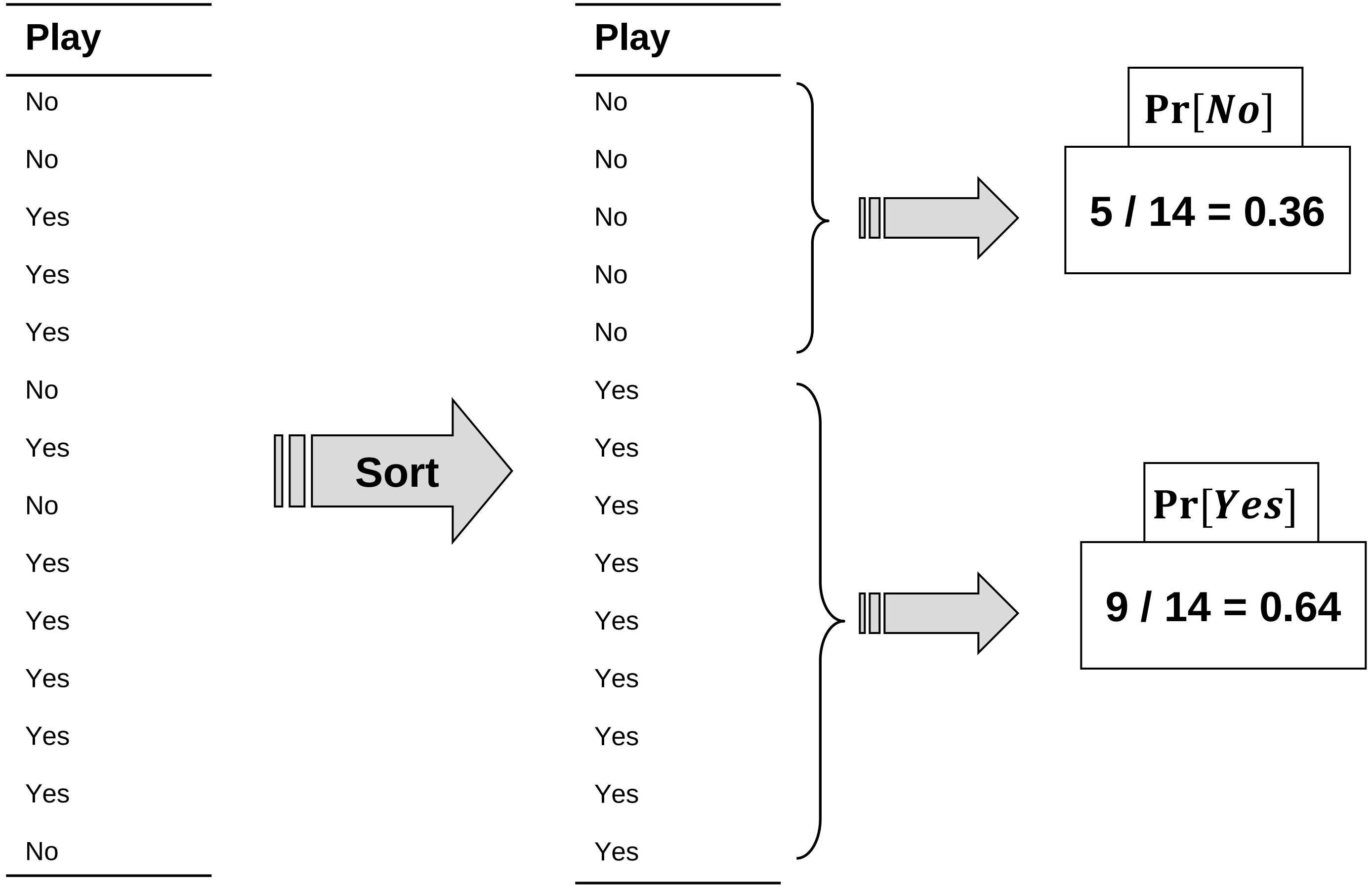
Evidence, e.g., weather=windy

Conditional probability of *h* given *e*

Dataset

Outlook	**Temp**	**Humidity**	**Windy**	**Play**
Sunny	Hot	High	False	**No**
Sunny	Hot	High	True	**No**
Overcast	Hot	High	False	**Yes**
Rainy	Mild	High	False	**Yes**
Rainy	Cool	Normal	False	**Yes**
Rainy	Cool	Normal	True	**No**
Overcast	Cool	Normal	True	**Yes**
Sunny	Mild	High	False	**No**
Sunny	Cool	Normal	False	**Yes**
Rainy	Mild	Normal	False	**Yes**
Sunny	Mild	Normal	True	**Yes**
Overcast	Mild	High	True	**Yes**
Overcast	Hot	Normal	False	**Yes**
Rainy	Mild	High	True	**No**

Frequency Tables



Play
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
No

Sort

Play
No
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes

$\Pr[No]$

5 / 14 = 0.36

$\Pr[Yes]$

9 / 14 = 0.64

Frequency Tables

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	**No**
Sunny	Hot	High	True	**No**
Overcast	Hot	High	False	**Yes**
Rainy	Mild	High	False	**Yes**
Rainy	Cool	Normal	False	**Yes**
Rainy	Cool	Normal	True	**No**
Overcast	Cool	Normal	True	**Yes**
Sunny	Mild	High	False	**No**
Sunny	Cool	Normal	False	**Yes**
Rainy	Mild	Normal	False	**Yes**
Sunny	Mild	Normal	True	**Yes**
Overcast	Mild	High	True	**Yes**
Overcast	Hot	Normal	False	**Yes**
Rainy	Mild	High	True	**No**

Play

Outlook	No	Yes
Sunny	3	2
Overcast	0	4
Rainy	2	3

Temp	No	Yes
Hot	2	2
Mild	2	4
Cool	1	3

Humidity	No	Yes
High	4	3
Normal	1	6

Windy	No	Yes
False	2	6
True	3	3

Naive Bayes – Probabilities

Frequency Tables

Outlook	No	Yes
Sunny	3	2
Overcast	0	4
Rainy	2	3

Temp.	No	Yes
Hot	2	2
Mild	2	4
Cool	1	3

Humidity	No	Yes
High	4	3
Normal	1	6

Windy	No	Yes
False	2	6
True	3	3

Outlook	No	Yes
Sunny	3/5	2/9
Overcast	0/5	4/9
Rainy	2/5	3/9

Temp.	No	Yes
Hot	2/5	2/9
Mild	2/5	4/9
Cool	1/5	3/9

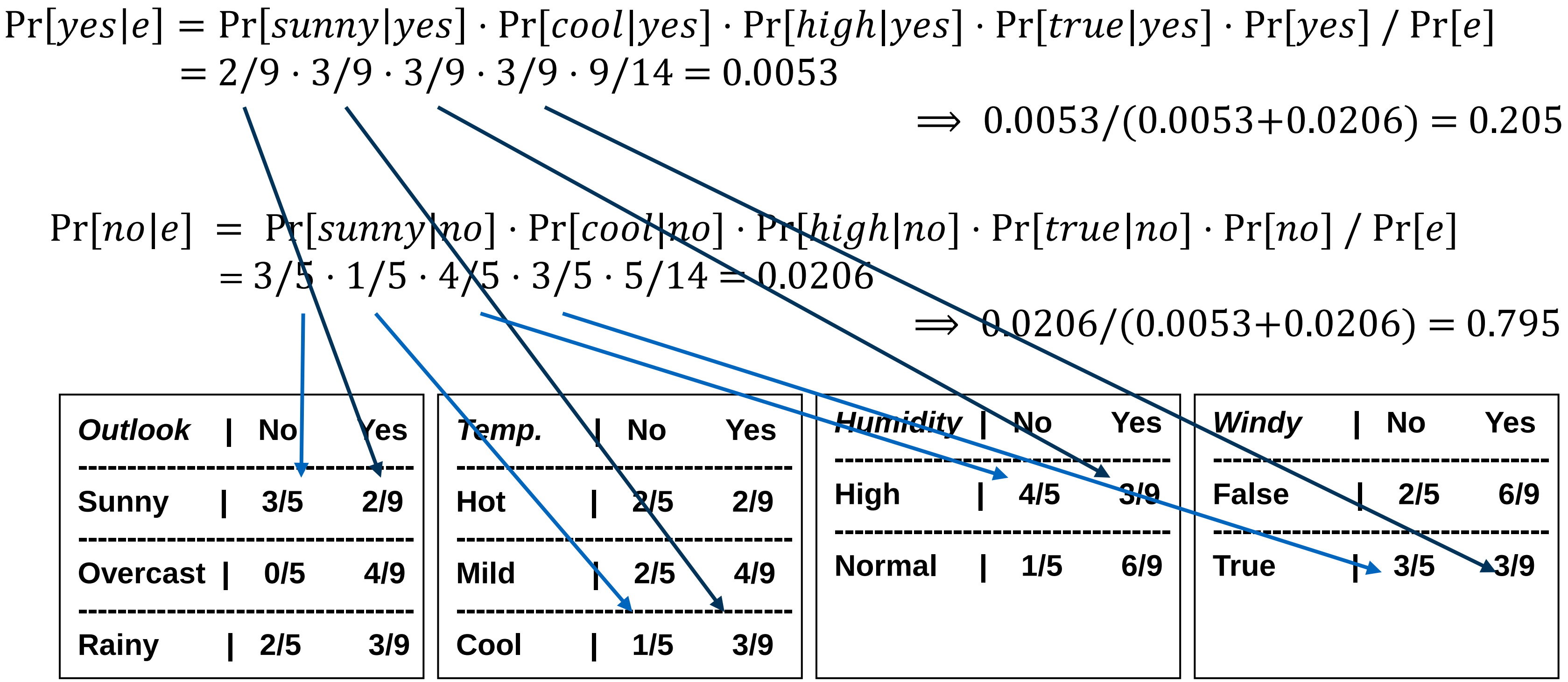
Humidity	No	Yes
High	4/5	3/9
Normal	1/5	6/9

Windy	No	Yes
False	2/5	6/9
True	3/5	3/9

Likelihood Tables

Predicting a New Day

Outlook	Temp.	Humidity	Windy	Play
sunny	cool	high	true	?



$$\Pr[yes|e] = \Pr[sunny|yes] \cdot \Pr[cool|yes] \cdot \Pr[high|yes] \cdot \Pr[true|yes] \cdot \Pr[yes] \;/\; \Pr[e]$$
$$= 2/9 \cdot 3/9 \cdot 3/9 \cdot 3/9 \cdot 9/14 = 0.0053$$
$$\Longrightarrow \; 0.0053/(0.0053{+}0.0206) = 0.205$$

$$\Pr[no|e] \;=\; \Pr[sunny|no] \cdot \Pr[cool|no] \cdot \Pr[high|no] \cdot \Pr[true|no] \cdot \Pr[no] \;/\; \Pr[e]$$
$$= 3/5 \cdot 1/5 \cdot 4/5 \cdot 3/5 \cdot 5/14 = 0.0206$$
$$\Longrightarrow \; 0.0206/(0.0053{+}0.0206) = 0.795$$

Outlook	**No**	**Yes**
Sunny	**3/5**	**2/9**
Overcast	**0/5**	**4/9**
Rainy	**2/5**	**3/9**

Temp.	**No**	**Yes**
Hot	**2/5**	**2/9**
Mild	**2/5**	**4/9**
Cool	**1/5**	**3/9**

Humidity	**No**	**Yes**
High	**4/5**	**3/9**
Normal	**1/5**	**6/9**

Windy	**No**	**Yes**
False	**2/5**	**6/9**
True	**3/5**	**3/9**

Note: The probability Pr[sunny, cool, high, true] is unknown. We use the law of total probability.

Predicting a New Day - Formulas

- Again, given a new instance with
  - outlook=sunny
  - temperature=cool
  - humidity=high
  - windy=true
- If all explanatory attributes are independent and equally important they can be multiplied

$$\Pr[Play = yes] \cdot \prod_i \Pr[e_i | Play = yes]$$

$$= \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} = 0.0053$$

Formulas Continued …

- Similarly

$$\Pr[Play = no] \cdot \prod_i \Pr[e_i | Play = no]$$

$$= \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} = 0.0206$$

- Thus

$$h_{MAP} = \underset{h \in \{Play=yes,\, Play=no\}}{arg \max} \Pr[h] \cdot \prod_i \Pr[e_i | h]$$

$$= \{Play = no\}$$

Normalization

Note that we can normalize to get the *probabilities*:

$$\Pr[h|e_1, e_2, \dots, e_n] = \frac{\Pr[e_1, e_2, \dots, e_n|h] \cdot \Pr[h]}{\Pr[e_1, e_2, \dots, e_n]}$$

$$\frac{0.0053}{0.0053 + 0.0206} = 0.205 \quad h = \{Play = yes\}$$

$$\frac{0.0206}{0.0053 + 0.0206} = 0.795 \quad h = \{Play = no\}$$

$$\Pr[h \mid e] = \frac{\Pr[e_1 \mid h] \Pr[e_2 \mid h] \dots \Pr[e_n \mid h] \Pr[h]}{\Pr[e]}$$

Naive Bayes - Summary

- Want to classify a new instance $(e_1, e_2, \ldots, e_n)$ into finite number of categories from the set h.
  - Choose the most likely classification using Bayes theorem
  - MAP (maximum a posteriori classification)
- Assign the most probable category h_{MAP} given $(e_1, e_2, \ldots, e_n)$, i.e. the maximum likelihood

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} \Pr[h | e_1, e_2, \ldots, e_n] \\ &= \arg\max_{h \in H} \frac{\Pr[e_1, e_2, \ldots, e_n | h] \cdot \Pr[h]}{\Pr[e_1, e_2, \ldots, e_n]} \\ &= \arg\max_{h \in H} \Pr[e_1, e_2, \ldots, e_n | h] \cdot \Pr[h] \end{aligned}$$

- “Naive Bayes” since the attributes are treated as independent: Only then you can multiply the probabilities

$$\Pr[e_1, e_2, \ldots, e_n | h] = \Pr[e_1 | h] \cdot \Pr[e_2 | h] \cdots \Pr[e_n | h]$$

The Weather Data (yet again)

Outlook			Temperature			Humidity			Windy			Play	
	Yes	***No***		***Yes***	***No***		***Yes***	***No***		***Yes***	***No***	***Yes***	***No***
Sunny	2	3	Hot	2	2	High	3	4	FALSE	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	TRUE	3	3		
Rainy	3	2	Cool	3	1								

$$\Pr[Play = no] = {}^{5}\!/_{14}$$

$$\Pr[Outlook = overcast \mid Play = no] = {}^{0}\!/_{5}$$

$$\Pr[Temperature = cool \mid Play = no] = {}^{1}\!/_{5}$$

$$\Pr[Humidity = high \mid Play = no] = {}^{4}\!/_{5}$$

$$\Pr[Windy = true \mid Play = no] = {}^{3}\!/_{5}$$

The Zero Frequency Problem

What if an attribute value doesn't occur with every class value

$$\Pr[Outlook = overcast \mid no] = 0$$

Remedy: add 1 to the numerator for every attribute value-class combination, and the probability can never be zero

$$\Pr[no|e] = 5/14 \cdot 1/8 \cdot 2/8 \cdot 5/7 \cdot 4/7 \qquad = 0.00456 \qquad \Longrightarrow 27.84\%$$

$$\Pr[yes|e] = 9/14 \cdot 5/12 \cdot 4/12 \cdot 4/11 \cdot 4/11 = 0.01181 \qquad \Longrightarrow 72.16\%$$

Outlook	**No**	**Yes**
Sunny	**3+1**	**2+1**
Overcast	**0+1**	**4+1**
Rainy	**2+1**	**3+1**

Temp.	**No**	**Yes**
Hot	**2+1**	**2+1**
Mild	**2+1**	**4+1**
Cool	**1+1**	**3+1**

Humidity	**No**	**Yes**
High	**4+1**	**3+1**
Normal	**1+1**	**6+1**

Windy	**No**	**Yes**
False	**2+1**	**6+1**
True	**3+1**	**3+1**

Modified Probability Estimates

- In some cases adding a constant different from 1 might be more appropriate
- Example: attribute outlook for class no

$$\frac{3+\mu/3}{5+\mu} \qquad \frac{0+\mu/3}{5+\mu} \qquad \frac{2+\mu/3}{5+\mu}$$

Sunny *Overcast* *Rainy*

- Weights (w_p) don't need to be equal (as long as their sum to 1)

$$\frac{3+\mu w_1}{5+\mu} \qquad \frac{0+\mu w_2}{5+\mu} \qquad \frac{2+\mu w_3}{5+\mu}$$

Which assumptions does the Naive Bayes classifier require?

Missing Values

- Training: instance is not included in frequency count for attribute value-class combination
- Classification: attribute will be omitted from calculation
- Example:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood of “yes” = $3/9 \cdot\ 3/9 \cdot 3/9 \cdot 9/14\ =\ 0.0238$

Likelihood of “no” = $\ 1/5 \cdot 4/5 \cdot 3/5 \cdot 5/14\ =\ 0.0343$

$\Pr[\text{Play} = \text{"yes"} \mid e_2, e_3, e_4] = 0.0238\ /\ (0.0238 + 0.0343) = 41\%$

$\Pr[\text{Play} = \text{"no"} \mid e_2, e_3, e_4] = 0.0343\ /\ (0.0238 + 0.0343) = 59\%$

Dealing with Numeric Attributes

- Usual assumption: attributes have a normal or Gaussian probability distribution (given the class)
- The probability density function for the normal distribution is defined by two parameters:
  - The sample mean μ: $$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$
  - The standard deviation σ: $$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$
- The density function $f(x)$: $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Statistics for the Weather Data

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3		83	85		86	85	False	6	2	9	5
Overcast	4	0		70	80		96	90	True	3	3		
Rainy	3	2		68	65		80	70					
				…	…		…	…					
Sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	*std dev*	6.2	7.9	*std dev*	10.2	9.7	True	3/9	3/5		
Rainy	3/9	2/5											

- Calculate mean and std for Temperature when Play = yes; mean = 73, std = 6.2
- Example density value for Temperature = 66:

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Classifying a New Day

- A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood of “yes” $= \frac{2}{9} \cdot\ 0.0340 \cdot 0.0221 \cdot \frac{3}{9} \cdot \frac{9}{14} = 0.000036$

Likelihood of “no” $= \frac{3}{5} \cdot 0.0291 \cdot 0.0380 \cdot \frac{3}{5} \cdot \frac{5}{14} = 0.000136$

$\Pr[\text{Play} = \text{"yes"} | e_1, .., e_4] = 0.000036\ /\ (0.000036 + 0.\ 000136) = 20.9\%$

$\Pr[\text{Play} = \text{"no"} | e_1, .., e_4] = 0.\ 000136\ /\ (0.000036 + 0.\ 000136) = 79.1\%$

- Missing values during training: not included in calculation of mean and standard deviation

Numeric Data: Unknown Distribution

- What if the data distribution does not follow a known distribution
- In this case we need a mechanism to estimate the density distribution
- A simple and intuitive approach is based on kernel density estimation

- Consider a random variable X whose distribution $f(X)$ is unknown but a sample with a non-uniform distribution

$$\{x_1, x_2, \dots, x_n\}$$

Kernel Density Estimation

We want to derive a function $f(x)$ such that

(1) $f(x)$ is a probability density function, i.e.

$$\int f(x)dx = 1$$

(2) $f(x)$ is a smooth approximation of the data points in X

(3) $f(x)$ can be used to estimate values x^* which are not in

$$\{x_1, x_2, \dots, x_n\}$$

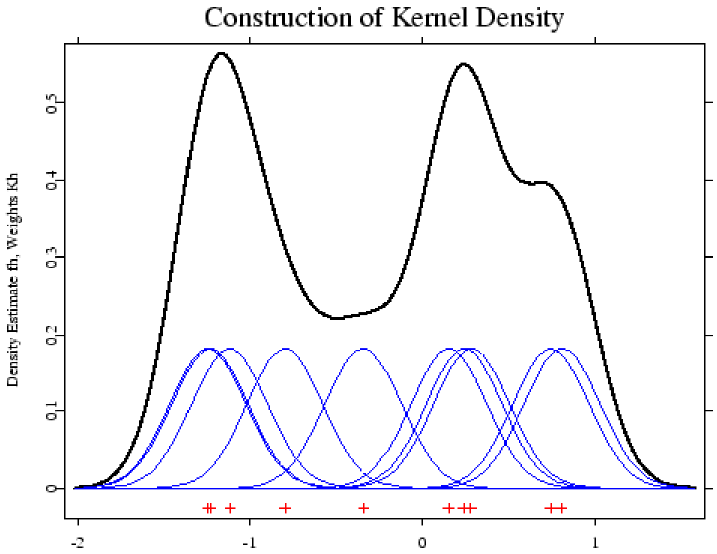
Kernel Density Estimate

Rosenblatt-Parzen Kernel-Density-Estimator:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i, h)$$

where

$$K(t,h) = \frac{1}{\sqrt{2\pi h}} e^{-\frac{1}{2}(\frac{t}{h})^2}$$



Adjust “h” (aka bandwidth) to fit data as a parameter

Discussion of Naïve Bayes

- Naïve Bayes works surprisingly well (even if independence assumption is clearly violated)
  - Domingos, Pazzani: On the Optimality of the Simple Bayesian Classifier under Zero-One-Loss, Machine Learning (1997) 29.
- However: adding too many redundant attributes will cause problems (e.g., identical attributes)
  - Note also: many numeric attributes are not normally distributed
- Time complexity
  - Calculating conditional probabilities: Time $O(n)$ where n is the number of instances
  - Calculating the class: Time $O(cp)$ where c is the number of classes, p the attributes

What would we do, if the independence assumption was violted?

Bayesian (Belief) Networks: Multiple Variables with Dependency

- Naïve Bayes assumption of conditional independence is often too restrictive
- Bayesian Belief network (Bayesian net) describe conditional independence among subsets of attributes: combining prior knowledge about dependencies among variables with observed training data
- Graphical representation: directed acyclic graph (DAG), one node for each attribute
  - Overall probability distribution factorized into component distributions
  - Graph’s nodes hold component distributions (conditional distributions)

Probability Laws

- Chain rule
  - $\Pr[e_1, e_2, \ldots, e_n] = \prod_{i=1,\ldots,n} \Pr[e_{\ i} | e_{\ i\ -\ 1}, \ldots, e_1]$
  - $\Pr[A, B, C, D, E] = \Pr[A]\Pr[B|A]\Pr[C|\mathrm{A}, \mathrm{B}]\Pr[D|A, B, C]\Pr[E|A,B,C,D]$
  - The joint distribution is independent of the ordering

- Conditional independence
  - $\Pr[h|e_1, e_2\ ] = \Pr[h|e_2]$
  - Example:
    - Rain causes people to use an umbrella and traffic to slow down
    - Umbrella is conditionally independent of traffic given rain
      - $Umbrella \perp\!\!\!\perp Traffic \mid Rain$
      - $\Pr[Umbrella, Traffic|Rain] = \Pr[Umbrella|Rain] * \Pr[Traffic|Rain]$
      - $\Pr[Umbrella|Rain, Traffic] = \Pr[Umbrella|Rain]$

The Full Joint Distribution

$$
\begin{aligned}
&\Pr[e_1,\ldots,e_n]\\
&=\Pr[e_n \mid e_{n-1},\ldots,e_1]\Pr[e_{n-1},\ldots,e_1]\\
&=\Pr[e_n \mid e_{n-1},\ldots,e_1]\Pr[e_{n-1} \mid e_{n-2},\ldots,e_1]\Pr[e_{n-2},\ldots,e_1]\\
&=\Pr[e_n \mid e_{n-1},\ldots,e_1]\Pr[e_{n-1} \mid e_{n-2},\ldots,e_1]\ldots\Pr[e_2 \mid e_1]P[e_1]\\
&=\prod_{i=1}^{n}\Pr[e_i \mid e_{i-1},\ldots,e_1]\\
&=\prod_{i=1}^{n}\Pr[e_i \mid parents(e_i)]
\end{aligned}
$$

(Chain Rule)

From the chain rule
to Bayesian networks

Bayesian Network Assumptions

- $\Pr[e_1, e_2, \ldots, e_n] = \Pr[\mathrm{E}_1 = e_1 \wedge \ldots \wedge \mathrm{E}_n = e_n]$

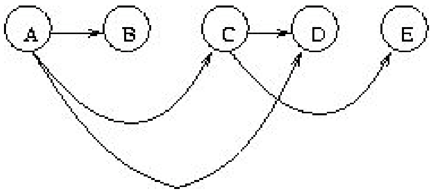
$$= \prod_{i=1,\ldots,n} \Pr[e_i | e_{i-1}, \ldots, e_1] = \prod_{i=1,\ldots,n} \Pr[e_i | Parents(e_i)]$$

- Pr[Traffic, Rain, Umbrella]= $\Pr[T, R, U]$

  $= \Pr[R] \cdot \Pr[T|R] \cdot \Pr[U|R, T]$

  – Considering conditional independence:

  Pr[T,R,U]= Pr[R] · Pr[T|R] · Pr[U|R]



- $\Pr[A, B, C, D, E] = \Pr[A]\Pr[B|A]\Pr[C|A, B]\Pr[D|A, B, C]\Pr[E|A, B, C, D]$

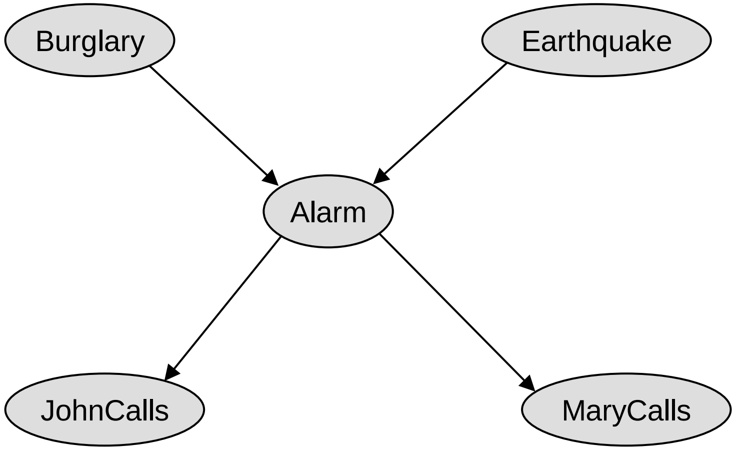
  $= \Pr[A]\Pr[B|A]\Pr[C|A]\Pr[D|A, C]\Pr[E|C]$

Example: Alarm

Your house has an alarm system against burglary. The house is located in the seismically active area and the alarm system can get occasionally set off by an earthquake. You have two neighbors, Mary and John, who do not know each other. If they hear the alarm they call you, but this is not guaranteed. They also call you from time to time just to chat.

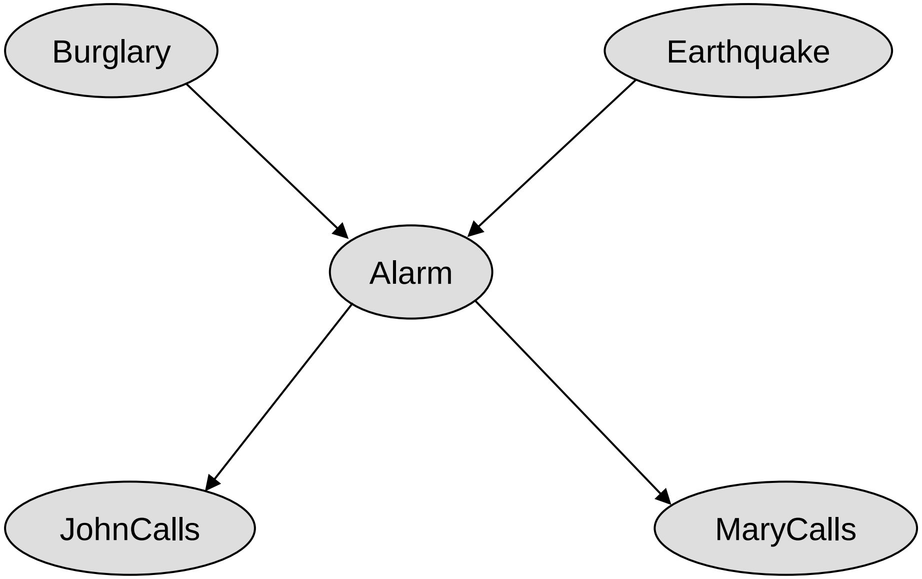
Five random variables

- A: Alarm
- B: Burglary
- E: Earthquake
- J: JohnCalls
- M: MaryCalls



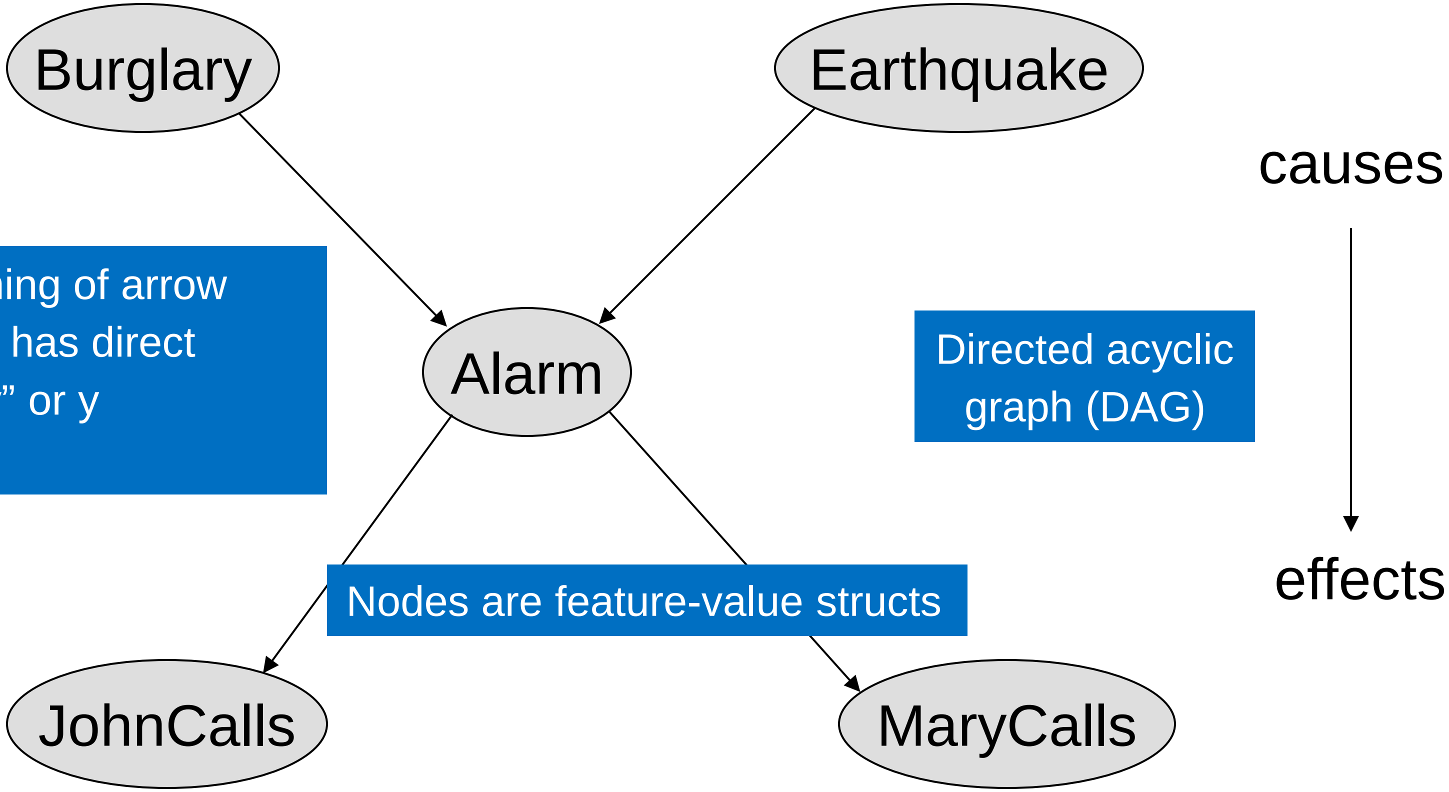
[illustration by Kevin Murphy]

Example



$$\Pr(JohnCalls, MaryCalls, Alarm, Burglary, Earthquake)$$
$$= \Pr(JohnCalls \mid Alarm) \Pr(MaryCalls \mid Alarm)$$
$$\Pr(Alarm \mid Burglary, Earthquake) \Pr(Burglary) \Pr(Earthquake)$$

A Simple Bayes Net



Burglary

Earthquake

Alarm

JohnCalls

MaryCalls

causes

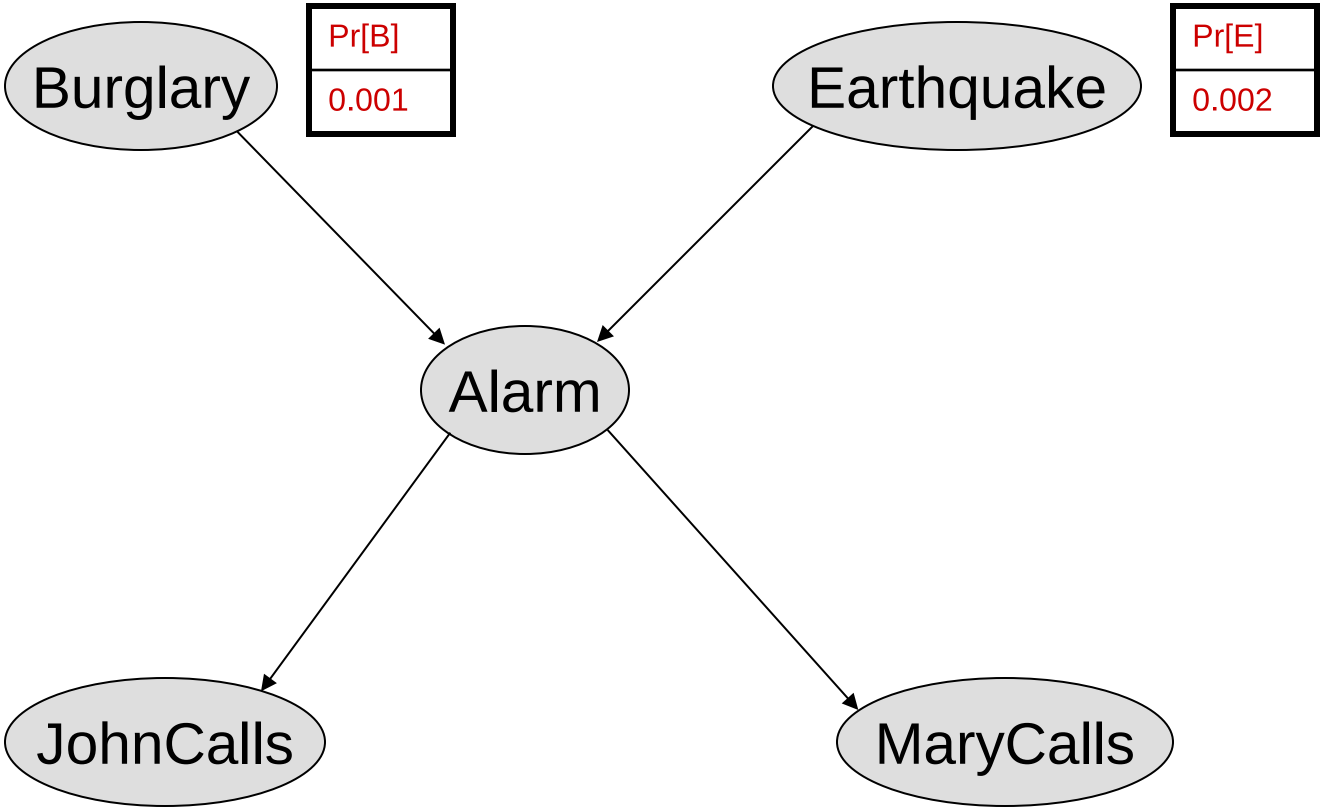
effects

Intuitive meaning of arrow from x to y: “x has direct influence on y” or y depends on x

Directed acyclic graph (DAG)

Nodes are feature-value structs

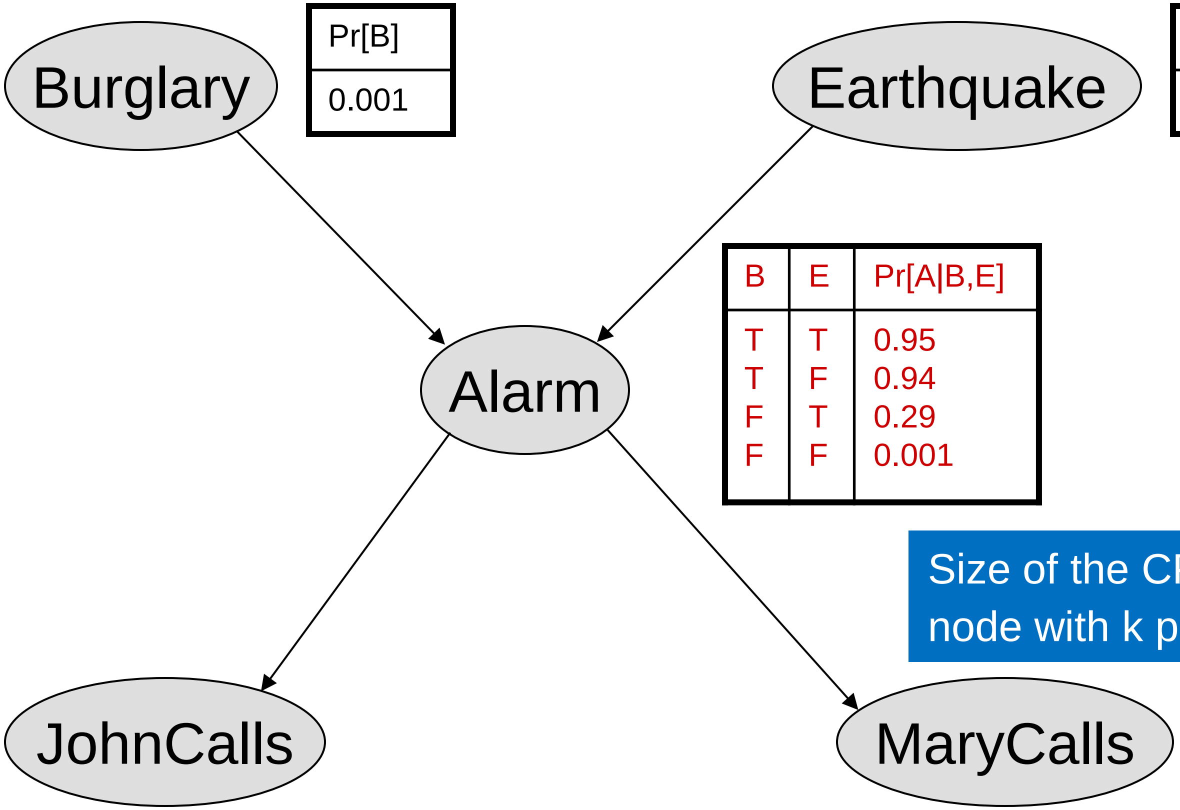
Assigning Probabilities to Roots



Pr[B]
0.001

Pr[E]
0.002

Conditional Probability Tables



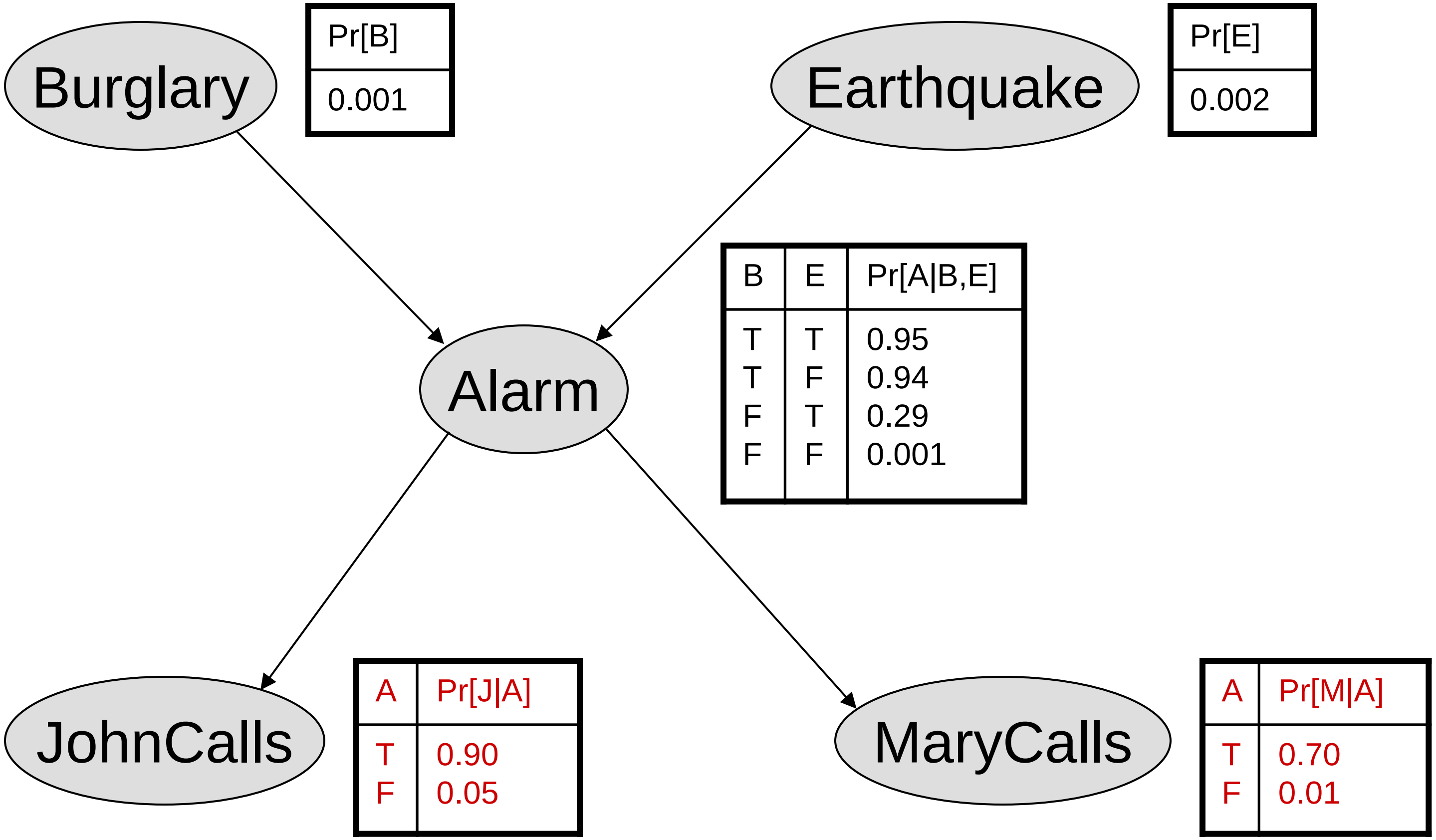
Pr[B]
0.001

Pr[E]
0.002

B	E	Pr[A|B,E]
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

Size of the CPT for a node with k parents: 2^k

Conditional Probability Tables



Burglary

Pr[B]
0.001

Earthquake

Pr[E]
0.002

Alarm

B	E	Pr[A|B,E]
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

JohnCalls

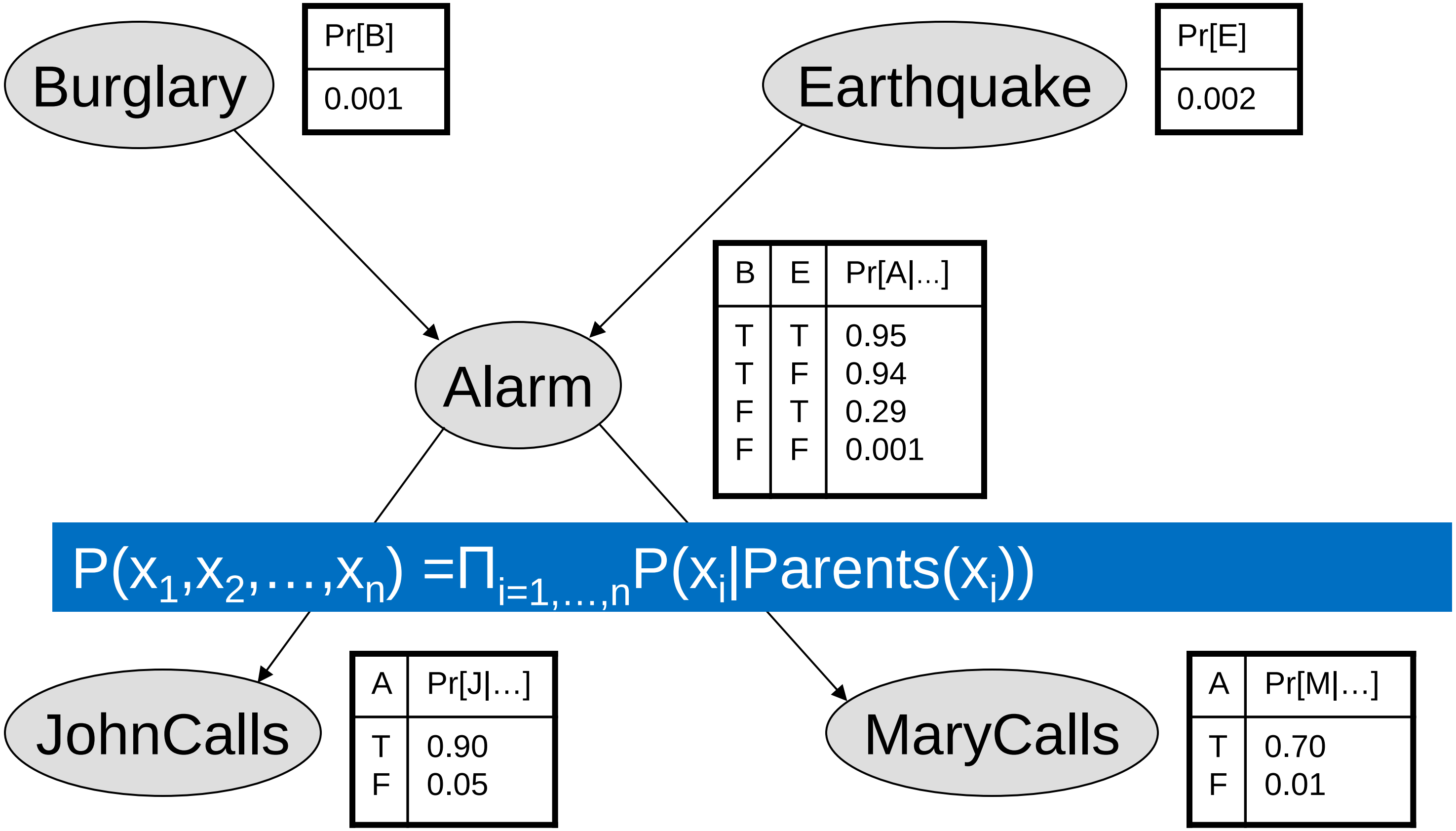
A	Pr[J|A]
T	0.90
F	0.05

MaryCalls

A	Pr[M|A]
T	0.70
F	0.01

Note: $\Pr[J|A] + \Pr[\neg J|A] = 1$, but $\Pr[J|A] + \Pr[J|\neg A] \neq 1$

What the BN Means



Pr[B]
0.001

Pr[E]
0.002

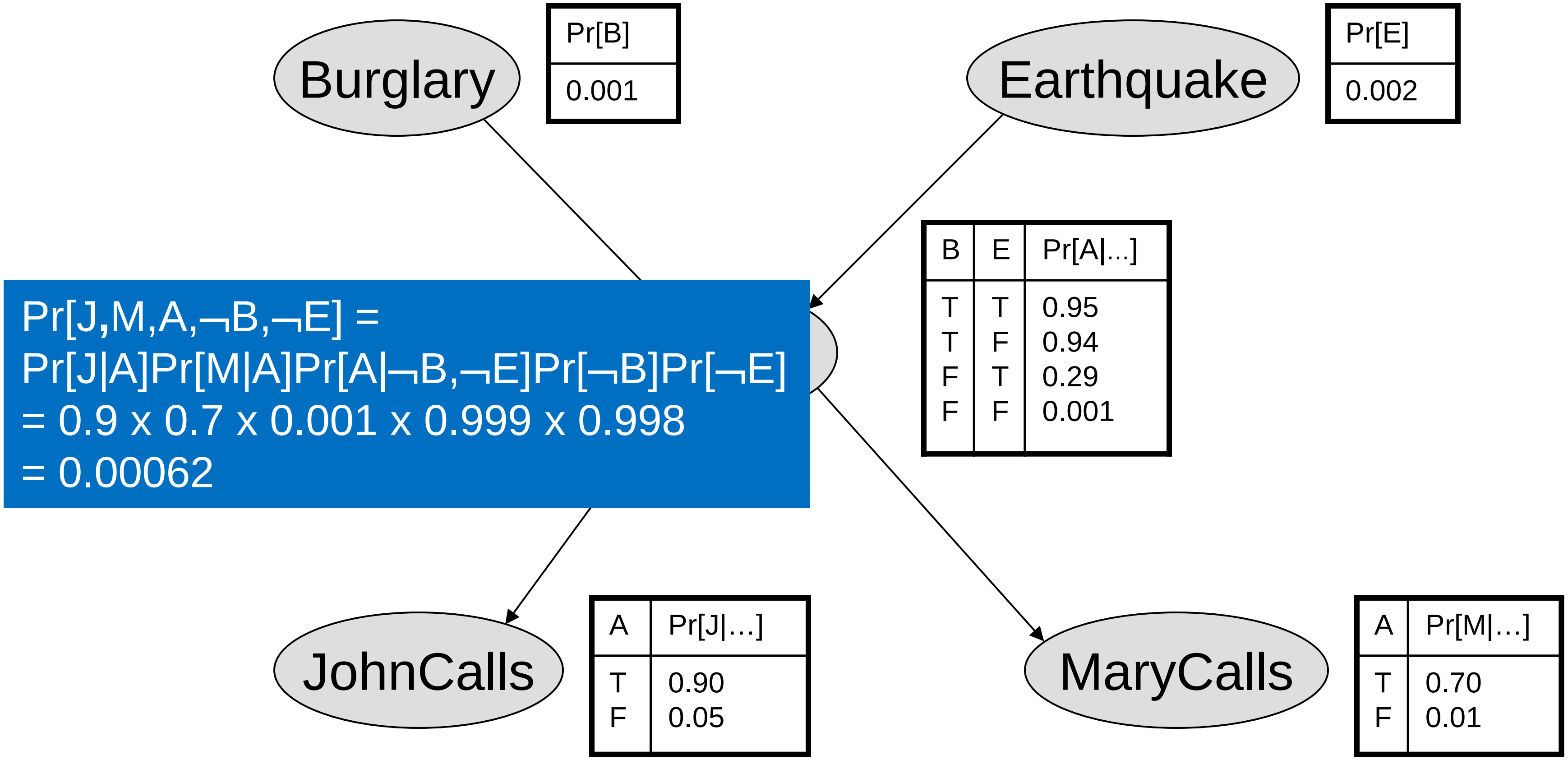
B	E	Pr[A|...]
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

$$P(x_1,x_2,\ldots,x_n) = \Pi_{i=1,\ldots,n} P(x_i|Parents(x_i))$$

A	Pr[J|…]
T	0.90
F	0.05

A	Pr[M|…]
T	0.70
F	0.01

Calculation of Joint Probability



Pr[B]
0.001

Pr[E]
0.002

B	E	Pr[A|…]
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	Pr[J|…]
T	0.90
F	0.05

A	Pr[M|…]
T	0.70
F	0.01

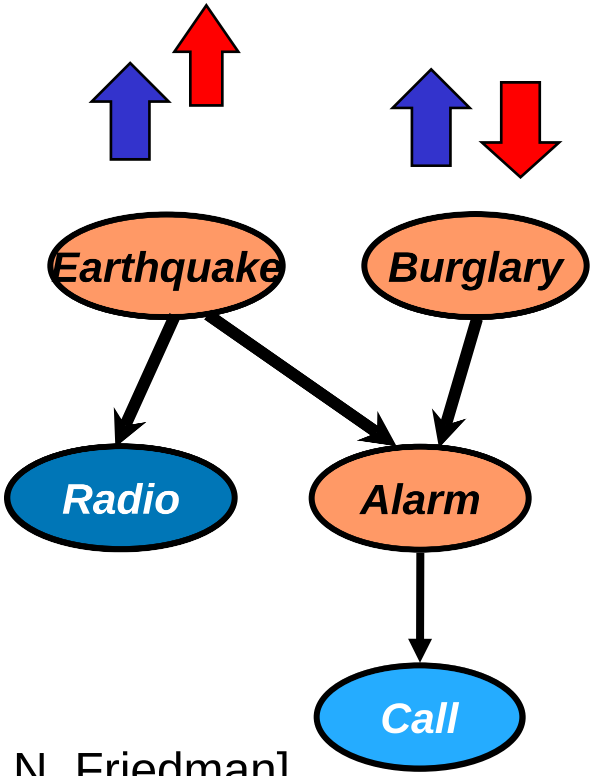
$\Pr[J,M,A,\neg B,\neg E] =$
$\Pr[J|A]\Pr[M|A]\Pr[A|\neg B,\neg E]\Pr[\neg B]\Pr[\neg E]$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

Inference in Bayesian Networks

- Other than the joint probability of specific events, we may want to infer the probability of an event, given observations about a subset of other variables

- For such inference on the Bayesian Network, we need to consider the evidence and the topology of the network

Explaining away effect



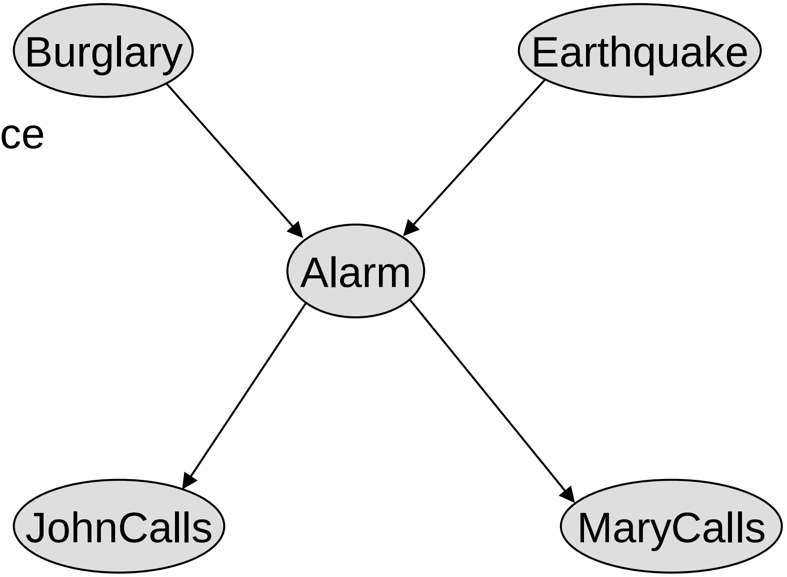
[Figure by N. Friedman]

Inference Rules: An Example

- If alarm is not observed, B and M calls are dependent:
  - My knowing that B has taken place increases my belief on M
  - My knowing that M called increases my belief on B

- If alarm is observed, B and M are conditionally independent:
  - If I already know that the alarm went off,
    - My further knowing that a burglary has taken place would not increase my belief on Mary‘s call
    - My further knowing that Mary called would not increase my belief on burglary

- d-separation determines whether a set of nodes *X* is independent of another set *Y* given a third set *E.* We’ll not discuss this further in this course.



Learning Bayes Nets

- Method for evaluating the goodness of a given network
  - Measures maximize the joint probability of training data given the network (or the sum of the logarithms thereof)
    - Summarize the Log-Likelihood of training data based on the network
  - Example:
    - Akaike information Criterion (AIC): $-2LL + 2K$
    - Minimize AIC, with K=number of parameters

- Method for searching through space of possible networks
  - Amounts to searching through sets of edges because nodes are fixed
  - Examples: K2, Tree Augmented Naive Bayes (TAN)

Bayes Nets Summary

- Bayes Nets can handle dependencies among attributes
- Learning Bayes Nets is computationally complex
  - Network structure is given or not
  - All or only part of the variable values are observable in the training data
- Bayes Networks are the subject of much current research