




Bayesian Learning



Outline

- Introduction: probabilistic (Bayesian) methods
- MAP and ML hypotheses
- Minimum description length principle
- Bayes optimal classifier
- Naïve Bayes learner
 - example: learning over text data
- Bayesian belief networks
- (Expectation Maximization (EM): see later)
-  Mitchell Ch. 6



Two Roles for Bayesian Methods

- Provides practical learning algorithms:
 - Naive Bayes learning
 - Bayesian belief network learning
 - Combine prior knowledge (prior probabilities) with observed data
 - Requires prior probabilities
- Provides useful conceptual framework
 - Provides "gold standard" for evaluating other learning algorithms: VS, FindS, MSE, Cross Entropy
 - Additional insight into Occam's razor: MDL
- *Bayes' theorem* plays a central role



Basics of probability

- $P(A)$: probability that A happens
- $P(A|B)$: probability that A happens, given that B happens (“conditional probability”)
- Some rules:
 - complement: $P(\text{not } A) = 1 - P(A)$
 - disjunction: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
 - conjunction: $P(A \text{ and } B) = P(A) P(B|A)$
 - $= P(A) P(B)$ if A and B independent
 - total probability: $P(A) = \sum P(A|B_i) P(B_i)$



Bayes' Theorem

- $P(A|B) = P(B|A) P(A) / P(B)$
- Mainly 2 ways of using Bayes' theorem:
 - Applied to learning a hypothesis h from data D :
 - $P(h|D) = P(D|h) P(h) / P(D) \sim P(D|h)P(h)$
 - $P(h)$: priori probability that h is correct
 - $P(h|D)$: posteriori probability that h is correct
 - $P(D)$: priori probability of obtaining data D
 - $P(D|h)$: probability of obtaining data D if h is correct
 - Applied to classification of a single example e :
 - $P(\text{class}|e) = P(e|\text{class})P(\text{class})/P(e)$



Bayes' theorem: Example

- assume some lab test for a particular form of cancer has 98% chance of giving positive result if the cancer is present, and 97% chance of giving negative result if the cancer is absent
- assume furthermore 0.8% of population has this cancer: $P(\text{cancer})=0.008$ and $P(\sim\text{cancer})=0.992$
- the most probable hypothesis is $h = \sim\text{cancer}$
- given positive result, what is probability that the cancer is present?
 - $P(\text{cancer})=0.008$, $P(P|\text{cancer})=0.98$, $P(P|\sim\text{cancer})=0.03$,
 $P(P)=P(P|\text{cancer})P(\text{cancer})+P(P|\sim\text{cancer})P(\sim\text{cancer})=0.98*0.008+0.03*0.992$
 - $P(\text{cancer}|P) = P(P|\text{cancer})P(\text{cancer}) / P(P) = 0.98*0.008 / (0.98*0.008 + 0.03*0.992)=0.21$



Naïve Bayes classifier

- Simple & popular classification method
- Based on Bayes' rule + assumption of *conditional independence*
 - assumption often violated in practice
 - even then, it usually works well
- Successful application:
 - classification of text documents
 - diagnosis



Classification using Bayes rule

- Given attribute values, what is most probable value of target variable?

$$\begin{aligned}v_{MAP} &= \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n) \\ &= \arg \max_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n | v_j)P(v_j)}{P(a_1, a_2, \dots, a_n)} \\ &= \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j)P(v_j)\end{aligned}$$

- Problem: too much data needed to estimate $P(a_1 \dots a_n | v_j)$ and too many parameter estimation ($\prod |A_i|$) (2^n for n binary attributes)



The Naïve Bayes classifier

- *Naïve Bayes assumption*: attributes are *independent, given the class*
 - $P(a_1, \dots, a_n | v_j) = P(a_1 | v_j)P(a_2 | v_j) \dots P(a_n | v_j)$
 - also called *conditional independence* (given the class)
 - reduce parameter estimation from $\prod |A_i|$ ($=O(2^n)$) to $\sum |A_i|$ ($=O(n)$)
- Under that assumption, v_{MAP} becomes

$$v_{\text{NB}} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$



Naïve Bayes: Algorithm

NaiveBayesLearn(examples)

For each target value v_j

$$\hat{P}(v_j) = \text{estimate } P(v_j)$$

For each attribute value a_i of each attribute a

$$\hat{P}(a_i|v_j) = \text{estimate } P(a_i|v_j)$$

ClassifyNewInstance(x)

$$v_{NB} = \arg \max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i | v_j)$$



Naïve Bayes: Estimation

- How to estimate $P(v_j)$ and $P(a_{ij}/v_j)$?
 - standard estimate from statistics
 - estimate probability from sample proportion
 - estimate $P(v)$ as $\text{count}(v) / N$
 - estimate $P(A|B)$ as $\text{count}(A \text{ and } B) / \text{count}(B)$
 - example: 100 examples with 70 + and 30 –
 - $P(+)=0.7$ and $P(-)=0.3$
 - among 70 positive examples, 35 with $a_1=\text{SUNNY}$
 - $P(a_1=\text{SUNNY} | +)=0.5$



Examples

$$P(Y) = 9/14, P(\text{sunny} | Y) = 2/9, P(\text{cool} | Y) = 3/9$$

$$P(\text{high} | Y) = 3/9, P(\text{strong} | Y) = 3/9$$

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Naïve Bayes: Example

- Consider *PlayTennis* again, and new instance $\langle \text{Outlk}=\text{sun}, \text{Temp}=\text{cool}, \text{Humid}=\text{high}, \text{Wind}=\text{strong} \rangle$
- Want to compute:

$$v_{NB} = \arg \max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i | v_j)$$

- $\hat{P}(Y) \hat{P}(\text{sun} | Y) \hat{P}(\text{cool} | Y) \hat{P}(\text{high} | Y) \hat{P}(\text{strong} | Y) = 0.005$
 $\hat{P}(N) \hat{P}(\text{sun} | N) \hat{P}(\text{cool} | N) \hat{P}(\text{high} | N) \hat{P}(\text{strong} | N) = 0.021$
 $\Rightarrow v_{NB} = N$



Naïve Bayes: Subtleties

- What if assumption violated?
 - i.e. $P(a_1, \dots, a_n | v_j) \neq P(a_1 | v_j)P(a_2 | v_j) \dots P(a_n | v_j)$
- Prediction still equivalent to Bayes prediction as long as the following (weaker) condition holds:

$$\begin{aligned} & \arg \max_{v_j \in V} P(a_1 | v_j)P(a_2 | v_j) \dots P(a_n | v_j)P(v_j) \\ &= \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j)P(v_j) \end{aligned}$$

- But *probabilities* associated with prediction may be unrealistically close to 0 or 1



Naïve Bayes: Subtleties

- What if attribute value a_i never observed for class v_j ?
 - Estimate $P(a_i|v_j)=0$ because $\text{count}(a_i \text{ and } v_j) = 0$?
 - Effect is too strong: this 0 makes the whole product 0!
- Solution: use m-estimate smoothing
 - $$\hat{P}(a_i | v_j) = \frac{n_c + mp}{n + m}$$
 - n number of training examples with $v = v_j$
 - n_c number of examples with $v = v_j$ and $a = a_j$
 - p prior estimate for $P^{\wedge}(a_i/v_j)$ (uniform in general)
 - m number of virtual examples



Learning to classify text

- Example application:
 - Learn which news articles are of interest
 - Learn to tell which newsgroup a news article is taken from
 - Learn to classify web pages by topic
- Naïve bayes turns out to work well
 - How to apply NB?
 - Key issue : how do we represent examples?
what are the attributes?



Representation

- Attributes = word positions
 - i.e. attribute i represents i -th word in text
 - value of attribute = word that occurs there
 - $\text{doc} = (a_1=w_1, a_i=w_k, \dots, a_n=w_n)$
 - Note: could have chosen other representations; e.g. attribute = specific word, value = its frequency in the text
 - further assumption: probability of having a specific word is independent of position
 - $P(a_i=w_k | v_j) = P(a_m=w_k | v_j) = P(w_k | v_j) \quad \forall i, m$
 - $P(\text{doc} | v_j) = P(a_1=w_1, a_2=w_2, \dots, a_n=w_n | v_j)$
 $= P(w_1 | v_j)^{\text{TF}(w_1)} P(w_2 | v_j)^{\text{TF}(w_2)} \dots P(w_n | v_j)^{\text{TF}(w_n)}$
where $\text{TF}(w)$ is the term frequency of word w



Classify Texts by Naïve Bayes

- $$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$
$$= \arg \max_{v_j \in V} P(v_j) \prod_{w_k \in Voc} P(w_k | v_j)^{TF(w_k)}$$

- $$P(w_k | v_j) = \frac{n_{k,j} + 1}{n_j + |Voc|}$$



Algorithm

procedure learn_naive_bayes_text(E : set of articles, V : set of classes)

Voc = all words and tokens occurring in E

estimate $P(v_j)$ and $P(w_k|v_j)$ for all w_k in E and v_j in V :

N_j = number of articles of class j

N = number of articles

$P(v_j) = N_j/N$

n_{kj} = number of times word w_k occurs in text of class j

n_j = number of words in class j

$P(w_k|v_j) = (n_{kj}+1)/(n_j+|Voc|)$

procedure classify_naive_bayes_text(A : article)

remove from A all words/tokens that are not in Voc

return $\operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i|v_j)$

Twenty News Groups (Jochims 1996)

- Given 1000 training documents from each group
- Learn to classify new documents according to which newsgroup it came from
 - comp.graphics misc.forsale comp.os.ms-windows.misc rec.autos
 - comp.sys.ibm.pc.hardware rec.motorcycles comp.sys.mac.hardware
 - rec.sport.baseball comp.windows.x rec.sport.hockey alt.atheism
 - sci.space sci.med soc.religion.christian sci.crypt talk.religion.misc
 - sci.electronics talk.politics.mideast talk.politics.misc talk.politics.guns
- Naive Bayes: 89% classification accuracy
 - 100 most frequent words (the and of ...) are removed
 - any word occurring fewer than 3 times is removed
 - resulting vocabulary contains about 38,500 words



NewsWeeder (Lang 1995)

- learn the target concept “usenet articles that I find interesting”
- user rates netnews articles as reading them
- use rated articles as training examples
- the pool of the top 10% of automatically rated articles contains 3 to 4 times as many interesting articles as the general pool of articles read by the user



Bayesian Belief Networks



Bayesian Belief Networks

- Consider two extremes of spectrum:
 - guessing joint probability distribution
 - would yield optimal classifier
 - but infeasible in practice (too much data needed)
 - Naïve Bayes
 - much more feasible
 - but strong assumptions of conditional independence
- Can we find something in between?
 - make some independence assumptions, but only where reasonable



Representing Distributions

- G: pregnancy of a woman
- D: test by a doctor
- Joint Probability $P(G, D)$

G	D	$P(G, D)$
g^0	d^0	0.54
g^0	d^1	0.06
g^1	d^0	0.02
g^1	d^1	0.38

independent parameters: 3



Alternative Representation

- $P(G, D) = P(G)P(D|G)$

g^0	g^1	G	d^0	d^1
0.6	0.4	g^0	0.9	0.1
		g^1	0.05	0.95

$P(d^0|g^1)$: probability of false negative (5%)

$P(d^1|g^0)$: probability of false positive (10%)

independent parameters: $1+2 = 3$



Independent Tests

- H: simple but less reliable home test
- G, D and H are not independent (i.e. the outcome of one variable will affect outcomes of the others)
- Given the woman is pregnant, outcomes of D and H are independent
- Conditional Independence:
$$P(H,D|G) = P(H|G)P(D|G)$$
- $P(G,H,D) = P(G)P(H|G)P(D|G)$



Conditional Independence

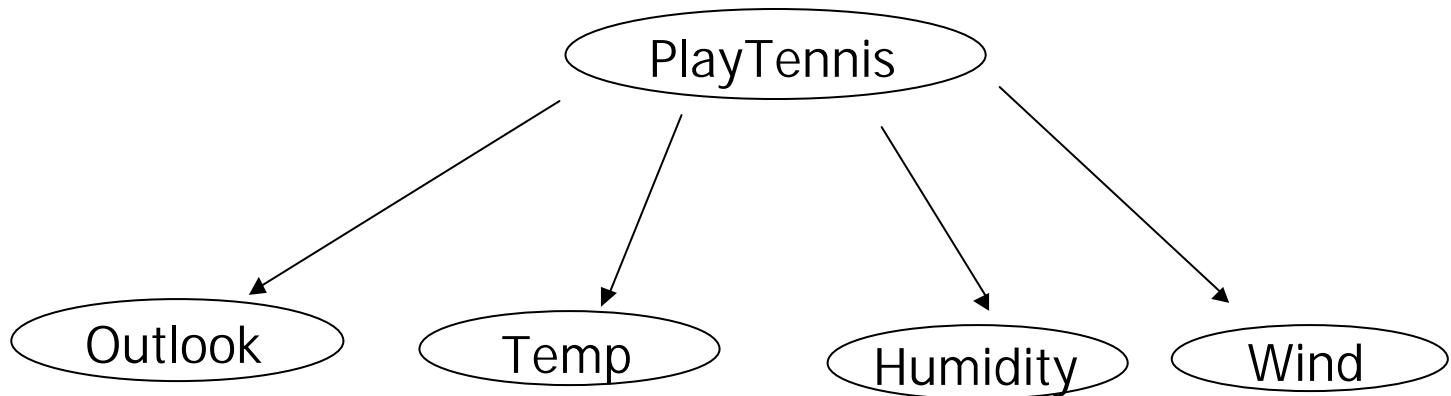
- $P(G,H,D) = P(G) P(H,D|G)$
 $= P(G) P(H|G)P(D|G)$

g^0	g^1	G	d^0	d^1	G	h^0	h^1
0.6	0.4	g^0	0.9	0.1	g^0	0.9	0.1
		g^1	0.05	0.95	g^1	0.2	0.8

Independent Parameters: $1+2+2 = 5 (<7)$

Naïve Bayes

- $I(X_i; X_j | C)$ for all i, j (i.e. attributes are independent of each other given the class)



Arcs denote direct influence



Bayesian belief networks

- Bayesian belief network consists of
 - 1: *graph*
 - intuitively: indicates which variables “directly influence” which other variables
 - arrow from A to B: A has direct effect on B
 - $\text{parents}(X)$ = set of all nodes directly influencing X
 - X is influenced only by its parents
 - formally: each node is *conditionally independent of each of its non-descendants, given its parents*
 - conditional independence: cf. Naïve Bayes
 - *X conditionally independent of Y given Z iff*
$$P(X|Y,Z) = P(X|Z)$$
 - 2: *conditional probability tables*

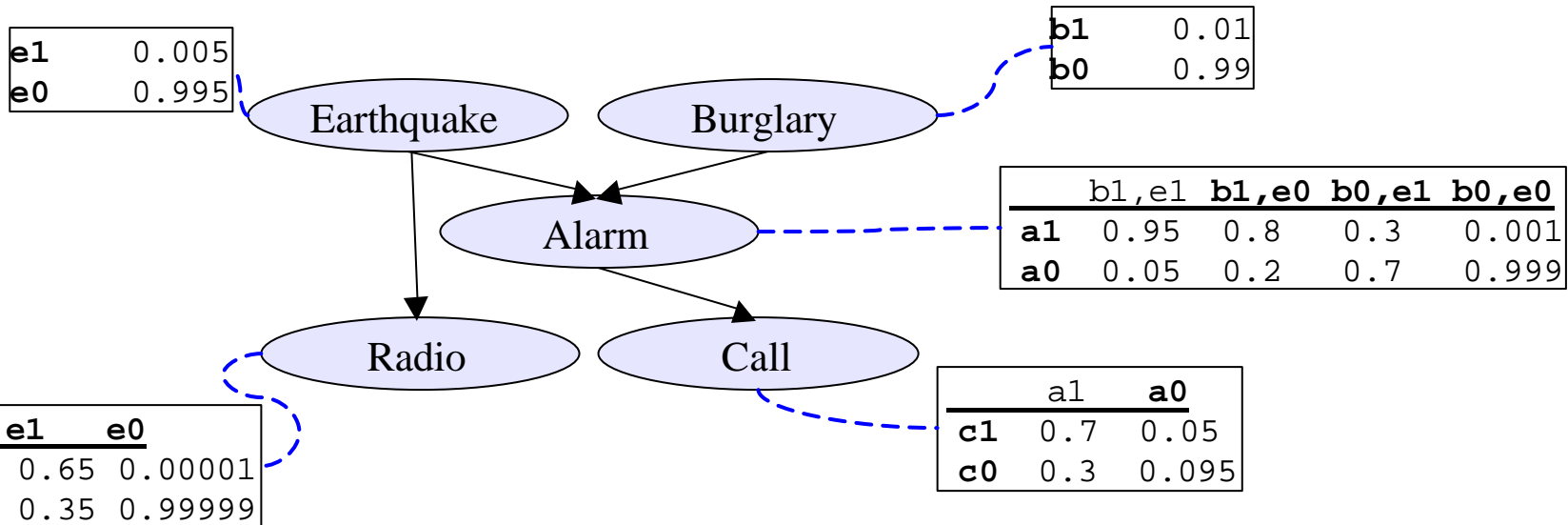


Two Ways to View BBN

- a data structure that provides the skeleton for representing a joint distribution compactly in a factorized way
- a compact representation for a set of conditional independence assumptions

Example

- Burglary or earthquake may cause alarm to go off
- Alarm going off may cause neighbor to call
- Radio report depends only on earthquake



$$P(b1, e0, a1, c1, r0) = P(b1)P(e0)P(a1 | b1, e0)P(c1 | a1)P(r0 | e0) = 0.005566428$$



Intuitive Explanation

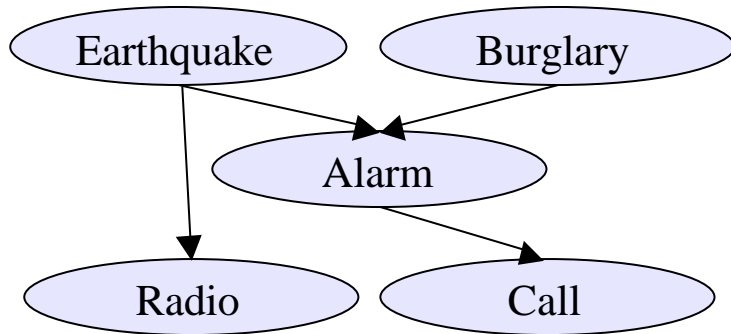
- The choice of burglary does not depend on any of the other variables
- Earthquake happens separately and independently of everything else
- Alarm can be set off either by a burglary or by an earthquake
- Neighbor calls based on whether he hears the alarm
- Radio report depends only on whether there was an earthquake



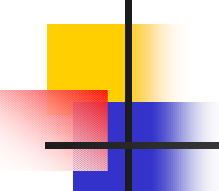
Reasoning

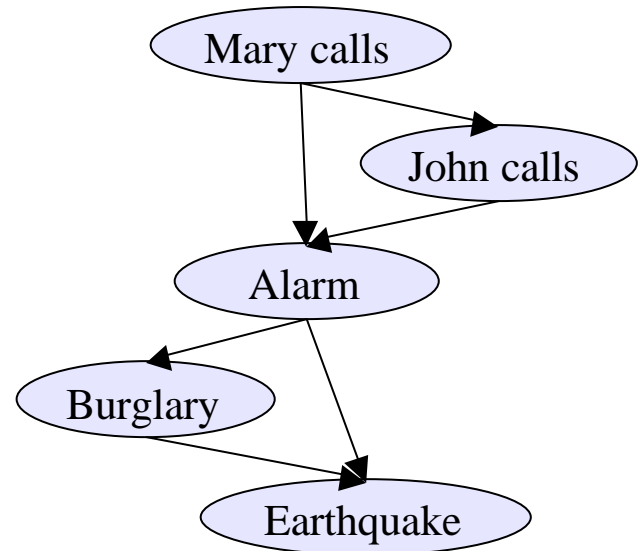
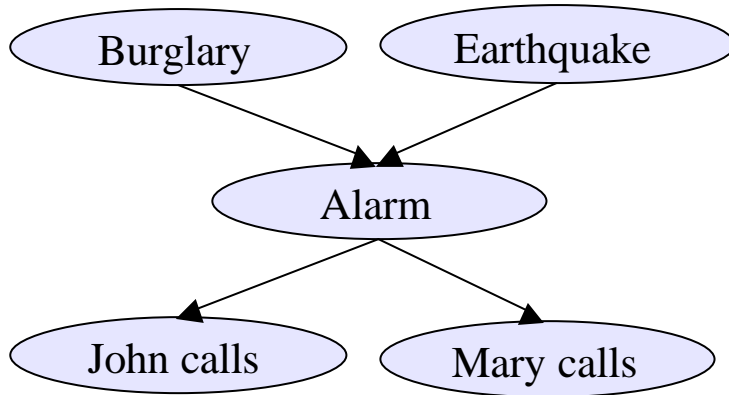
- *causal reasoning or prediction*
 - $P(c1|b1)=0.5705 > P(b1)=0.01$
 - $P(c1|b0)=0.0516$ (false positive: ~5%)
- *evidential reasoning or explanation*
 - $P(b1|c1)=0.3251 > P(b1)=0.01$
 - $P(e1|c1)=0.1034 > P(c1)=0.005$
- *explaining away or intercausal reasoning*
 - $P(e1|c1,r1)=0.9993 > P(e1|c1)=0.021$
 - $P(b1|c1,r1)=0.0268 < P(b1|c1)=0.3251$
 - earthquake provides explanation for phone call

Independence

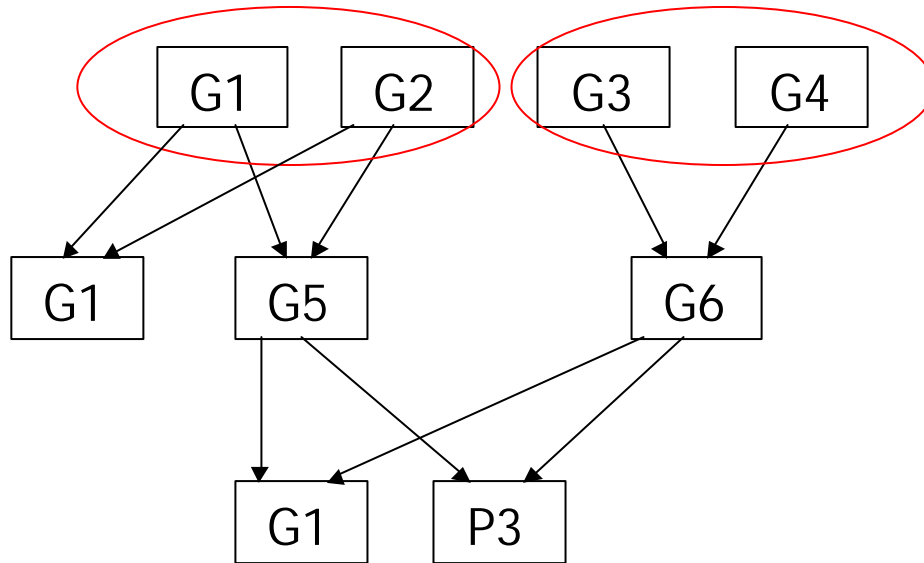


- Each variable is independent of its nondescendants given its parents
 - $I(R;B,A,C|E)$, $I(C;B,E,R|A)$
- Each variable is dependent on its descendants given its parents
 - if c_1 is observed, our probability in a_1 should go up
 - $P(a_1|b_0,e_1,c_1) > P(a_1|b_0,e_1)$

- 
- Network topology usually reflects *direct causal influences*
 - other structure also possible
 - but may render network more complex



Genetic Example



•G = A, B or O

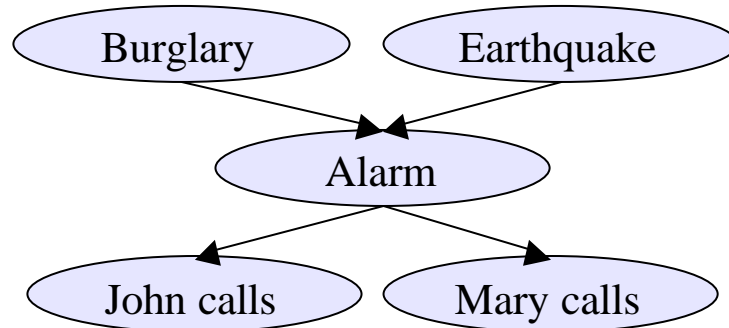
•P = A, B, AB or O



Bayesian Belief Network

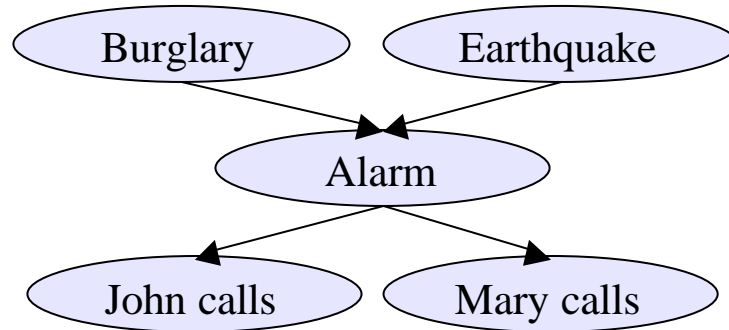
- Graph + conditional probability tables allow to construct joint probability distribution of all variables
 - $P(X_1, X_2, \dots, X_n) = \prod_i P(X_i | \text{parents}(X_i))$
 - In other words: bayesian belief network carries full information on joint probability distribution

Example



- Joint probability distribution from conditional ones:
 - $P(J,M,A,B,E) = P(J|A) P(M|A) P(A|B,E) P(B) P(E)$
 - to see this: start with $P(B)$ and $P(E)$
 - “conditionally independent from each other given their parents” = unconditionally independent, hence $P(B,E) = P(B) P(E)$
 - $P(A,B,E) = P(A|B,E) P(B,E)$ (by definition)
 - $P(J|M,A,B,E) = P(J|A)$

Example



- $P(M, J, A, B, E) = P(M|J, A, B, E)P(J, A, B, E)$
 $= P(M|A)P(J|A, B, E)P(A, B, E)$
 $= P(M|A)P(J|A)P(A|B, E)P(B, E)$
 $= P(M|A)P(J|A)P(A|B, E)P(B)P(E)$

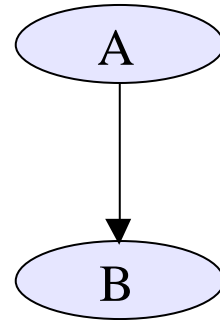


Inference

- Given values for certain nodes, infer probability distribution for values of other nodes
- General algorithm quite complicated
 - see Russel & Norvig, 1995: *Artificial Intelligence, a Modern Approach*

Simplest case: 2 nodes

- Given: $p(A)$, $p(B|A)$
 - A known, infer $p_{A=a}(B)$
 - directly from $p(B|A)$
 - A unknown, infer $p_{A=?}(B)$
 - "total probability" rule
 - B known, infer $p_{B=b}(A)$
 - Bayes' rule
 - B unknown, infer $p_{B=?}(A)$
 - $= p(A)$



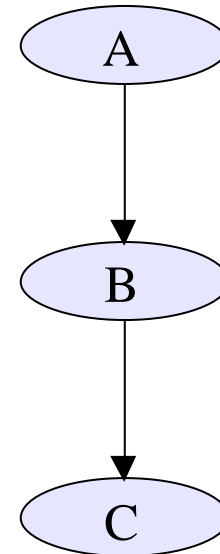
A simple 3 nodes network

- Given: $p(A)$, $p(B|A)$, $p(C|B)$
- E.g., $A=a$ and $C=c$ known:

$$P(B=b|A=a, C=c)$$

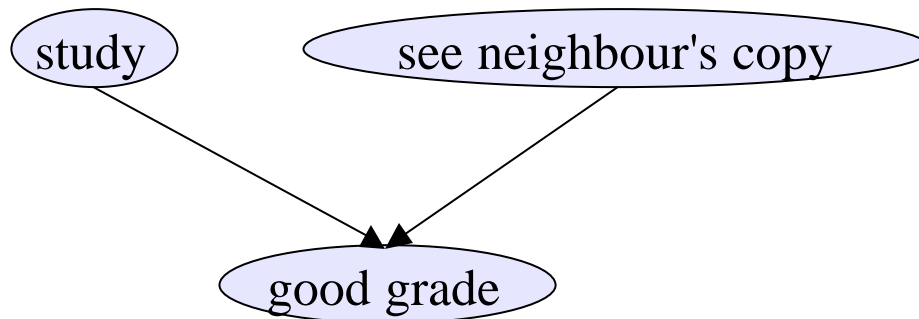
$$= \frac{P(B=b, A=a, C=c)}{\sum_i P(A=a, B=b_i, C=c)}$$

$$= \frac{P(A=a)P(B=b|A=a)P(C=c|B=b)}{\sum_i P(A=a)P(B=b_i|A=a)P(C=c|B=b_i)}$$

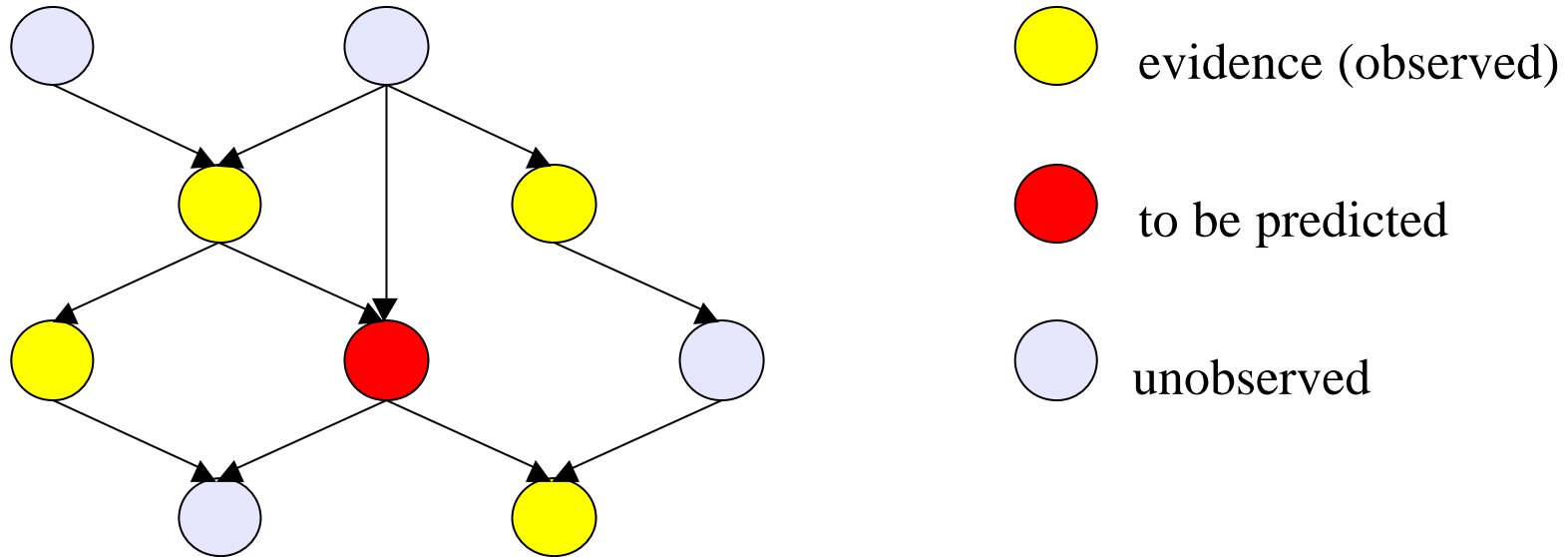


More nodes

- In general, relatively complex reasoning can be achieved
 - forward and backward reasoning
 - "explaining away"
 - "good grade" is evidence for "studied hard"
 - but is less strong evidence if known that person looked at neighbour's copy



General case

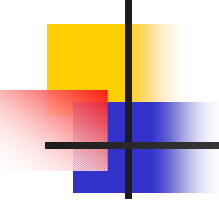


- In general: inference is NP-complete
 - approximating methods, e.g. Monte-Carlo



Learning bayesian networks

- Assume structure of network given:
 - only conditional probability tables to be learnt
 - training examples may include values for all variables, or just for some of them
 - when all variables observable:
 - estimating probabilities as easy as for Naïve Bayes
 - e.g. estimate $P(A|B,C)$ as $\text{count}(A,B,C)/\text{count}(B,C)$
 - when not all variables observable:
 - notice similarity with training neural networks (hidden units)
 - methods exist, based on gradient descent or EM (see later)

- 
-
- When structure of network not given:
 - search for structure + tables
 - e.g. propose structure, learn tables
 - propose change to structure, relearn, see whether better results
 - active research topic



Sample complexity

- When structure known and all variables observable:
 - how many examples needed for learning?
 - accurate estimates of conditional probability tables needed
 - complexity of learning linear in size of largest probability table
 - i.e. exponential in number of parent variables of node
 - compare with estimating joint distribution
 - exponential in total number of variables
 - and with naïve bayes
 - always only 1 “parent variable”, i.e. the class



To remember

- Importance of Bayes' theorem
- MAP, ML, MDL
 - definitions, characterising learners from this perspective, relationship MDL-MAP
- Bayes optimal classifier, Gibbs classifier
- Naïve Bayes: how it works, assumptions made, application to text classification
- Bayesian networks: representation, inference, learning