6.036 Introduction to Machine Learning

(meets with 6.862)

Bayesian Networks (Chapter 11 in notes)

Administrivia

HW4 will be out by the end of the week.

Project 3 Due Friday 5/5 at 9PM.

Drop Date: this Thursday 4/27.

As always:

- Check LMOD/Piazza for announcements.
- To contact staff, use Piazza (6036-staff@lists.csail.mit.edu for exceptions only)

Probabilistic models

- **Probabilistic** models to explain the structure of data

- E.g., mixture models (e.g., mixture of Gaussians), models with latent variables, Hidden Markov models…

- Want to learn how to:
  - Specify them (joint distribution, parameter values)
  - Sample from them (as generative models)
  - Estimate them from data

- Today: **Bayesian Networks**

Bayesian networks

Rich class of generative models, combining **graphs** and **probability**

Two main elements in a Bayesian network:

- A **directed acyclic graph (DAG)** over the variables
- An associated **probability distribution**

Why both?

- Graph makes explicit and summarizes useful properties of the underlying distribution
- We can understand how to use the graph structure for efficient inference (marginals and conditionals).

Bayesian networks

Nodes of the graph are associated to **random variables**
Arcs of the graph represent **dependencies** between vars

We've already seen a few examples!

- Mixtures of distributions
- Hidden Markov Models (HMM)

Bayesian Networks subsume these, and many more…

Example (I)

Three binary variables (coin flips H/T, and True/False)

- Person 1 flips a fair coin: variable $\mathbf{X_1}$
- Person 2 flips a fair coin: variable $\mathbf{X_2}$
- Person 3 checks whether the coin flips resulted in the same value: variable $\mathbf{X_3} = [[\mathbf{X_1} = \mathbf{X_2}]]$

Examples:

- $\mathbf{X_1} = H$, $\mathbf{X_2} = T$, $\mathbf{X_3} = F$
- $\mathbf{X_1} = H$, $\mathbf{X_2} = H$, $\mathbf{X_3} = T$

Example (II)

- Can easily describe the distributions of $\mathbf{X_1}$ and $\mathbf{X_2}$ (e.g., P($\mathbf{X_1}$=H)=P($\mathbf{X_1}$=T) = 1/2 — and similarly for $\mathbf{X_2}$)

$$X_1 : \begin{array}{c|cc} & H & T \\ \hline & 0.5 & 0.5 \end{array}, \quad X_2 : \begin{array}{c|cc} & H & T \\ \hline & 0.5 & 0.5 \end{array}$$

- For $\mathbf{X_3}$, need to specify the *conditional* distribution P($\mathbf{X_3}$=x_3 | $\mathbf{X_1}$=x_1, $\mathbf{X_2}$ = x_2)

$$X_3|X_1, X_2 : \begin{array}{c|cc} X_1, X_2 & \mathrm{T} & F \\ \hline H, H & 1 & 0 \\ H, T & 0 & 1 \\ T, H & 0 & 1 \\ T, T & 1 & 0 \end{array}$$

Example (III)

Recall that $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent coin flips.

From this, can write the *joint distribution* over the three variables

$$P(\mathbf{X_1}=x_1, \mathbf{X_2} = x_2, \mathbf{X_3}=x_3) =$$
$$P(\mathbf{X_1}=x_1)\, P(\mathbf{X_2} = x_2)\, P(\mathbf{X_3}=x_3 \mid \mathbf{X_1}=x_1, \mathbf{X_2} = x_2)$$

Notice that it *factors,* since the first two coin flips are <u>independent</u>.

Can we represent this in terms of a graph?

Example (IV)

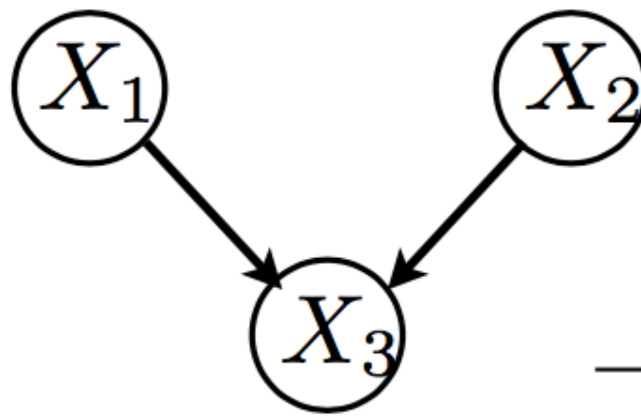
A more convenient way: *in addition* to the distribution, use a directed graph that makes the structure obvious:

X_1 :

H	T
0.5	0.5

X_2 :

H	T
0.5	0.5



$X_3 | X_1, X_2$:

X_1, X_2	T	F
H, H	1	0
H, T	0	1
T, H	0	1
T, T	1	0

Properties

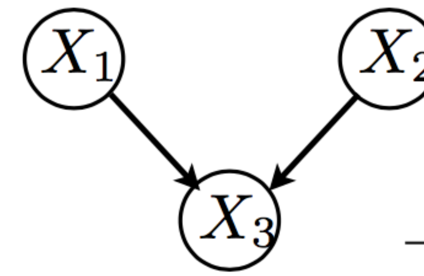
X_1 :

H	T
0.5	0.5

X_2 :

H	T
0.5	0.5

Factorization of the distribution determined by the graph



$X_3 | X_1, X_2$:

X_1, X_2	T	F
H, H	1	0
H, T	0	1
T, H	0	1
T, T	1	0

$$P(\mathbf{X_1}=x_1, \mathbf{X_2} = x_2, \mathbf{X_3}=x_3) = P(\mathbf{X_1}=x_1)\, P(\mathbf{X_2} = x_2)\, P(\mathbf{X_3}=x_3 \mid \mathbf{X_1}=x_1, \mathbf{X_2} = x_2)$$

Notice graph has *no cycles*

We say that $\mathbf{X_1}$ (or $\mathbf{X_2}$) is a **parent** of $\mathbf{X_3}$

Similarly, $\mathbf{X_3}$ is a **child** of $\mathbf{X_1}$ (or $\mathbf{X_2}$)

General Bayesian Networks

- Always defined by **acyclic graphs** (no directed cycles)

- Distribution factors according to the graph:
  - If no parents, write $P(\mathbf{X_i}=x_i)$
  - Otherwise, product of conditional probabilities of variables, given parents, e.g., $P(\mathbf{X_i}=x_i \mid \mathbf{X_j}=x_j, \mathbf{X_k} = x_k, \mathbf{X_l} = x_l)$

- Describes how to *generate (sample)* from the model

- Graph structure yields useful insights about dependence and independence of the variables

- E.g.: *marginal independence* and *induced dependence*

Marginal Independence

Recall our model:

$P(\mathbf{X_1}=x_1, \mathbf{X_2} = x_2, \mathbf{X_3}=x_3) =$

$\qquad P(\mathbf{X_1}=x_1)\ P(\mathbf{X_2} = x_2)\ P(\mathbf{X_3}=x_3 \mid \mathbf{X_1}=x_1, \mathbf{X_2} = x_2)$

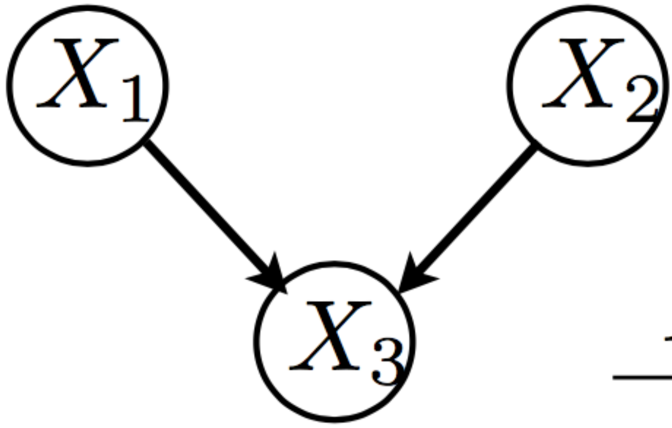
- What is the <u>marginal distribution</u> of $(\mathbf{X_1}, \mathbf{X_2})$?

$P(\mathbf{X_1}=x_1, \mathbf{X_2} = x_2) =$

$\quad = \Sigma_{x3}\ P(\mathbf{X_1}=x_1)\ P(\mathbf{X_2} = x_2)\ P(\mathbf{X_3}=x_3 \mid \mathbf{X_1}=x_1, \mathbf{X_2} = x_2)$

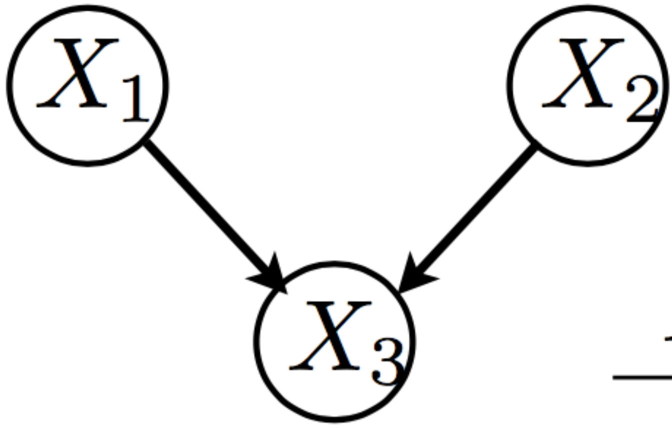
$\quad = P(\mathbf{X_1}=x_1)\ P(\mathbf{X_2} = x_2)\ \Sigma_{x3}\ P(\mathbf{X_3}=x_3 \mid \mathbf{X_1}=x_1, \mathbf{X_2} = x_2)$

$\quad = P(\mathbf{X_1}=x_1)\ P(\mathbf{X_2} = x_2)$



Thus, $\mathbf{X_1}$ and $\mathbf{X_2}$ are marginally independent.
Easy to see directly from the graph!

Induced Dependence



Recall that $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent.

What if we measure $\mathbf{X_3}$? Say, $\mathbf{X_3}$=T ?

What do we know now?

Either $\mathbf{X_1}$=$\mathbf{X_2}$=H or $\mathbf{X_1}$=$\mathbf{X_2}$=T.

Values are now *dependent,* and this dependence is induced by the additional knowledge (measuring $\mathbf{X_3}$).

Again, easy to see directly from the graph.

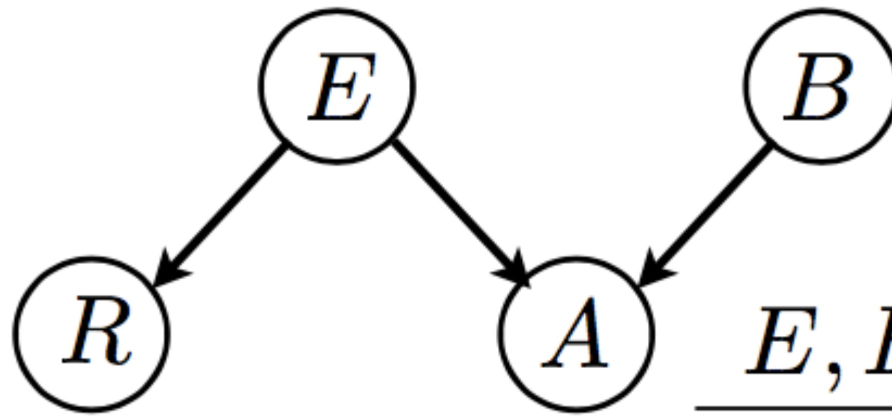
Alarm Example

E:

	T	F
	0.01	0.99

B:

	T	F
	0.01	0.99



R:

E	T	F
T	1	0
F	0	1

$A|E,B$:

E,B	T	F
T,T	1	0
T,F	1	0
F,T	1	0
F,F	0	1

Binary (T/F) variables:

- **B**: Burglary
- **E**: Earthquake
- **R**: Radio Report
- **A**: Alarm

Alarm example (II)

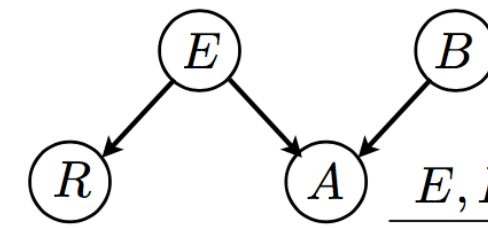
- Can write the joint distribution:

$E:$

	T	F
	0.01	0.99

$B:$

	T	F
	0.01	0.99



$R:$

E	T	F
T	1	0
F	0	1

$A|E,B:$

E,B	T	F
T,T	1	0
T,F	1	0
F,T	1	0
F,F	0	1

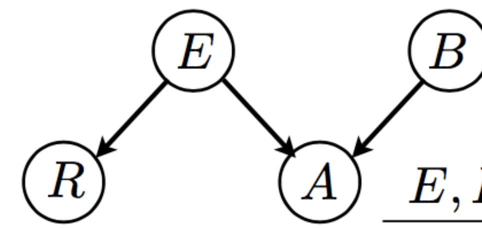
P(**E**=e, **B**=b, **A**=a, **R**=r) =
P(**E**=e) P(**B**=b) P(**A**=a|**E**=e, **B**=b) P(**R**=r|**E**=e)

As before, factors along terms “variable given its parents”.

Reasoning in BN

$E:$	T	F
	0.01	0.99

$B:$	T	F
	0.01	0.99



$R:$

E	T	F
T	1	0
F	0	1

$A|E,B:$

E,B	T	F
T,T	1	0
T,F	1	0
F,T	1	0
F,F	0	1

If alarm goes off (**A**=T),
what can we deduce?

Either **E** (earthquake) or
B (burglary) occurred — or both.

Two competing explanations, equally likely.

Let's compute the posterior probability that there was a burglary…

(Can always use brute force, but can we be a bit more clever?)

Reasoning

Marginal over (**B**,**A**):

$$
\begin{aligned}
&P(B=b, A=T) = \\
&= \sum_{e\in\{T,F\}} \sum_{r\in\{T,F\}} P(E=e)P(B=b)P(A=T|E=e,B=b)P(R=r|E=e) \\
&= \sum_{e\in\{T,F\}} P(E=e)P(B=b)P(A=T|E=e,B=b) \sum_{r\in\{T,F\}} P(R=r|E=e) \\
&= \sum_{e\in\{T,F\}} P(E=e)P(B=b)P(A=T|E=e,B=b) \\
&= P(B=b) \sum_{e\in\{T,F\}} P(E=e)P(A=T|E=e,B=b)
\end{aligned}
$$

Notice that **R** drops out (why?)

The *conditional* (prob. burglary, given alarm) is now:

$$P(B=T|A=T) = \frac{P(B=T, A=T)}{\sum_{b\in\{T,F\}} P(B=b, A=T)}$$

Intuitively, what do you think it should be? Let's compute it!

"Explaining away"

- Now we hear an earthquake radio report (i.e., **R**=T).
- How do our beliefs change?
- In our case, **R**=T implies **E**=T (earthquake occurred).
- Thus, this explains the alarm, and removes any evidence of burglary (**B**=T).
- Additional info (report) *explained away* evidence of burglary. Now, we have:

$$P(\mathbf{B}=T \mid \mathbf{A}=T, \mathbf{R}=T) = P(\mathbf{B}=T)=0.01$$

$$P(\mathbf{E}=T \mid \mathbf{A}=T, \mathbf{R}=T) = 1$$

(show formally!)

Summary - Bayesian Networks

- Rich class of generative models
- Two key elements: a directed acyclic graph, and a (compatible) probability distribution
- Dependence/independence properties are directly reflected in (and can be read from) graph structure
- Makes possible systematic, efficient algorithms for reasoning and inference