Directed Graphical Models

Felipe José Bravo Márquez

September 2, 2021

Directed Graphical Models

- Probabilistic graphical models (PGMs) provide a visual representation of the underlying structure of a joint probablity distribution [Ruozzi,].
- In this class we will focus on directed graphical models (DGMs), which are one type of PGM.
- Directed graphical models (DGMs) are a family of probability distributions that admit a compact parametrization that can be naturally described using a **directed acyclic graph**.
- DGMs are also known as **Bayesian networks**.
- We won't use that term in this class because statistical inference for DGMs can be performed using frequentist or Bayesian methods [Wasserman, 2013].
- These types of graphs are also called **causal graphs** or **causal diagrams**.

Directed Graphical Models

- DGMs link two very different branches of mathematics: probability and graph theory.
- They also have intriguing connections to philosophy, particularly the question of **causality**.
- At the same time, they are widely used in statistics and machine learning.
- DGMs can be used to solve problems in fields as diverse as medicine, language processing, vision, and many others [Ermon and Kuleshov,].
- But before introducing them more formally we need to learn the following mathematical concepts:
  1. Conditional independence
  2. The chain rule of probability
  3. Directed acyclical graphs (DAGs)

Conditional Independence

- Two random variables X and Y are independent, written $X \perp Y$ if $f(x, y) = f(x)f(y)$ for all values of x, y.
- This also implies that $f(x|y) = f(x)$ and $f(y|x) = f(y)$.
- Notice that f can be either a density function for continous random variables or a probability mass function for discrete random variables.
- In the same way $f(x|y)$ and $f(y|x)$ correspond to conditional densities or mass functions.
- Now, suppose we have three random variables A, B, C.
- A and B are conditionally independent given C, written $A \perp B|C$, if:

$$f(a|b, c) = f(a|c)$$

for all a, b and c.

Conditional Independence

- Intuitively, $A \perp B|C$ means that, once you know C, B provides no extra information about A.
- Notice that $A \perp B|C$ doesn't necessarily imply that $A \perp B$.

Example

- Let H, V, A be three random variables representing a person's height, vocabulary and age.
- H and V are dependent $f(v|h) \neq f(v)$ since very small people tend to be children, known for their more basic vocabularies.
- But knowing that two people are 19 years old (i.e., conditional on age) there is no reason to think that one person's vocabulary is larger if we are told that they are taller:

$$f(v|h,a) = f(v|a) \Leftrightarrow V \perp H|A$$

The chain rule of probability

- For a set of random variables $X_1, \ldots, X_n$, the chain rule of probability allow us to express the joint probability function $f(x_1, x_2, \ldots, x_n)$ as a product of n conditional probabilities:

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2|x_1)f(x_3|x_2, x_1)\ldots f(x_n|x_{n-1}, \ldots, x_2, x_1).$$

- For example for the case of $n = 3$

$$f(x_1, x_2, x_3) = f(x_1)f(x_2|x_1)f(x_3|x_2, x_1)$$

using the definition of conditional probabilities we have that

$$f(x_2|x_1) = f(x_1, x_2)/f(x_1)$$

and

$$f(x_3|x_2, x_1) = f(x_1, x_2, x_3)/f(x_1, x_2)$$

- By replacing these expressions into the chain of products, many expressions cancel out and we obtain $f(x_1, x_2, x_3)$.

Joint Probabilities

- Expressing a joint probability function can be expensive.
- For example if there are m binary random variables, the complete distribution is specificied by $2^m - 1$ joint probabilities.
- For example, if we have two Boolean variables (A, B), we need the probabilities $\mathbb{P}(A, B)$, $\mathbb{P}(\neg A, B)$, $\mathbb{P}(A, \neg B)$, and $\mathbb{P}(\neg A, \neg B)$.
- A joint distribution for a set of random variables gives all the information there is about the distribution.
- Note that for m boolean variables, the joint distribution contains 2^m values.
- However, the sum of all the joint probabilities must be 1 because the probability of all possible outcomes must be 1.
- Thus, to specify the joint distribution, one needs to specify 2^{m-1} numbers.

Joint Probabilities

- The main idea of DGMs is that by assuming that some variables are conditionally independent we can use the probability chain rule to represent the joint distribution more compactly.
- For example in the previous example, a conditional independence assumption could be: $X_3 \perp X_2 | X_1$: $f(x_3|x_2, x_1) = f(x_3|x_2)$.
- Then our joint distribution would be reduced to:

$$f(x_1, x_2, x_3) = f(x_1)f(x_2|x_1)f(x_3|x_2)$$

- In the following slides we will clarify these concepts with the Student example given in [Koller and Friedman, 2009].

The Student Example

- Consider the problem faced by a company trying to hire a recent college graduate.
- The company's goal is to hire intelligent employees, but there is no way to test intelligence directly.
- However, the company has access to the student's SAT scores[1], which are informative but not fully indicative.
- Thus, our probability space is induced by the two random variables Intelligence (I) and SAT (S).
- For simplicity, we assume that each of these takes two values: $Val(I) = \{i_1, i_0\}$, which represent the values high intelligence (i_1) and low intelligence (i_0).

[1] The SAT is a standardized test widely used for college admissions in the United States.

The Student Example

- Similarly $Val(S) = \{s_1, s_0\}$, which also represent the values high (score) and low (score), respectively.
- Thus, our joint distribution in this case has four entries.
- For example, one possible joint distribution f would be:

I	S	$f(i,s)$
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

- Notice that we would need 3 parameters to specify this joint distribution (2^{m-1}).
- Alternatively, we could use the the chain rule of conditional probabilities to represent the joint distribution as follows:

$$f(i,s) = f(i)f(s|i)$$

The Student Example

- Other factorizations obtained by changing the order of the variables are also valid (e.g., $f(i,s) = f(s)f(i|s)$).
- However, it is convenient to represent the process in a way that is more compatible with causality.
- Various factors (e.g., genetics, upbringing) first determined (stochastically) the student's intelligence.
- Her/his performance on the SAT is determined (stochastically) by her/his intelligence.
- We note that the models we construct are not required to follow causal intuitions, but they often do.

The Student Example

- From a mathematical perspective, $f(i,s) = f(i)f(s|i)$ leads to an alternative way of representing the joint distribution.
- Instead of specifying the various joint entries $f(i,s)$, we would specify it in the form of $f(i)$ and $f(s|i)$.
- Thus, we could represent the joint distribution using two tables.
- The first one representing the marginal distribution of I.

i^0	i^1
0.7	0.3

- These numbers can be easily verified from the previous table. For example

$$f(i^0) = f(i^0, s^0) + f(i^0, s^1) = 0.665 + 0.035 = 0.7$$

The Student Example

- The second table represents the conditional probability distribution (CPD) of S given I:

I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

- which can also be verified from previous table using the Bayes theorem. For example:

$$f(s^0|i^0) = f(s^0, i^0)/f(i^0) = 0.665/0.7 = 0.95$$

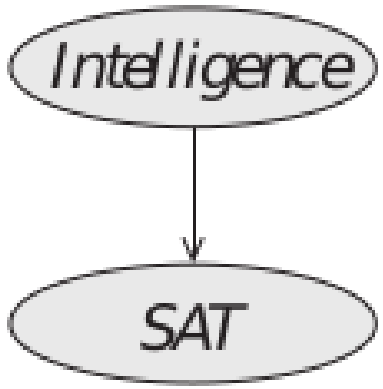
- The CPD $f(s|i)$ represents the probability that the student will succeed on his SATs in the two possible cases:
  1. The case where the student's intelligence is low.
  2. The case where it is high.

The Student Example

- The CPD asserts that a student of low intelligence is extremely unlikely to get a high SAT score ($f(s_1|i_0) = 0.05$).
- On the other hand, a student of high intelligence is likely, but far from certain, to get a high SAT score ($f(s_1|i_1) = 0.8$).
- It is instructive to consider how we could parameterize this alternative representation.
- Here, we are using three Bernoulli distributions, one for $f(i)$, and two for $f(s|i_0)$ and $f(s|i_1)$.
- Hence, we can parameterize this representation using three independent parameters, say θ_{i^1} ,$\theta_{s^1|i^1}$, and $\theta_{s^1|i^0}$
- Recall that the original representation also required 3 parameters.

The Student Example

- Thus, although the conditional representation is more natural than the explicit representation of the joint, it is not more compact.
- However, as we will soon see, the conditional parameterization provides a basis for our compact representations of more complex distributions.
- Although we haven't defined directed acyclical graphs (DAGs) yet, it is instructive to see how this example would be represented as one.



- The DAG from above has a node for each of the two random variables I and S, with an edge from I to S representing the direction of the dependence in this model.

The Student Example

- Let's assume now that the company also has access to the student's grade G in some course.
- In this case, our probability space is the joint distribution over the three relevant random variables I, S, and G.
- We will assume that I and S are as before, and that G takes on three values g^1, g^2, g^3, representing the grades A, B, and C, respectively.
- So, the joint distribution has twelve entries.
- We can see that for any reasonable joint distribution function f, there are no independencies that hold.
- The student's intelligence is clearly correlated both with SAT score and grade.

The Student Example

- The SAT score and grade are also not independent.
- A high SAT score increases the chances of getting a high grade.
- Thus, we may assume that, for our particular distribution f, $f(g^1|s^1) > f(g^1|s^0)$.
- However, it is quite plausible that our distribution f in this case satisfies a conditional independence property.
- If we know that the student has high intelligence, a high grade on the SAT no longer gives us information about the student's performance in the class.
- More formally: $f(g|s,i) = f(g|i)$ or $G \perp S|I$
- So the join distribution can be reduced from

$$f(i,s,g) = f(i)f(s|i)f(g|i,s)$$

to

$$f(i,s,g) = f(i)f(s|i)f(g|i)$$

The Student Example

- Now our joint distribution is specified by $f(i)$, $f(s|i)$, and $f(g|i)$.
- While $f(i)$, $f(s|i)$ might be the same as before, and $f(g|i)$ might be:

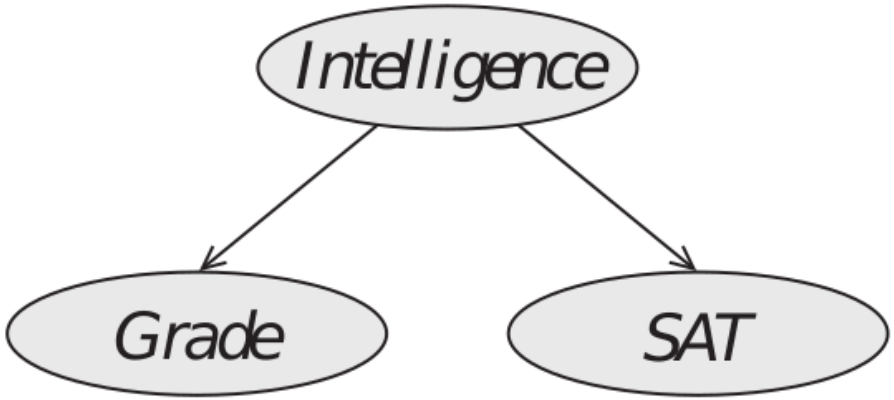
I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

- Now we can calculate the joint probability of any combinatio of imputs.
- For example

$$
\begin{aligned}
f(i^1, s^1, g^2) &= f(i^1) f(s^1|i^1) f(g^2|i^1) && (1)\\
&= 0.3 * 0.8 * 0.17 && (2)\\
&= 0.0408 && (3)
\end{aligned}
$$

The Student Example

- The corresponding DAG would be as follows:



- In this case, the alternative parameterization is more compact than the joint.
- We now have three Bernoulli distributions $f(i)$, $f(s|i^1)$ and $f(s|i^0)$, and two three-valued categorical distributions $f(g|i_1)$ and $f(g|i^0)$.
- A categorical distribution is discrete probability distribution that describes the possible results of a random variable that can take on one of K possible categories, with the probability of each category separately specified[2].

[2] https://en.wikipedia.org/wiki/Categorical_distribution

The Student Example

- Each of the Bernoullis requires one independent parameter, and each three-valued categorical requires two independent parameters, for a total of seven.
- By contrast, our joint distribution has twelve entries, so that eleven independent parameters are required to specify an arbitrary joint distribution over these three variables.
- It is important to note another advantage of this way of representing the joint: modularity.
- When we added the new variable G, the joint distribution changed entirely.
- Had we used the explicit representation of the joint, we would have had to write down twelve new numbers.
- In the factored representation, we could reuse our local probability models for the variables I and S, and specify only the probability model for G, the CPD $f(G|I)$.
- This property will turn out to be invaluable in modeling real-world systems.

Directed Acyclic Graphs (DAGs)

- Now that we understand how the chain rule of probability and conditional independence assumptions enable us to obtain a compact representation of a joint distribution, we are in a position to introduce the DAGs more formally.
- A directed graph consists of a set of nodes with arrows between some nodes.
- Graphs are useful for representing independence relations between variables.
- More formally, a directed graph G consists of a set of vertices V and an edge set E of ordered pairs of vertices.
- For our purposes, each vertex corresponds to a random variable.
- If $(Y, X) \in E$ then there is an arrow pointing from Y to X.

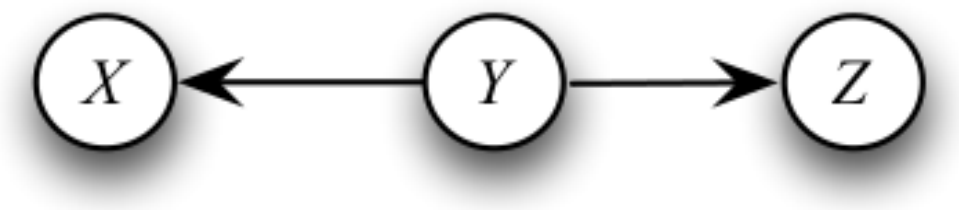


Figure: A directed graph with vertices $V = \{X, Y, Z\}$ and edges $E = \{(Y, X), (Y, Z)\}$.

Directed Acyclic Graphs (DAGs)

- If an arrow connects two variables X and Y (in either direction) we say that X and Y are adjacent.
- If there is an arrow from X to Y then X is a parent of Y and Y is a child of X.
- The set of all parents of X is denoted by π_X or $\pi(X)$.
- A directed path between two variables is a set of arrows all pointing in the same direction linking one variable to the other such as the chain shown below:

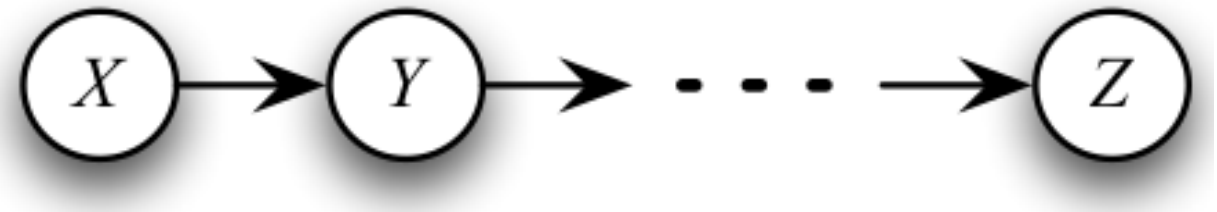


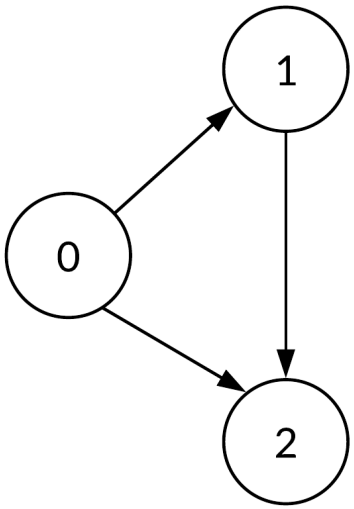
Figure: A chain graph with a directed path.

- X is an ancestor of Y if there is a directed path from X to Y (or X = Y).
- We also say that Y is a descendant of X.

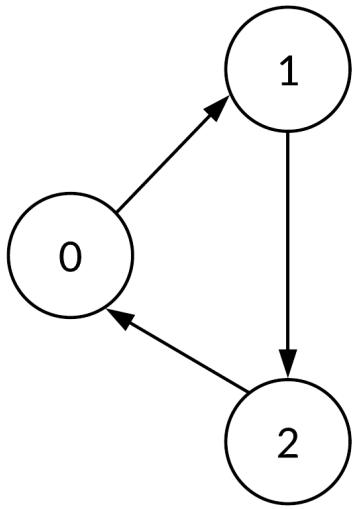
Directed Acyclic Graphs (DAGs)

- A directed path that starts and ends at the same variable is called a cycle.
- A directed graph is acyclic if it has no cycles.
- In this case we say that the graph is a directed acyclic graph or DAG.

Acyclic Graph



Cyclic Graph



- From now on, we only deal with directed acyclic graphs since it is very difficult to provide a coherent probability semantics over graphs with directed cycles.

Probability and DAGs

- Let G be a DAG with vertices $V = (X_1, \ldots, X_i, \ldots, X_d)$.
- Each vertex X_i is random variable
- The order of vertices $(1, \ldots, i, \ldots, d)$ forms a topological sort on the random variables.
- This means that every variable comes before all its descendants in the graph.
- If F is a distribution for V with probability function $f(x)$ (density or mass), we say that G represents F, if

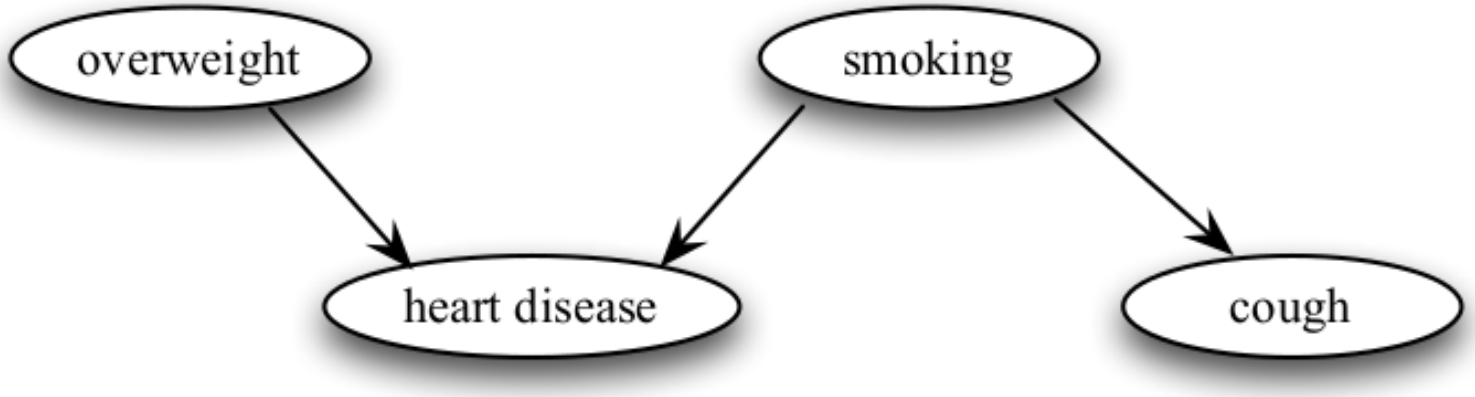
$$f(x) = \prod_{j=1}^{d} f(x_j|\pi_{x_j})$$

where π_{x_j} is the set of parent nodes of X_j

- So, a DAG used to encode a multivariate probability distribution F with given conditional independence relations is a Directed Graphical Model or Bayesian Network.
- Warning: It is common to find in the literature the terms DAG, DGM or Bayesian network used interchangeably.

Probability and DAGs

- The next figure shows a DAG with four variables.



- The probability function takes the following decomposition:
- f(overweight, smoking, heart disease, cough) $=$ f(overweight) $\times$ f(smoking) $\times$ f(heart, disease| overweight, smoking) $\times$ f(cough|smoking).

Causal DAGs

- The interpretation of direct acyclic graphs as carriers of independence assumptions does not necessarily imply causation.
- In fact, it will be valid for any set of recursive independencies along any ordering of the variables, not necessarily causal or chronological.
- However, the ubiquity of DAG models in statistical and AI applications stems (often unwittingly) primarily from their causal interpretation
- That is, as a system of processes, one per family, that could account for the generation of the observed data.
- It is this causal interpretation that explains why DAG models are rarely used in any variable ordering other than those which respect the direction of time and causation. [Pearl, 2009]

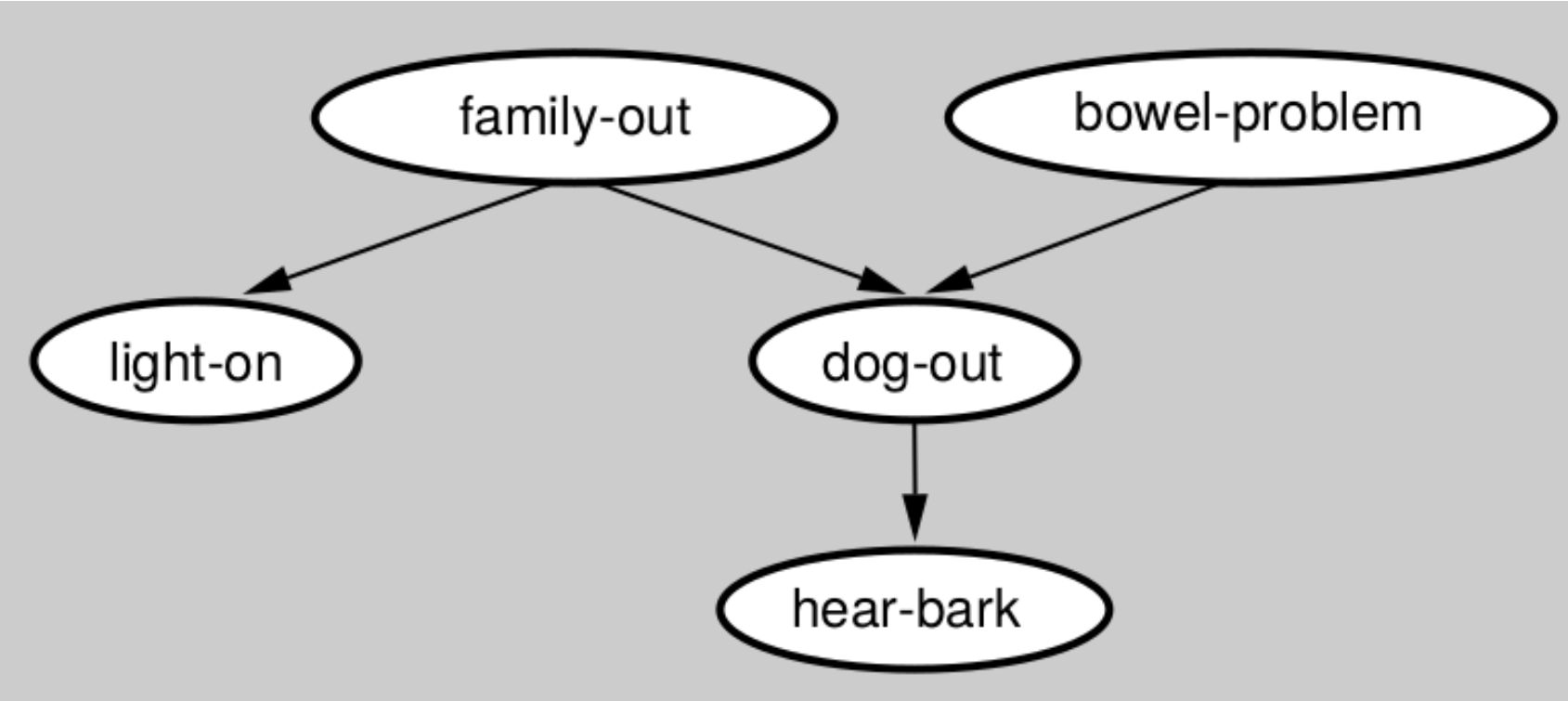
An Example

- The best way to understand DAGs is to imagine trying to model a situation in which causality plays a role.
- And also our understanding of what is actually going on is incomplete
- So we need to describe things probabilistically.
- The following example is based on [Charniak, 1991].
- Eugene Charniak is a famous AI researcher who's got the following situation.
- When he goes home at night, he wants to know if his family is home before trying the doors.
- Often when his wife leaves the house, she turns on an outdoor light.

An Example

- Eugene's wife can also turn on the outdoor light if she is expecting a guest.
- Also, they have a female dog.
- When nobody is home, the dog is put in the back yard.
- The same is true if the dog has bowel troubles.
- Finally, if the dog is in the backyard, Eugene's will probably hear her barking
- The next slide shows a DAG encoding all the above causal relationships.

An Example



- The DAG can help to predict what will happen in a particular scenario (if his family goes out, the dog goes out)
- Or to infer causes from observed effects (if the light is on and the dog is out, then his family is probably out).

D-separation

- sdsad

Plate Notation

- sdsad

Estimation for DAGs

- Two estimation questions arise in the context of DAGs.
- First, given a DAG $\mathcal{G}$ and data $d_1, \ldots, d_n$ from a distribution f consistent with $\mathcal{G}$, how do we estimate f?
- Second, given data $d_1, \ldots, d_n$ how do we estimate $\mathcal{G}$?
- The first question is pure estimation while the second involves model selection.
- These are very involved topics and are beyond the scope of this course.
- We will just briefly mention the main ideas.

Estimation for DAGs

- If we are doing frequentist inference, we typically use some parametric model $f(x|\pi_x;\theta_x)$ for each conditional density.
- The likelihood function is then

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(d_i;\theta)) = \prod_{i=1}^{n}\prod_{j=1}^{m} f(X_{ij}|\pi_j;\theta_j)$$

- where X_{ij} is the value of X_j for the ith data point and j are the parameters for the-jth conditional density.
- We can then estimate the parameters by maximum likelihood.
- On the other hand, if we want to perform Bayesian inference we must set priors for all our variables $X_1, \ldots, X_m$ and estimate the posterior accordingly.

Estimation for DAGs

- To estimate the structure of the DAG itself, we could fit every possible DAG using maximum likelihood and use AIC (or some other method) to choose a DAG.
- However, there are many possible DAGs so we would need much data for such a method to be reliable.
- Also, searching through all possible DAGs is a serious computational challenge.
- Producing a valid, accurate confidence set for the DAG structure would require astronomical sample sizes.
- If prior information is available about part of the DAG structure, the computational and statistical problems are at least partly ameliorated [Wasserman, 2013].

Conclusions

- Blabla

References I


Charniak, E. (1991).
Bayesian networks without tears.
AI magazine, 12(4):50–50.


Ermon, S. and Kuleshov, V.
Cs228 notes.


Koller, D. and Friedman, N. (2009).
Probabilistic graphical models: principles and techniques.
MIT press.


Pearl, J. (2009).
Causality.
Cambridge university press.


Ruozzi, N.
Cs 6347: Statistical methods in artificial intelligence and machine learning.


Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.