Lectures: Bayes Nets

## Directed Graphical Models: Bayes Nets

Probabilistic graphical models represent a joint distribution over a set of variables.

$$P(X_1, \ldots, X_m)$$

Given a joint distribution, we can reason about unobserved variables given observations (evidence). Here's an **applet** (model file: **http://www.cs.cmu.edu/~javabayes/Examples/CarStarts/car-starts.bif**) that illustrates reasoning with a Bayes Net. Here's an example from my own work (with Ben Sapp) of using graphical models for detection and parsing of human figures:

Pictorial Structure model (Bayes net)

Posterior distribution over head/torso/limbs locations
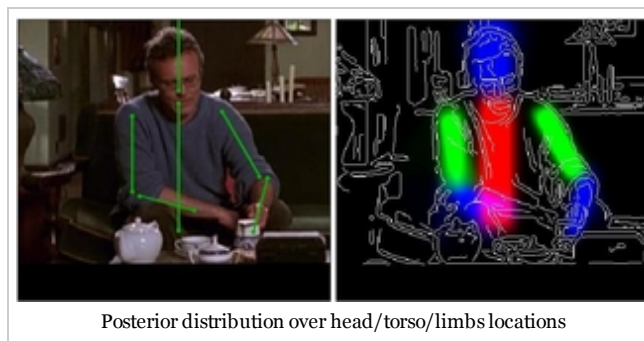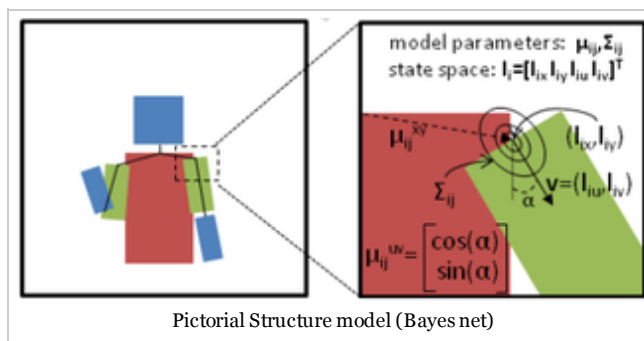
The **Naive Bayes model** is an example of a joint distribution model (where one of the variables is the class variable Y):

$$P(Y, X_1, \ldots, X_m) = P(Y) \prod_j P(X_j|Y)$$

We used the Naive Bayes model *diagnostically*: Given the effects (features) $X_1 = x_1, \ldots, X_m = x_m$, we used Bayes rule to compute $P(Y|x_1, .., x_m)$, the posterior over the *cause*. In one of the midterm problems, you also saw that we don't need to know all the features, but for example can compute $P(Y|x_1)$. In general, the joint distribution allows us to compute the posterior distribution over any set of variables given another set of variables.

Representing the Joint

The joint distribution of a set of variables is an exponentially-sized object. If all the variables are binary, the joint over m variables has $2^m - 1$ parameters. The Naive Bayes model is much more concise: it uses conditional independence assumptions to restrict the types of distributions it can represent and uses $O(m)$ parameters. The key to concise representation is *factoring*: representing the whole with a product of interrelated parts. The key ingredients in this representation are the chain rule, Bayes Rule and conditional independence.

Chain Rule : $P(X_1, \ldots, X_m) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1) \ldots P(X_m|X_1, \ldots, X_{m-1})$
.

The chain rule by itself does not lead to concise representation: we need to assume conditional independencies to reduce the size of the representation. Let's consider some very simple examples. We want to reason about three variables: Traffic (whether there will be traffic on the roads tomorrow), Rain (whether it will rain tomorrow), Umbrella (whether I'll bring my umbrella with me tomorrow). These random variables are clearly correlated. Using the chain rule, we have:

$$P(R, T, U) = P(R)P(T|R)P(U|T, R)$$

Now if we assume that my decision about the umbrella does not depend on traffic once I know if it's raining, i.e. $P(U|T, R) = P(U|R)$, then we have

$$P(R, T, U) = P(R)P(T|R)P(U|R)$$

Conditional Independence:
X_i is *conditionally independent* of X_j given X_k, denoted as
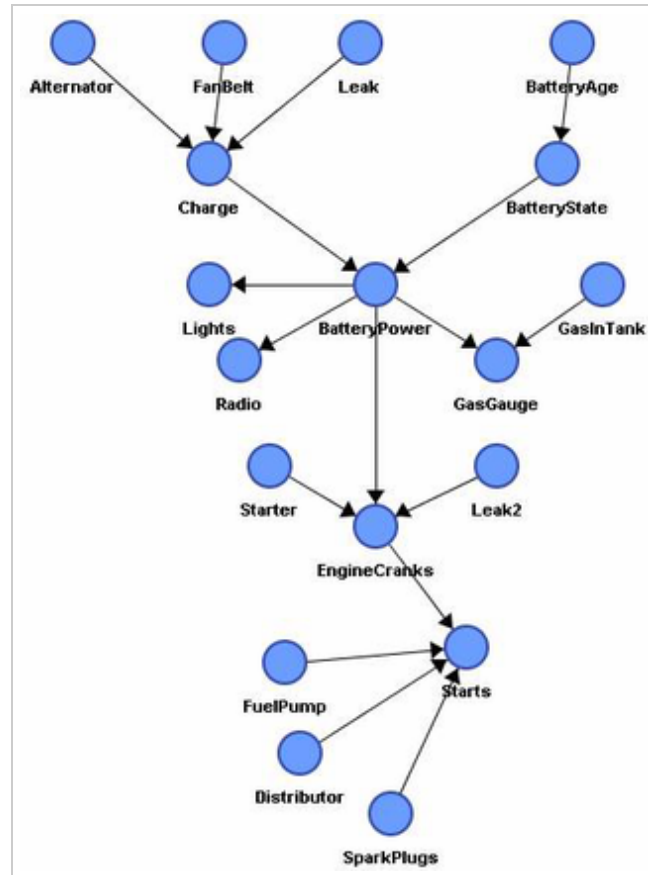
$$X_i \perp X_j | X_k$$

if

$$P(X_i = x_i \mid X_k = x_k, X_j = x_j) = P(X_i = x_i \mid X_k = x_k), \ \forall x_i, x_j, x_k$$

or equivalently:

$$P(X_i = x_i, X_j = x_j \mid X_k = x_k) = P(X_i = x_i \mid X_k = x_k)P(X_j = x_j \mid X_k = x_k), \ \forall x_i, x_j, x_k.$$

Graphical models use conditional independence assumptions for efficient representation, inference and learning of joint distributions. The above model for R,T,U can be represented as a directed graph, $U \leftarrow R \rightarrow T$, where nodes correspond to variables and edges correspond to direct correlations, which we will define precisely below.

Here's a toy graphical model for a car operation.

Alternator FanBelt Leak BatteryAge

Charge BatteryState

Lights Battery Power GasInTank

Radio GasGauge

Starter Leak2

EngineCranks

FuelPump Starts

Distributor

SparkPlugs

Graphical models are simplified descriptions of how some portion of the world works. They will usually not include every possible variable or all interactions between variables, but focus on the most relevant variables and strongest interactions. Their probabilistic nature allows reasoning about unknown variables given partial evidence about the world. Directed graphical models, know as Bayesian Networks, Bayes Nets, Belief Nets, etc, can be thought of as causal models (with some important caveats). Typically, causal models (if they exist) are the most concise representation of the underlying distribution. There are several types of reasoning we can do with such models:

- diagnostic reasoning: from effects to causes
- causal reasoning: from causes to effects
- mixtures of the two

Graphical Models represent joint distributions using simple, local interactions models which describe how variables depend on each other. Local interactions are then chained together to give global, indirect interactions. In case of directed graphical models (Bayes nets), local interactions are conditional distributions. The graphical representation encodes conditional independence assumptions.
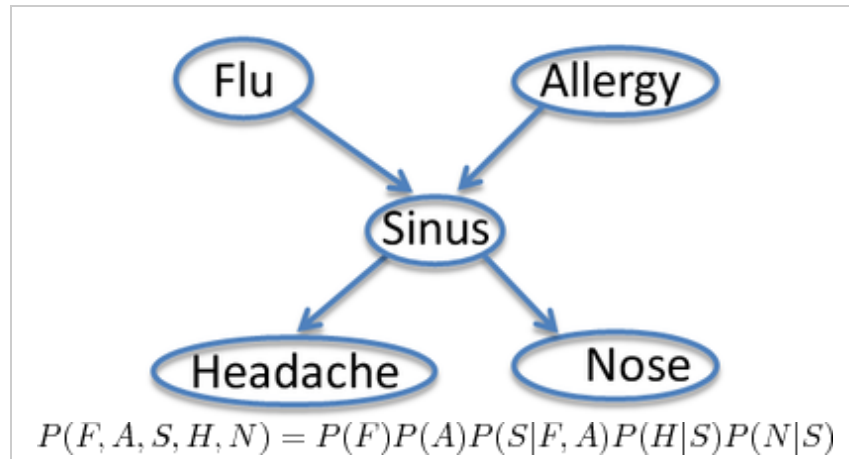
Graphs and Conditional Independence Assumptions

Let's consider a graphical model for a slightly more complex example:
- The flu causes sinus inflammation
- Allergies cause sinus inflammation
- Sinus inflammation causes a runny nose
- Sinus inflammation causes headaches

A Bayes net which encodes the assumptions for this example is shown below. Nodes represent variables

$$P(F, A, S, H, N) = P(F)P(A)P(S|F, A)P(H|S)P(N|S)$$

We have 5 variables, so there are $2^5 - 1$ entries in the joint. How many parameters do we need for all the conditional distributions above?
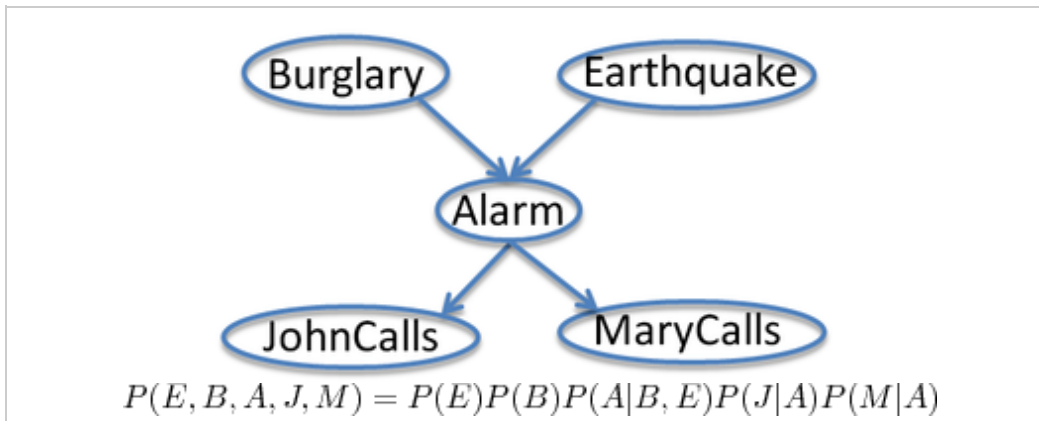
The graph encodes the following conditional independence assumptions (and more).
- $F \perp A$
- $\neg(F \perp N)$
- $F \perp N | S$
- $\neg(H \perp N)$
- $H \perp N | S$

Here's another classic example, due to Judea Pearl, who was instrumental to popularizing Bayes nets and probabilistic reasoning in AI. It helps to know that he lives in LA.

John calls to say my house alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar? The variables are: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls. Our model will reflect the following assumptions:
- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

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

In general, the graph encodes the following basic assumption, from which many others can be derived:

Local Markov Assumption (LMA): A variable X is independent of its nondescendants given its parents (and only its parents) :

$$X \perp NonDescendants_X | Pa_X$$

For our purposes here, a node X' is a non-descendant of X if X' is not X and not a parent of X and there is no directed path from X to X'.

Parameters: Conditional Distributions

The directed graph encodes independencies. The numerical value of the joint distribution is encoded in local conditional distributions. In case of discrete variables, these are called CPTs (conditional probability tables).

$$P(E, B, A, J, M) = P(E)P(B)P(A|B, E)P(J|A)P(M|A)$$
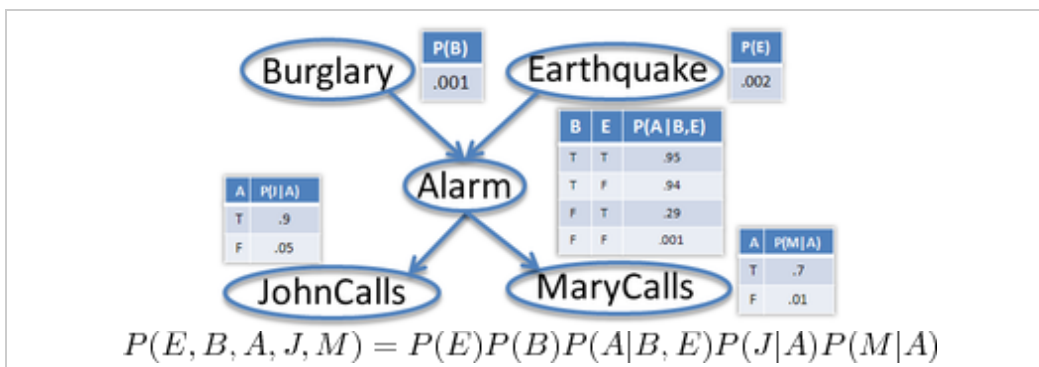
The Local Markov Assumption allows us to decompose the joint in terms of the CPTs:

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

Why? Let's look at the general case.

General Bayes Net Definition
- Set of random variables: $\{X_1, \ldots, X_m\}$
- Directed acyclic graph (DAG): nodes = variables

- Conditional Distributions: CPTs $P(X_j|Pa_{X_j})$
- Local Markov Assumption: A variable X is independent of its non-descendants given it parents: $(X \perp NonDescendants_X|Pa_X)$
- Joint distribution is given by: $P(X_1, \ldots, X_m) = \prod_j P(X_j|Pa_{X_j})$

The last statement about the joint distribution is due to the following construction. Take a *topological sort* [1] of the DAG and renumber the variables such that all the children of a node appear after it in the ordering. Then

$$P(X_1, \ldots, X_m) = \prod_j P(X_j|X_{j-1}, \ldots, X_1) = \prod_j P(X_j|Pa_{X_j})$$

by LMA: since none of j's children appear before j in the ordering, $X_{j-1}, \ldots, X_1$ only contains its parents and non-descendants.

Properties of Conditional Independence

The Local Markov Assumption specifies the set of basic independencies in a Bayes net. We can use properties of conditional independence to infer many others.

- Symmetry:

$$(X \perp Y|Z) \rightarrow (Y \perp X|Z)$$

- Decomposition:

$$(X \perp Y, W|Z) \rightarrow (X \perp Y|Z)$$

- Weak union:

$$(X \perp Y, W|Z) \rightarrow (X \perp Y|Z, W)$$

- Contraction:

$$(X \perp W|Y, Z), (X \perp Y|Z) \rightarrow (X \perp Y, W|Z)$$

- Intersection (only for positive distributions P()>0)

$$(X \perp Y|W, Z), (X \perp W|Y, Z) \rightarrow (X \perp Y, W|Z)$$

Conditional Independence In Bayes Nets

For the Flu-Sinus network, the LMA (plus symmetry and decomposition) gives the following indepedencies:
- $F \perp A$
- $F \perp N|S$
- $H \perp N|S$

Explaining Away

Using the Flu-Sinus network, a reasonable setting of parameters could give:

$$P(F = t) = 0.1, \quad P(F = t|A = t) = 0.1, \quad P(F = t|S = t) = 0.8, \quad P(F = t|S = t, A = t) = 0.5$$

that is $(F \perp A)$, $\neg(F \perp A \mid S)$. This property of distributions, called explaining away, can be encoded in the V-structures of a Bayes net. The CPTs shown for the Burglar-alarm network above exhibit exactly this explaining away property.

3-Node BNs

For 3 node BNs, there are several possible DAG structures: fully disconnected (1), fully connected (6), partially disconnected (6) and the following 4 cases:

1. X has indirect causal effect on Z: $X \to Y \to Z$
2. X has indirect evidential effect on Z: $X \leftarrow Y \leftarrow Z$
3. X has common cause with Z: $X \leftarrow Y \to Z$
4. X has common effect with Z: $X \to Y \leftarrow Z$

In the first 3 cases, we have $\neg X \perp Z$ but $X \perp Z \mid Y$, and in the fourth, the V-structure, we have $X \perp Z$ but $\neg(X \perp Z \mid Y)$

In the first three cases, X influences Z (without observations) but the influence is stopped by observing Y, while in cases like 4, X does not influence Z (without observations) but the influence is *enabled* by observing Y. In the Flu-Sinus network, observing S correlates F and A because of possible explaining away. More generally, observing any descendant of S correlates F and A since we have evidence about S which could be explained away by one of the causes of S.

Active Trails

Definition: A simple trail $\{X_1, X_2, \cdots, X_k\}$ in the graph (no cycles) is an **active trail** when variables $O \subset \{X_1, \cdots, X_m\}$ are observed if for each consecutive triplet in the trail:

- $X_{i-1} \to X_i \to X_{i+1}$, and X_i is not observed ($X_i \notin O$)

- $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$, and X_i is not observed ($X_i \notin O$)

- $X_{i-1} \leftarrow X_i \to X_{i+1}$, and X_i is not observed ($X_i \notin O$)

- $X_{i-1} \to X_i \leftarrow X_{i+1}$, and X_i is observed ($X_i \in O$), or one of its descendants is observed

Directed Separation (d-sep) Theorem: Variables X_i and X_j are independent given $O \subset \{X_1, \cdots, X_m\}$ if there is no active trail between X_i and X_j when variables O are observed.

The same theorem applies for sets of variables A,B: if there are no active trails between any pairs of variables $(X_i \in A, X_j \in B)$ when O are observed, then $(A \perp B|O)$.

Here are some examples:

$$(A{\perp}G),\ (A{\perp}C|B),\ (A{\perp}C|B,D),\ \neg(A{\perp}C\mid B,D,I)$$
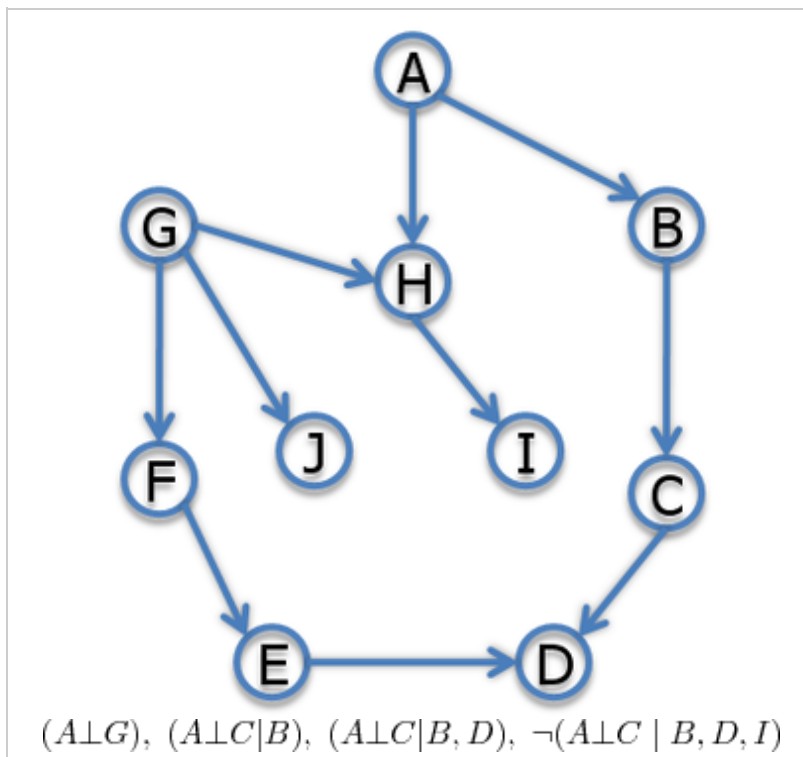
Representation Theorem

Definition: $I(G)$ is the set of all independencies implied by directed separation on graph G.

Definition: $I(P)$ is the set of all independencies in distribution P.

BN Representation Theorem

$$I(G) \subseteq I(P) \Leftrightarrow P(X, \ldots, X_m) = \prod_i P(X_i | Pa_{X_i})$$

This theorem is important because it tells us that every P has at least one BN structure G that represents it and conversely if P *factors* over G, then we can read (a subset of) I(P) from G.

Copyright © 2005–2013 the Main wiki and its authors

Links

1. **en.wikipedia.org/wiki/Topological_sorting**