Tufts

Lecture 12: Independence and Bayes Nets

Artificial Intelligence (CS 131)

1

Absolute Independence

- Two random variables are independent when knowing one *tells us nothing* about the other
- Dice example: just because one die comes up 5, this tells us nothing about what the next die will be
  - The conditional probability of the second die being 5, given that the first one was, is the *same as before*
- To get the probability of two 5's, we multiply
  - $1/6$ of the time, the 1st die comes up 5
  - Of those cases, $1/6$ of the time, the 2nd die is 5 too
  - Thus, the joint probability is $(1/6 \times 1/6) = 1/36$

2

Defining Independence

- General Definition: X and Y are independent whenever the following holds (all are equivalent):

1. $P(X \mid Y) = P(X)$
2. $P(Y \mid X) = P(Y)$
3. $P(X, Y) = P(X)P(Y)$

3

Examples of Independence

- Given a fair coin, the result of the current flip is independent of all prior flips:
  - $P(H) = 0.5$
  - $P(H \mid T) = P(H) = 0.5$
  - $P(H \mid TTHHHTTTHH) = P(H) = 0.5$
  - $P(H \mid HHHHHHHHHHHHHHHHHHHH) = P(H) = 0.5$
  - Examples abound in games of chance (cards, dice, roulette)
  - Independence assumptions make many calculations and models much simpler (*if they correctly apply*)

4

Examples of Non-Independence

- Many real-life variables are NOT independent!
- Often, knowing that the one variable holds makes the other *much more likely* to hold true:

$$P(PassTest \mid StudyTest) \gtrsim P(PassTest)$$

$$P(Fire \mid Smoke) \gtrsim P(Fire)$$

$$P(TorontoWins \mid RedSoxError) \gtrsim P(TorontoWins)$$

- Other times, evidence can make it *much less likely*:

$$P(BucsWin \mid BradyRetires) \lesssim P(BucsWin)$$

$$P(Innocent \mid SmokingGun) \lesssim P(Innocent)$$

$$P(HealthCare \mid Poverty) \lesssim P(HealthCare)$$

5

Simplifying Joint Distributions Using Independence Relations

- Suppose we have a joint distribution on cavities and toothaches:

	$toothache$	$\neg toothache$
$cavity$	0.1	0.02
$\neg cavity$	0.08	0.8

- Now consider an independent weather variable with distribution:

$sunny$	$\neg sunny$
0.7	0.3

- We *could* combine the two into a single table:

	$sunny$		$\neg sunny$	
	$toothache$	$\neg toothache$	$toothache$	$\neg toothache$
$cavity$	0.07	0.014	0.03	0.006
$\neg cavity$	0.056	0.56	0.024	0.24

- Or, *save space* and simply multiply when we need to calculate one of the combinations, storing fewer values overall (4 vs. 7)

6

Conditional Independence

- Sometimes, we find that two variables are not independent when considered *by themselves*
- However, if we had some *other evidence* (i.e., knew the value of other variables), then conditional upon that new evidence, they *are in fact* independent
- Again, we can define this in three equivalent ways; we say that X and Y are conditionally independent given Z iff:

1. $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$
2. $P(X \mid Y, Z) = P(X \mid Z)$
3. $P(Y \mid X, Z) = P(Y \mid Z)$

7

An Example

- A classic case is multiple effects with a *common cause*
- Consider disease diagnosis, with 3 binary variables: B (patient has bubonic plague), P (patient has pustules), and G (patient has swollen glands)
  - Generally, there are all sorts of dependencies between the two symptoms: i.e., having one can make the other more likely
- However, if we actually *know* if someone has plague or not, then the chance of having pustules *does not* depend on whether they have swollen glands (i.e., it gives us no additional information)
  1. $P(P = true \mid B = true, G) = P(P = true \mid B = true)$
  2. $P(P = true \mid B = false, G) = P(P = true \mid B = false)$
  - Variables P and G are conditionally independent given B

8

Example, continued

- The full joint distribution on 3 binary variables (B, P, G) requires (2 x 2 x 2) – 1 = 7 values
- However, we can use the chain rule, along with conditional independence to reduce this number:

$$P(P,G,B) = P(P \mid G,B)P(G \mid B)P(B)$$

$$= P(P \mid B)P(G \mid B)P(B)$$

  - Now we only need 2 values for each of the first 2 distributions, plus 1 for the second: **5 total values**
  - As the number of variables increases in a model or system, this sort of savings can grow very large

9

Naïve Bayesian Models

- Plague example reflects a common assumption that effects (symptoms) are all completely independent of one another, given a common cause (disease)
- If we let D be a disease, and s_1, s_2, …, s_n be the symptoms, this naïve Bayes assumption gives us:

$$P(D, s_1, s_2, \ldots, s_n) = P(D)P(s_1 \mid D)P(s_2 \mid D)\ldots P(s_n \mid D)$$

- Thus, we can calculate probability of disease given symptoms:

$$P(D \mid s_1, \ldots, s_n) = \frac{P(D, s_1, \ldots, s_n)}{P(s_1, \ldots, s_n)}$$

$$= \frac{P(D)P(s_1 \mid D)P(s_2 \mid D)\ldots P(s_n \mid D)}{P(s_1, \ldots, s_n)}$$

10

Updating Naïve Beliefs

- This basic factored representation allows us to *update* our beliefs easily given new evidence
- Our *current belief* in the joint probability of disease and symptoms is given by the formula:

$$P(D, s_1, \ldots, s_n) = P(D)\prod_{i=1}^{n} P(s_i \mid D)$$

- When a new symptom (s_{n+1}) appears, we can then easily account for it as follows, with minimal need to re-calculate:

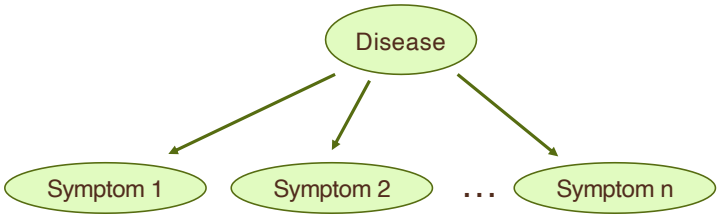
$$P(D, s_1, \ldots, s_n, s_{n+1}) = P(D)\prod_{i=1}^{n+1} P(s_i \mid D)$$

$$= P(D, s_1, \ldots, s_n)\, P(s_{n+1} \mid D)$$

11

A Graphical Representation

- We could represent our assumption about connections between disease and symptoms using a graphical model
  - Nodes are the variables (disease, symptoms)
  - Edges/arrows show connections between them
    - No edge if things are independent of one another



12

Bayesian Networks (BNs)

- Generalizes the naïve disease example:

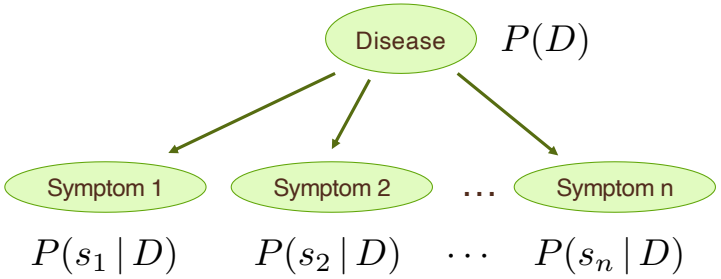
1. Set of nodes, one per variable.
2. Edges compose a directed acyclic graph (DAG).
   - Each arrow-link, $A \rightarrow B$, between variables represents *influence* of variable A on variable B.
   - Cycles (loops) are forbidden, which both represents common sense, and makes certain algorithms (which we will see later) work properly
3. Conditional probability tables (CPTs) for each node, conditioned on its parents (if it has none, this is just a simple, non-conditional, one-variable distribution):

$$P(Child \mid Parent_1, Parent_2, \ldots, Parent_n)$$

13

A Graphical Representation

- In disease example, with n possible symptoms we would have (n + 1) CPTs, including one for disease itself:



14

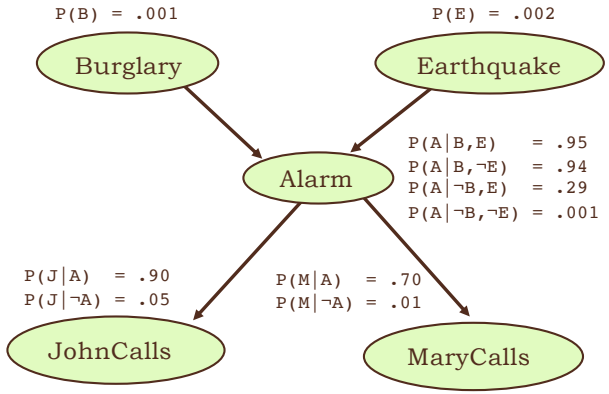
Joint Probability in BNs

- Each BN is capable of representing a full joint probability distribution over its node variables, which means:

1. All probability questions of interest can be answered by looking at the numbers from the BN
2. Memory cost of individual CPTs can be significantly smaller than that for the explicit full joint distribution
3. Inference can be made much more simple

15

An Example: Earthquakes & Burglaries



16

Parameters of a BN

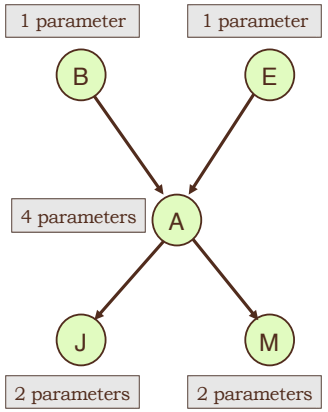
- In general, to specify a CPT in a Bayes Network, we need:
  - A CPT entry for every possible outcome of parent nodes
    - For nodes without parents, this is just the prior distribution
  - n–1 entries for each such, where n is the number of values the child node can have
- Thus, if $|X|$ is the number of values that a variable X has, the total number of parameters is:

$$\sum_{X \in BN} \left[(|X| - 1) \times \prod_{A \in Parents(X)} |A| \right]$$

17

Parameters of a BN, cont'd.

- Thus for the earthquake case, where all the variables are binary, we require:
  1. 1 value for P(B) and P(E)
  2. 4 values for P(A | B, E)
  3. 2 values for P(J | A) and P(M | A)
- The number of values needed gets smaller as the BN has more and more independence



18

General Principles for Building BN's

- To build a BN for a given domain, we need to:
  1. Figure out what *variables* we need (nodes)
  2. Figure out how they *depend upon* one another (edges)
  3. Give *probabilities* for these dependencies (CPTs)
- Often, identifying each of these is a matter of expert opinion or research

19

An Algorithm for Building BNs

Loop for ($i = 1 \ldots n$) variables:

1. Pick variable X_i to add to the graph.
2. Find *smallest set* of parents you can, choosing them so that the new variable is conditionally independent of *all others previously* added, *given* those parents:

$$P(X_i \mid X_1, X_2, \ldots X_{i-1}) = P(X_i \mid Parents(X_i))$$

3. Draw arrows from $Parents(X_i)$ to X_i
4. Specify the CPT: $P(X_i \mid Parents(X_i))$

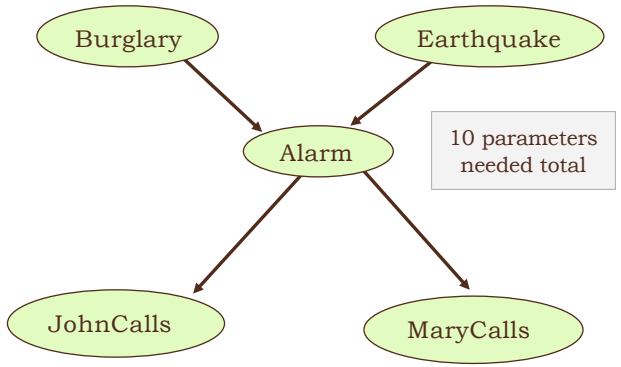
20

Properties of the Algorithm

- The graph is always acyclic
  - Arrows always go from existing prior nodes to new ones (and never vice-versa)
  - No looping back to nodes added before
- Result encodes all the probabilistic information we ever need to calculate any questions we might ask about the variables in the system
- *Compactness and sparseness* (number of connections) depends upon the order in which variables are considered when adding them to the BN, however

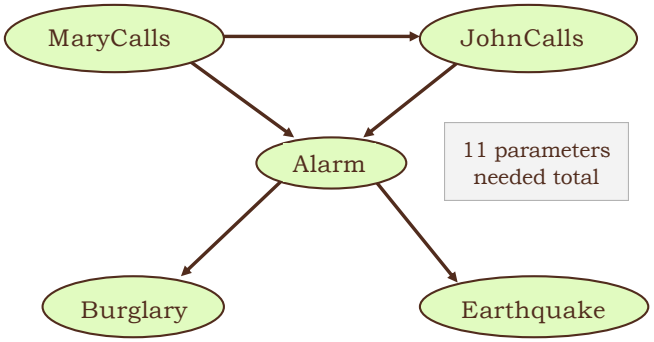
21

Original BN: Built Using Order (B, E, A, J, M)



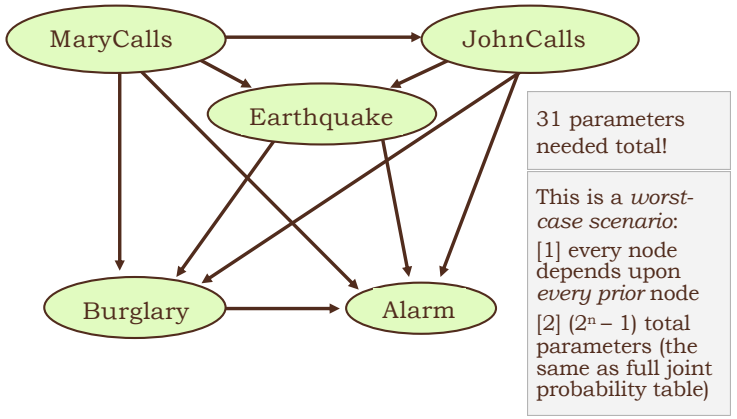
22

Another Order: (M, J, A, B, E)



23

Still Another Order: (M, J, E, B, A)



24