Business Analytics & Machine Learning

Naïve Bayes and Bayes Networks

Prof. Dr. Martin Bichler & Prof. Dr. Jalal Etesami

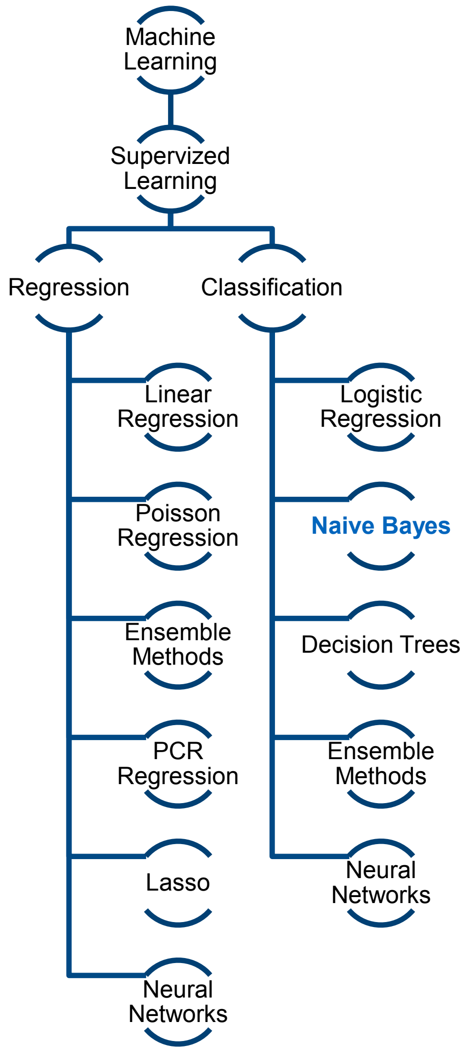
Department of Computer Science

School of Computation, Information, and Technology

Technical University of Munich

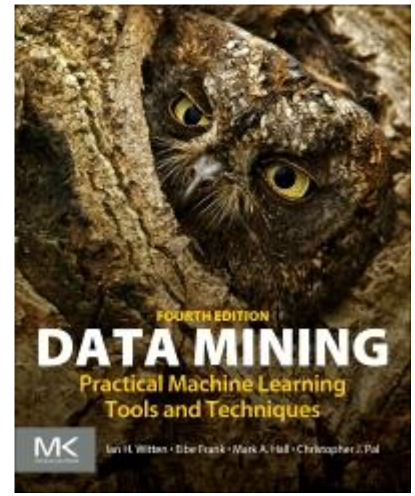
Course Content

- 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
- Dimensionality Reduction
- Association Rules and Recommenders
- Convex Optimization
- Neural Networks
- Reinforcement Learning



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



<u>Alternative literature:</u>

- **Machine Learning**
  – Tom M. Mitchell, 1997

- **Data Mining: Introductory and Advanced Topics**
  – Margaret H. Dunham, 2003

Formal Definition of Classification

Classification:

Given a database $D = \{x_1, x_2, \ldots, x_n\}$ of tuples (items, records) and a set of classes $C = \{C_1, C_2, \ldots, C_m\}$, the classification problem is to define a mapping $f: D \rightarrow C$ where each x_i is assigned to one class. A class, C_j, contains precisely those tuples mapped to it; that is,

$C_j = \{x_i \mid f(x_i) = C_j, 1 \leq i \leq n, \text{and } x_i \in D\}$.

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 are equally important.
- All attributes are conditionally 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?

joint

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

$a \perp\!\!\!\perp b$

Conditional Independence

First, two (sets of) random variables are independent if knowledge about one does not affect knowledge about another. In particular, we have that

- $P(A \cap C) = P(A)P(B)$

They are conditionally independent if they are unrelated after taking account of a 3rd variable.

- $P(A \cap B|C) = P(A|C)P(B|C)$

Conditional independence happens when we have three (sets of) random variables, and conditioning on one makes the other two independent.

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

- $\Pr[e \cap h] = \Pr[e|h]\Pr[h]$
- $\Pr[e \cap h] = \Pr[h|e]\Pr[e]$

posterior prior

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

Bayes‘ Theorem: An Example

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

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

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

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

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

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

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

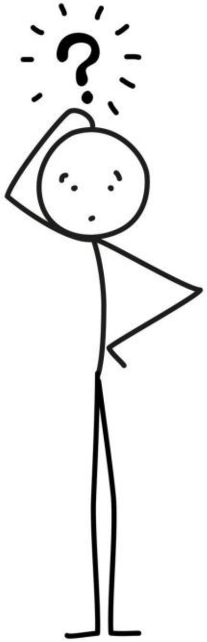
$$p(B) = \sum_{i=0} p(B|A_i)P(A_i)$$

- sum every B conditioned by the As times its prior

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

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.



Bayes Theorem

Posterior Probability

Likelihood

Hypothesis, e.g. play

Prior

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

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

priors : usually initialized by
· counting class intances

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

$\mathbf{Pr}[\boldsymbol{No}]$

5 / 14 = 0.36

$\mathbf{Pr}[\boldsymbol{Yes}]$

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

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

Humidity	I	No	Yes
High	I	4	3
Normal	I	1	6

Windy	I	No	Yes
False	I	2	6
True	I	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

P(Sunny | Yes)

Likelihood Tables

$p(p = yes \mid sunny, cool, high, windy) = \frac{p(sunny, cool, high, windy)\, P(p = yes)}{p(sunny, cool, high, windy)}$

Predicting a New Day

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

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

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

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: Pr[sunny, cool, high, true] (=$\Pr[e]$) is unknown. 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.

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

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

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

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

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

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 $(e_1, e_2, \dots, e_n)$ 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 $(e_1, e_2, \dots, e_n)$, i.e. the maximum likelihood.

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

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

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

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

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

Laplace smoothing

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

$$\Pr[no|e] \; = 5/14 \cdot 1/8 \cdot 2/8 \cdot 5/7 \cdot 4/7 \qquad\quad = 0.00456 \qquad \Longrightarrow 27.84\%$$
$$\Pr[yes|e] = 9/14 \cdot 5/12 \cdot 4/12 \cdot 4/11 \cdot 4/11 = 0.01181 \qquad \Longrightarrow 72.16\%$$

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

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

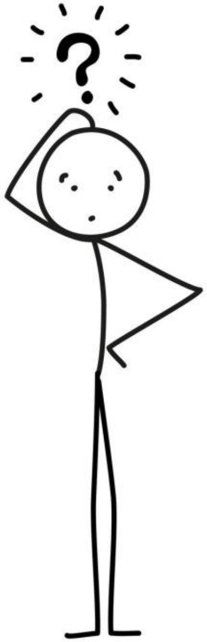
Sunny *Overcast* *Rainy*

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

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

- convex combination $w_1 + w_2 + w_3 = 1$ $0 \leq w_i \leq 1$

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

continuous variable

use normal distrib.

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:

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

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

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

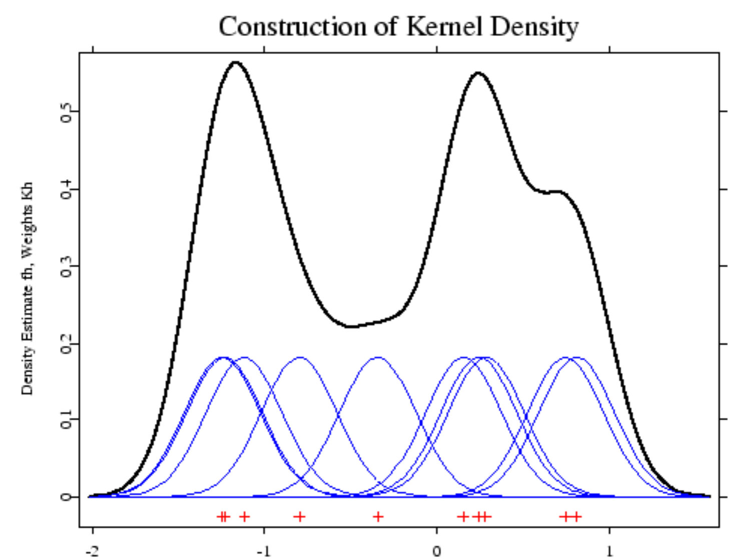
Kernel Density Estimate

Rosenblatt-Parzen Kernel-Density-Estimator:

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

Where

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



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.

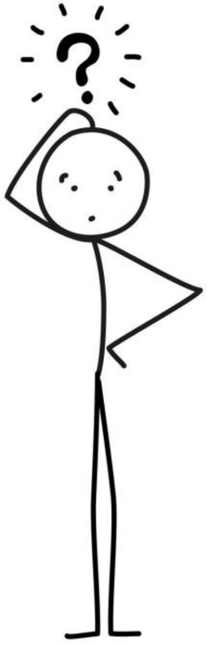
However: adding too many redundant attributes will cause problems (e.g., identical attributes).

- 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 violated?



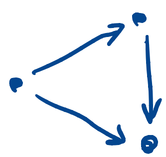
Bayesian (Belief) Networks: Multiple Variables with Dependency

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

- $\Pr[e_1, e_2, \ldots, e_n] = \prod_{i=1,\ldots,n} \Pr[e_i | e_{i-1}, \ldots, e_1]$
- $\Pr[A, B, C, D, E] = \Pr[A]\Pr[B|A]\Pr[C|A, B]\Pr[D|A, B, C]\Pr[E|A, B, C, D]$
- 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

conditional probs

$$
\begin{aligned}
&\Pr[e_1,\ldots,e_n] \\
&= \Pr[e_n \mid e_{n-1},\ldots,e_1]\Pr[e_{n-1},\ldots,e_1] \\
&= \Pr[e_n \mid e_{n-1},\ldots,e_1]\Pr[e_{n-1} \mid e_{n-2},\ldots,e_1]\Pr[e_{n-2},\ldots,e_1] \\
&= \Pr[e_n \mid e_{n-1},\ldots,e_1]\Pr[e_{n-1} \mid e_{n-2},\ldots,e_1]\ldots\Pr[e_2 \mid e_1]P[e_1]
\end{aligned}
$$

(Chain Rule)

$$
= \prod_{i=1}^{n} \Pr[e_i \mid e_{i-1},\ldots,e_1]
$$

From the chain rule

to Bayesian networks

$$
= \prod_{i=1}^{n} \Pr[e_i \mid parents(e_i)]
$$

Bayesian Network Assumptions

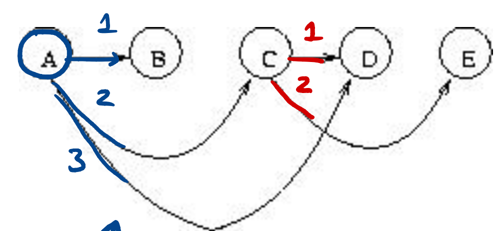
$$\Pr[e_1, e_2, \ldots, e_n] = \Pr[E_1 = e_1 \wedge \ldots \wedge E_n = e_n]$$

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

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]



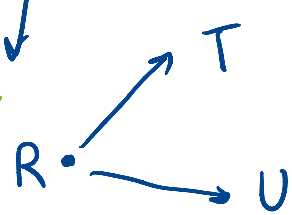
$$\Pr[A, B, C, D, E] = \Pr[A]\Pr[B|A]\Pr[C|A, B]\Pr[D|A, B, C]\Pr[E|A, B, C, D]$$

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

1 2 3 1 2

A is a parent of B

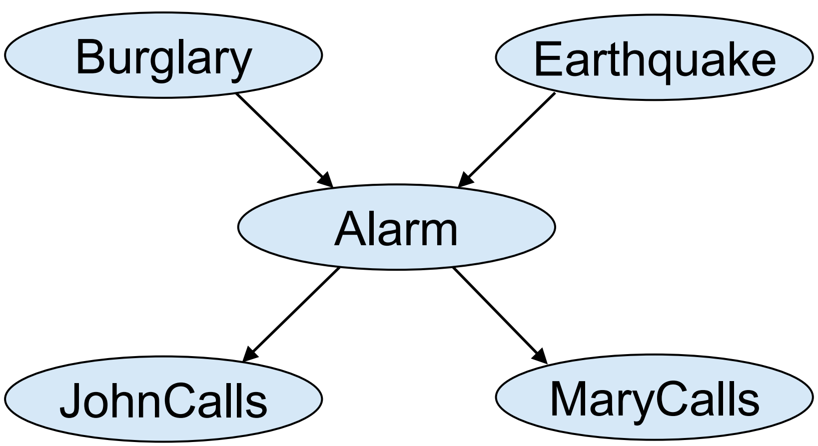
DAG:



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.

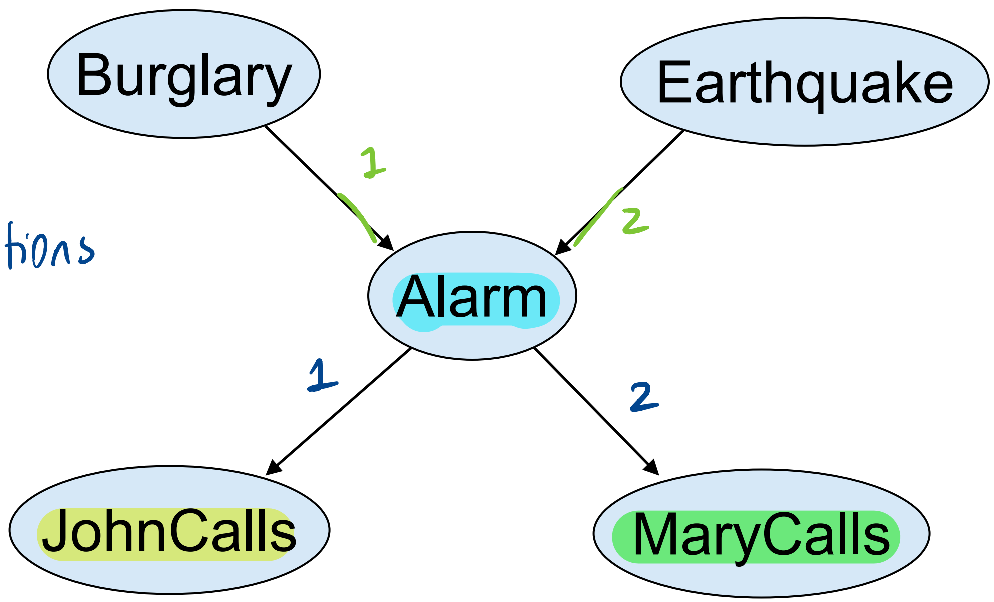
Five random variables
- A: Alarm
- B: Burglary
- E: Earthquake
- J: JohnCalls
- M: MaryCalls



[illustration by Kevin Murphy]

Example

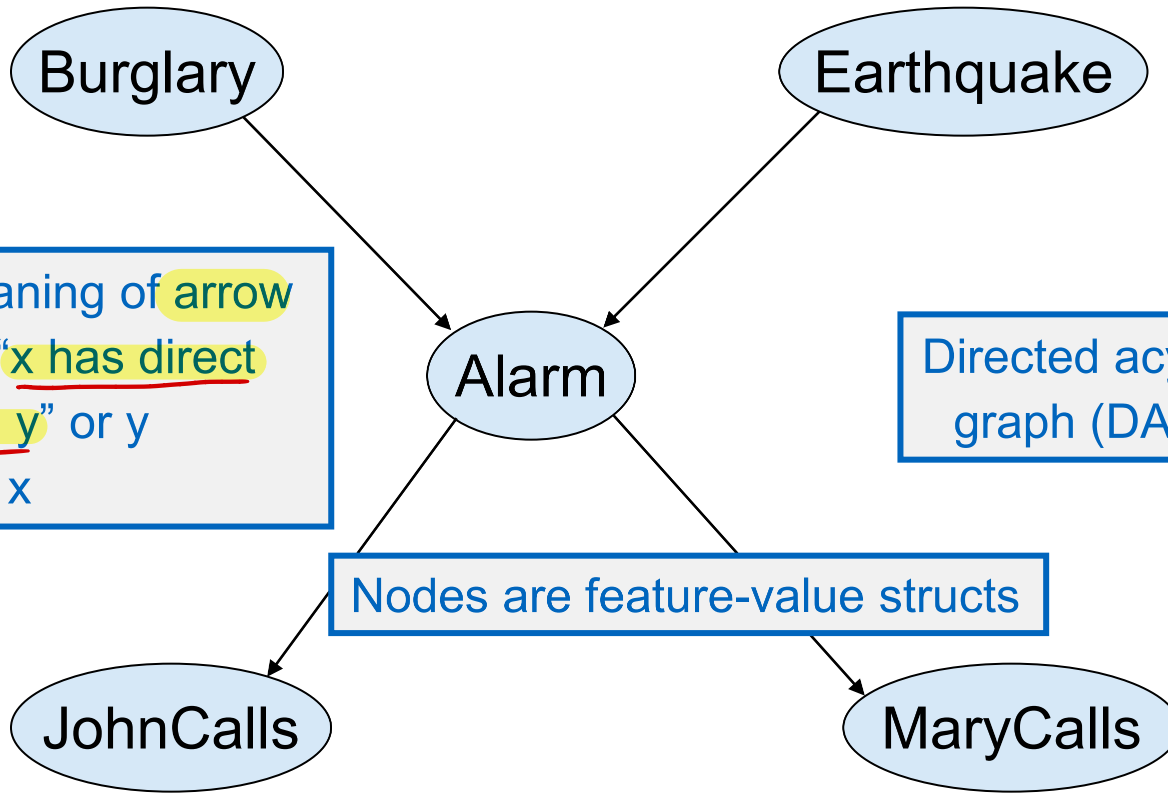
- every node conditions of its parents



$$
\begin{aligned}
&\Pr(JohnCalls, MaryCalls, Alarm, Burglary, Earthquake)\\
&= \Pr(JohnCalls \mid \overset{1}{Alarm}) \Pr(MaryCalls \mid \overset{2}{Alarm}\,)\\
&\Pr(Alarm \mid \underset{1}{Burglary}, \underset{2}{Earthquake}\,) \Pr(\,Burglary\,) \Pr(\,Earthquake\,)
\end{aligned}
$$

no parents

A Simple Bayes Net



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

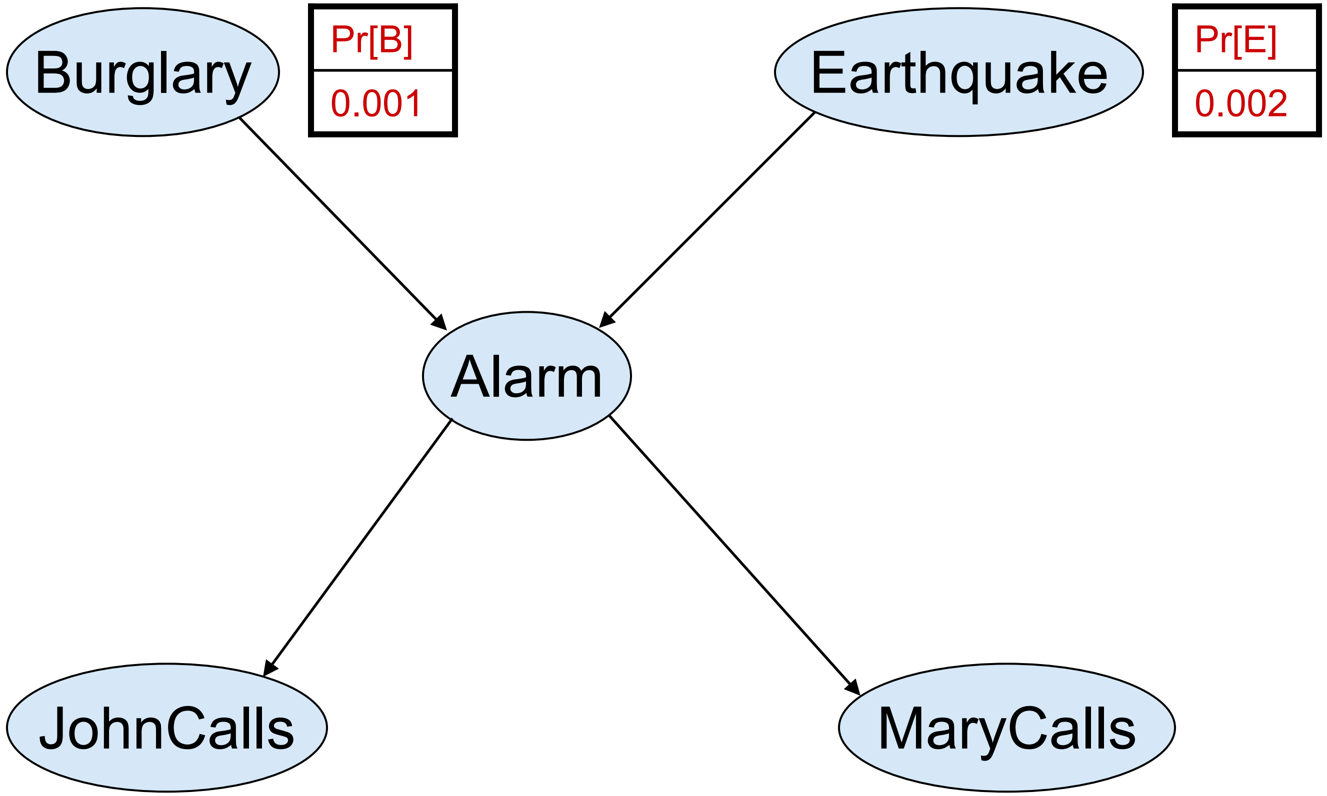
Nodes are feature-value structs

Directed acyclic graph (DAG)

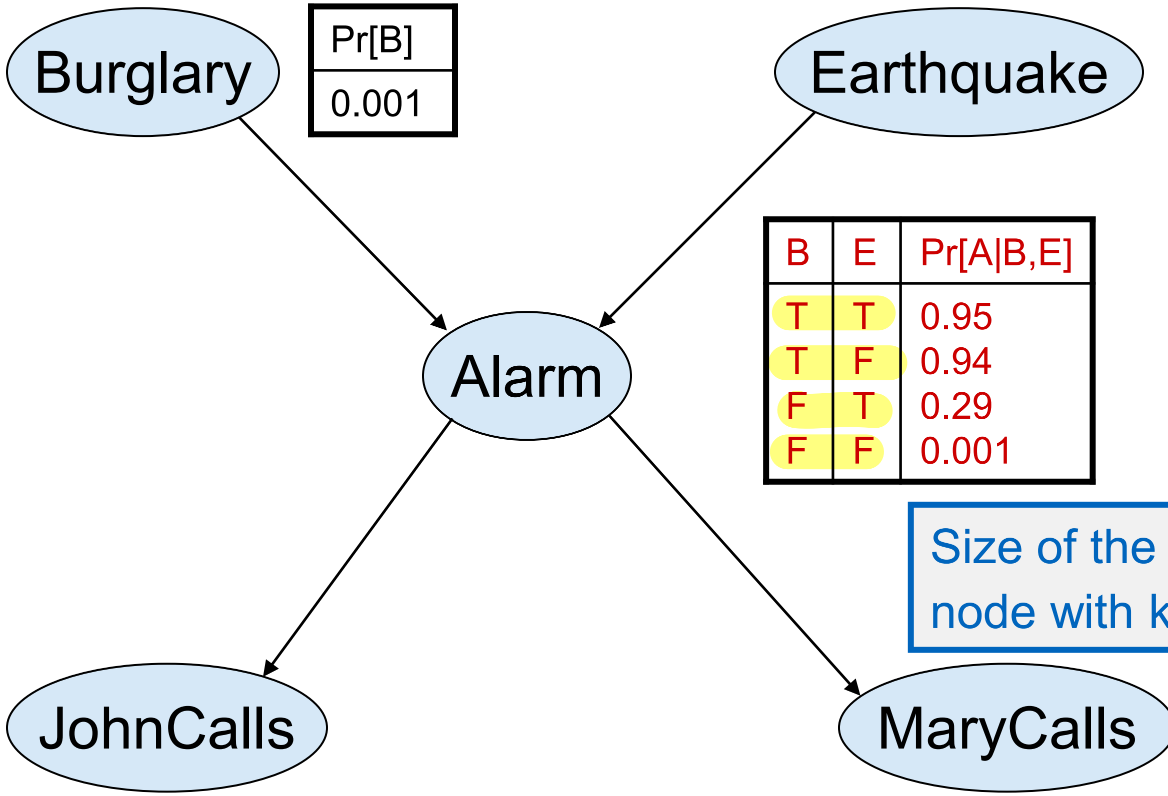
causes

effects

Assigning Probabilities to Roots



Conditional Probability Tables



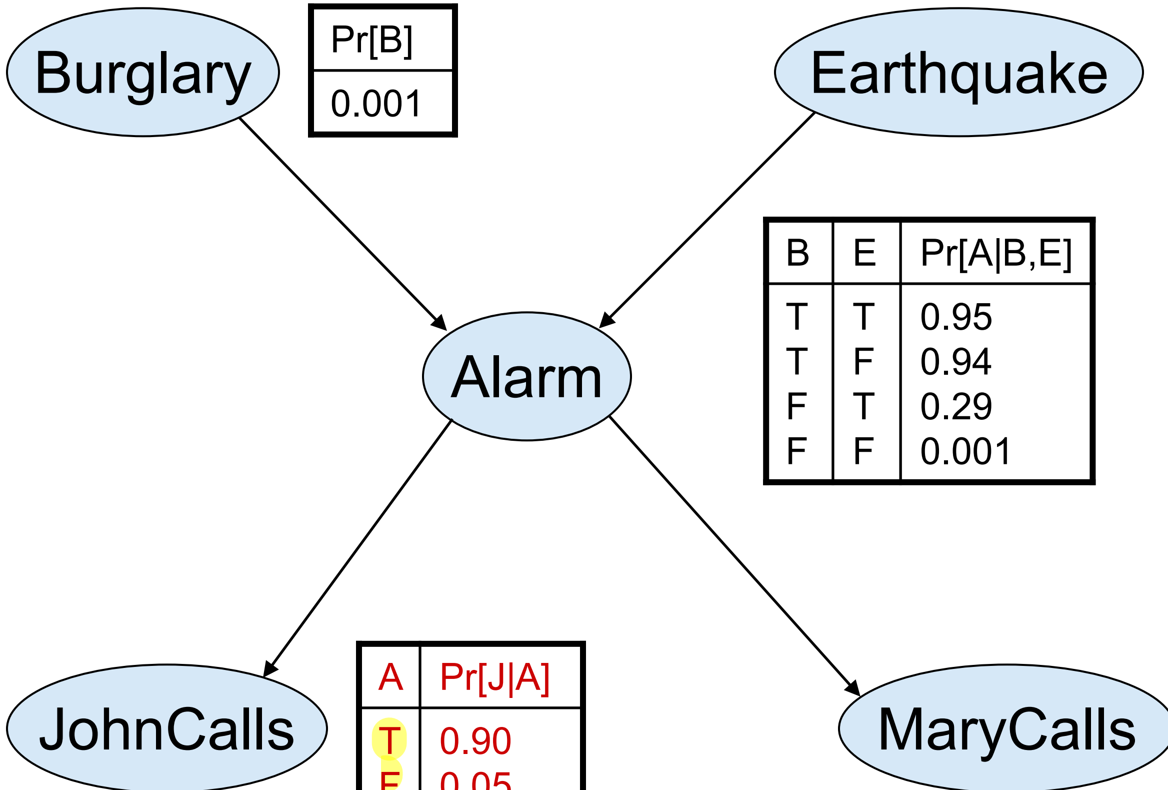
Pr[B]
0.001

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



Pr[B]
0.001

Pr[E]
0.002

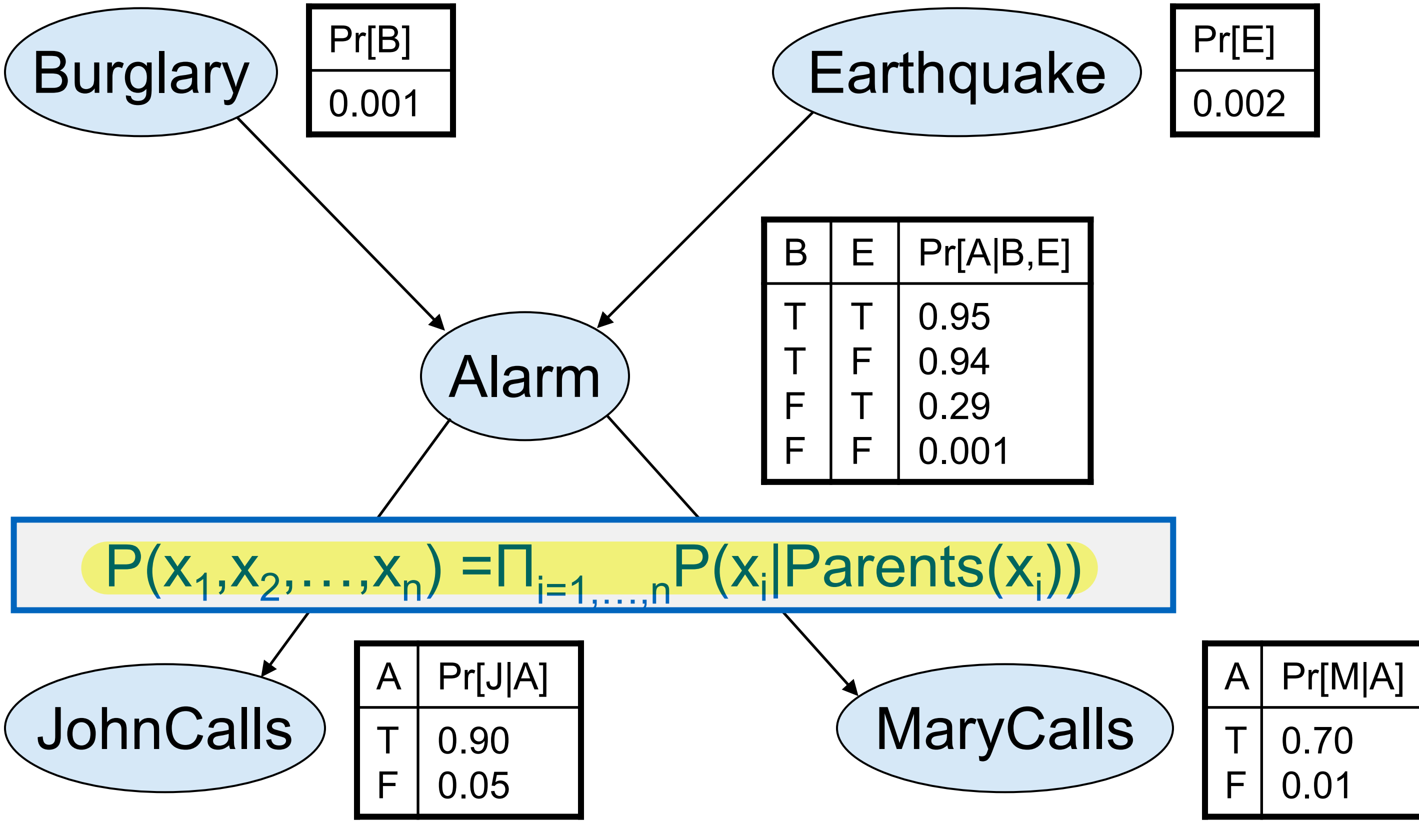
B	E	Pr[A|B,E]
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	Pr[J|A]
T	0.90
F	0.05

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

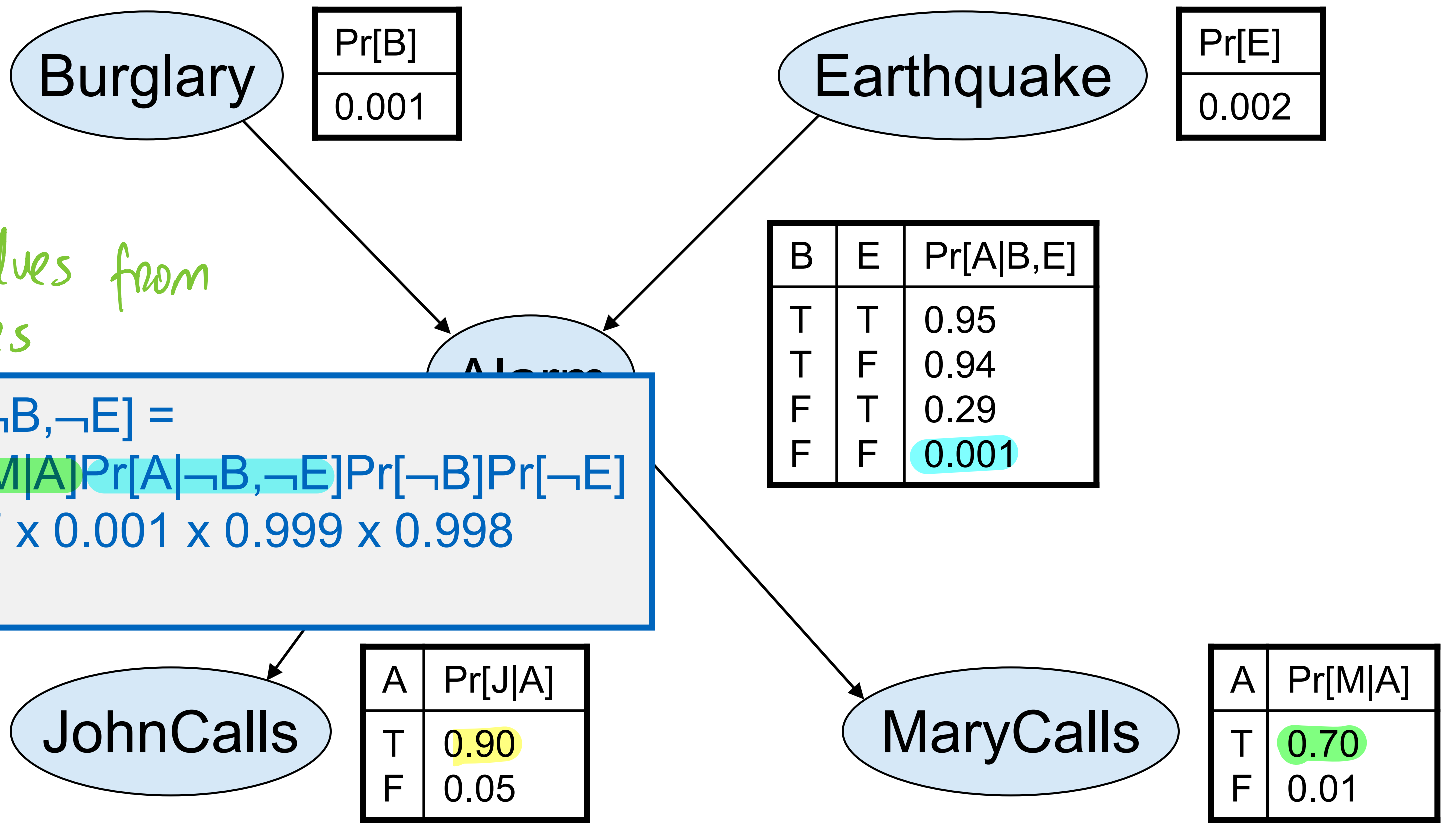
B	E	Pr[A|B,E]
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	Pr[J|A]
T	0.90
F	0.05

A	Pr[M|A]
T	0.70
F	0.01

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

Calculation of Joint Probability



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

A	Pr[J|A]
T	0.90
F	0.05

A	Pr[M|A]
T	0.70
F	0.01

use the values from the tables

Pr[J,M,A,¬B,¬E] =
Pr[J|A]Pr[M|A]Pr[A|¬B,¬E]Pr[¬B]Pr[¬E]
= 0.9 x 0.7 x 0.001 x 0.999 x 0.998
= 0.00062

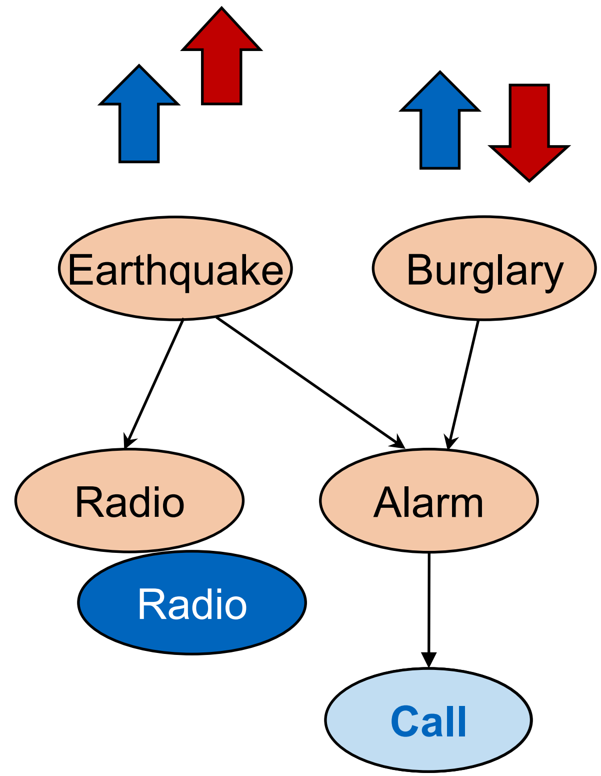
Inference in Bayesian Networks

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.

$$p(e_1, \dots, e_p \mid h) = \prod_i p(e_i \mid h, \text{parents}(e_i))$$

Explaining away effect



[Figure by N. Friedman]

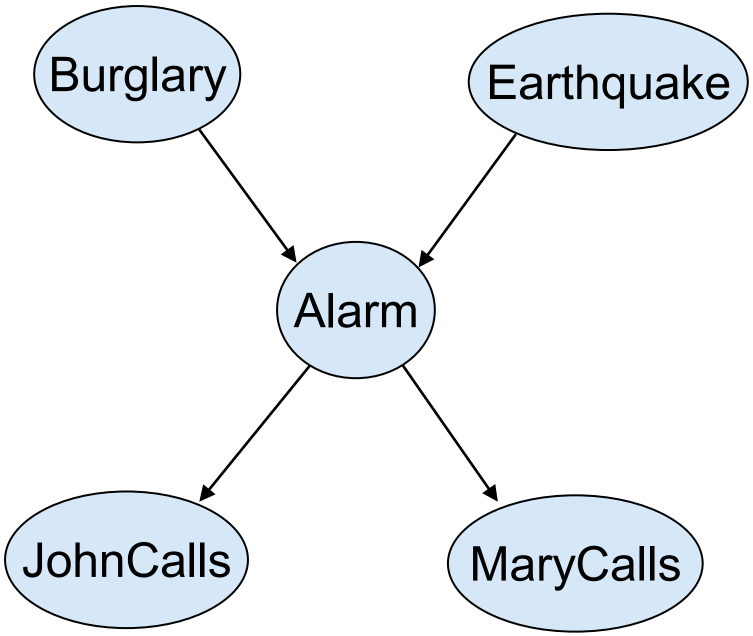
Inference Rules: An Example

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*.

Exact inference in Bayesian networks is #P-complete and needs approximation. We’ll not discuss this topic further.

Learning Bayes Nets

Parameter Learning: Method for evaluating the goodness of a given network

- Conditional distirubtions need to be learned from data
  - Maximize the joint probability of training data given the network via maximum likelihood estimation and summarize the log-likelihood of training data based on the network
- Evaluation criteria:
  - Akaike information Criterion (AIC): $-2LL + 2K$
  - Minimize AIC, with K=number of parameters

Structure Learning: 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.

Note: Bayesian vs. Frequentist Statistics

Frequentists: Models are fixed, but data varies

Bayesians: Data is fixed, but model parameters vary

The explicit consideration of *prior* distributional knowledge is a difference.
This allows us to model history. We can *update our beliefs* about parameters and derive *posterior* distributions

$$\boxed{\begin{gathered}\Pr[h|e] = \frac{\Pr[e|h]\Pr[h]}{\Pr[e]} \\ \textit{Posterior} = \textit{Likelihood} * \textit{Prior}\end{gathered}}$$

Despite the name, Bayesian networks do not necessarily imply a commitment to Bayesian statistics. Indeed, it is common to use frequentists methods to estimate the parameters of the CPTs.