CPSC 440: Advanced Machine Learning

More DAGs

Mark Schmidt

University of British Columbia

Winter 2021

Last Time: Directed Acyclic Graphical (DAG) Models

- DAG models use a factorization of the joint distribution,

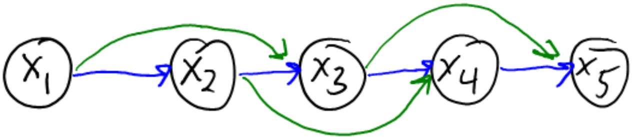
$$p(x_1, x_2, \dots, x_d) = \prod_{j=1}^{d} p(x_j | x_{\mathsf{pa}(j)}),$$

where $\mathsf{pa}(j)$ are called the "parents" of feature j.
  - We are using the order $1{:}d$, but note that you could use any order.

- This assumes a Markov property (generalizing Markov property in chains),

$$p(x_j | x_{1:j-1}) = p(x_j | x_{\mathsf{pa}(j)}),$$

- We visualize the assumptions made by the model as a graph:



Graph Structure Examples

- Instead of factorizing by variables j, could factor into blocks b:

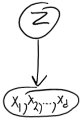
$$p(x) = \prod_b p(x_b \mid x_{\mathsf{pa}(b)}),$$

  and have the nodes be blocks.
  - Usually assuming full connectivity within the block.

- With mixture of Gaussian and full covariances we have

$$p(z, x) = p(z)p(x \mid z).$$

- The corresponding graph structure is:



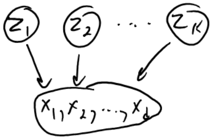
- Gaussian generative classifiers (GDA) have the same structure.
  - But using class lable y instead of cluster z.

Graph Structure Examples

With probabilistic PCA we have

$$p(z, x) = p(x \mid z) \prod_{c=1}^{k} p(z_c).$$

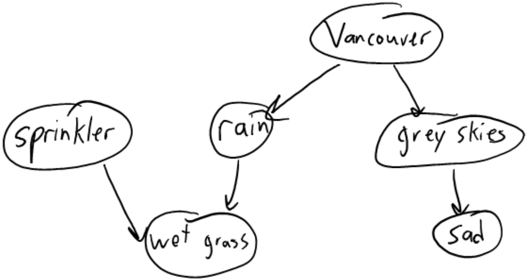
The corresponding graph structure is:



The data x comes from a set of independent parents (latent factors).

Graph Structure Examples

We can consider less-structured examples,

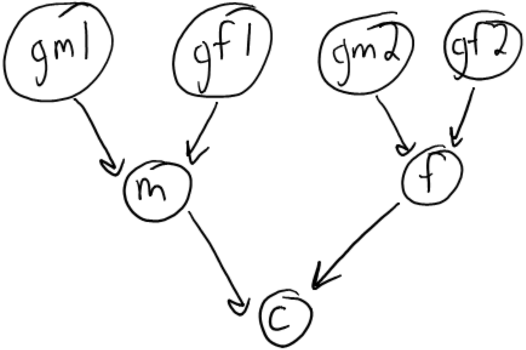


The corresponding factorization is:

$$p(S, V, R, W, G, D) = p(S)p(V)p(R \mid V)p(W \mid S, R)p(G \mid V)p(D \mid G).$$

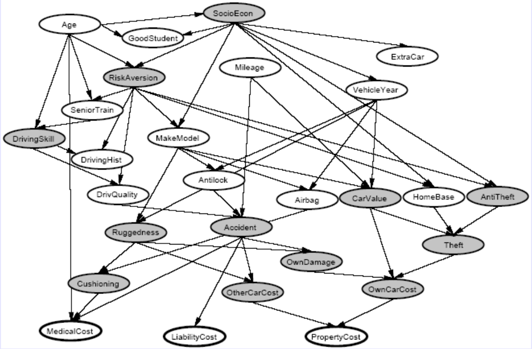
Graph Structure Examples

We can consider genetic phylogeny (family trees):



Example: Vehicle Insurance

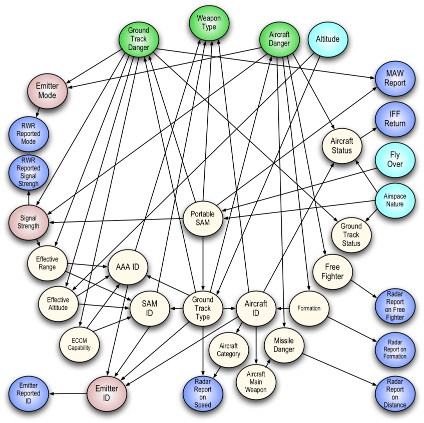
- Want to predict bottom three "cost" variables, given observed and unobserved values:



https://www.cs.princeton.edu/courses/archive/fall10/cos402/assignments/bayes

Example: Radar and Aircraft Control

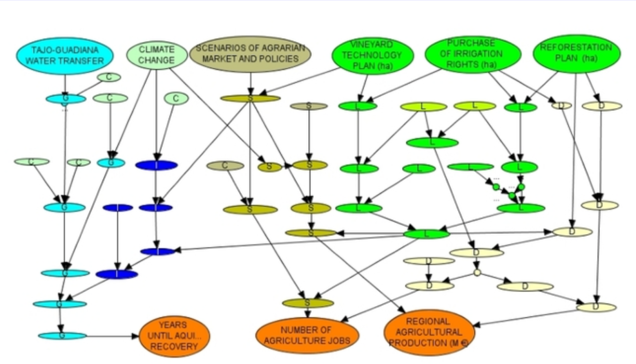
- Modeling multiple planes and radar signals:



https://pr-owl.org/basics/bn.php

Example: Water Resource Management

- Dependencies in environmental monitor and susatainability issues:



https://www.jstor.org/stable/26268156

Beware of the "Causal" DAG

- It can helpful to use the language of causality when reasoning about DAGs.
  - You'll find that they give the correct causal interpretation based on our intuition.
- However, keep in mind that the arrows are not necessarily causal.
  - "A causes B" has the same graph as "B causes A".
- There is work on causal DAGs which add semantics to deal with "interventions".
  - But these require extra assumptions: fitting a DAG to observational data doesn't imply anything about causality.

Outline

1 Conditional Independence

2 D-Separation

Review of Independence

- Let A and B are random variables taking values $a \in \mathcal{A}$ and $b \in \mathcal{B}$.
- We say that A and B are independent if we have

$$p(a, b) = p(a)p(b),$$

for all a and b.

- To denote independence of x_i and x_j we use the notation

$$x_i \perp x_j.$$

- In a product of Bernoullis, we assume $x_i \perp x_j$ for all i and j.

Review of Independence

- For independent a and b we have

$$p(a \mid b) = \frac{p(a,b)}{p(b)} = \frac{p(a)p(b)}{p(b)} = p(a).$$

- This gives us a more intuitive definition: A and B are independent if

$$p(a \mid b) = p(a)$$

  for all a and all b where $p(b) \neq 0$.
  - In words: knowing b tells us nothing about a (and vice versa).
    - This will tend to simplify calculations involving a.

- Useful fact: $a \perp b$ iff $p(a,b) = f(a)g(b)$ for some functions f and g.

Conditional Independence

- We say that A is conditionally independent of B given C if

$$p(a,b \mid c) = p(a \mid c)p(b \mid c),$$

  for all a, b, and $c \neq 0$.
- Equivalently, we have

$$p(a \mid b,c) = p(a \mid c), \quad \text{or} \quad p(b \mid a,c) = p(b \mid c).$$

- "If you know C, then *also* knowing B would tell you nothing about A"'.
  - In mixture of Bernoullis, given cluster there is no dependence between variables.

- We often write this as

$$A \perp B \mid C.$$

- In a naive Bayes, we assume $x_i \perp x_j \mid y$ for all i and j.
  - This simplifies calculations involving x_i and x_j, provided that we know y.

Extra Conditional Independences in Markov Chains

- In Markov chains, the Markov assumption is $x_j \perp x_1, x_2, \ldots, x_{j-2} \mid x_{j-1}$,

$$p(x_j \mid x_{j-1}, x_{j-2}, \ldots, x_1) = p(x_j \mid x_{j-1}).$$

- But note that this also implies the additional conditional independence that

$$p(x_j \mid x_{j-2}, x_{j-3}, \ldots, x_1) = p(x_j \mid x_{j-2}).$$

- We can use this property to easily compute $p(x_j \mid x_{j-2}, x_{j-3}, \ldots, x_1)$:

$$\begin{aligned}
p(x_j \mid x_{j-2}, x_{j-3}, \ldots x_1) &= p(x_j \mid x_{j-2}) \\
&= \sum_{x_{j-1}} p(x_j, x_{j-1} \mid x_{j-2}) \\
&= \sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2}) p(x_{j-1} \mid x_{j-2}) \\
&= \sum_{x_{j-1}} \underbrace{p(x_j \mid x_{j-1})}_{\text{tran prob}} \underbrace{p(x_{j-1} \mid x_{j-2})}_{\text{tran prob}}.
\end{aligned}$$

Extra Conditional Independences in Markov Chains

- Proof that x_j is independent of $\{x_1, x_2, \dots, x_{j-3}\}$ given x_{j-2}:

$$
\begin{aligned}
p(x_j \mid x_{j-2}, x_{j-3}, \dots, x_1) &= \frac{p(x_j, x_{j-2}, x_{j-3}, \dots, x_1)}{p(x_{j-2}, x_{j-3}, \dots, x_1)} \quad \text{(def'n cond. prob.)}\\
&= \frac{\sum_{x_{j-1}} p(x_j, x_{j-1}, x_{j-2}, \dots, x_1)}{p(x_{j-2} \mid x_{j-3}, x_{j-4}, \dots, x_1)p(x_{j-3} \mid x_{j-4}, x_{j-5}, \dots, x_1) \cdots p(x_1)} \quad \text{(marg. and chain rule)}\\
&= \frac{\sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2})p(x_{j-1} \mid x_{j-2}) \dots p(x_2 \mid x_1)p(x_1)}{p(x_{j-2} \mid x_{j-3})p(x_{j-3} \mid x_{j-4}) \cdots p(x_1)} \quad \text{(chain rule and Markov)}\\
&= \frac{p(x_1)p(x_2 \mid x_1) \cdots p(x_{j-2} \mid x_{j-3}) \sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2})p(x_{j-1} \mid x_{j-2})}{p(x_{j-2} \mid x_{j-3})p(x_{j-3} \mid x_{j-4}) \cdots p(x_1)} \quad \text{(take terms outside}\\
&= \sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2})p(x_{j-1} \mid x_{j-2}) \quad \text{(cancel out in numerator/denominator)}\\
&= \sum_{x_{j-1}} p(x_j, x_{j-1} \mid x_{j-2}) \quad \text{(product rule)}\\
&= p(x_j \mid x_{j-2}) \quad \text{(marg rule).}
\end{aligned}
$$

- Similar steps could be used to show $x_j \perp x_{j+2} \mid x_{j+1}$, and a variety of other conditional independences like $x_1 \perp x_{10} \mid x_5$.

DAGs and Conditional Independence

- So conditional independences can substantially simplify inference.
- But it's tedious to formally show that conditional independences hold.
  - See the last slide, and the EM notes.

- In DAGs we make the conditional independence assumption that

$$p(x_j \mid x_{j-1}, x_{j-2}, \ldots, x_1) = p(x_j \mid x_{\mathsf{pa}}(j)).$$

- Is there an easy way to find out what other independences are true?
  - If so, we could quickly find out which calculations are easy to do in a given DAG.

Outline

1 Conditional Independence

2 D-Separation

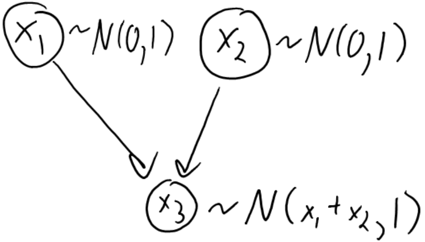
D-Separation: From Graphs to Conditional Independence

- All conditional independences implied by a DAG can be read from the graph.
- In particular: variables A and B are conditionally independent given C if:
  - "D-separation blocks all undirected paths in the graph from any variable in A to any variable in B."
- In the special case of product of independent models our graph is:

- Here there are no paths to block, which implies the variables are independent.
- Checking paths in a graph tends to be faster than tedious calculations.
  - We can start connecting properties of graphs to computational complexity.

D-Separation as Genetic Inheritance

- The rules of d-separation are intuitive in a simple model of gene inheritance:
  - Each node/person has single number, which we'll call a "gene".
  - If you have no parents, your gene is a random number.
  - If you have parents, your gene is a sum of your parents plus noise.
- For example, think of something like this:



- Graph corresponds to the factorization $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$.
  - In this model, does $p(x_1, x_2) = p(x_1)p(x_2)$? (Are x_1 and x_2 independent ?)

D-Separation as Genetic Inheritance

- Genes of people are independent if knowing one says nothing about the other.
- Your gene is dependent on your parents:
  - If I know your parent's gene, I know something about yours.
- Your gene is independent of your (unrelated) friends:
  - If you know your friend's gene, it doesn't tell me anything about you.
- Genes of people can be conditionally independent given a third person:
  - Knowing your grandparent's gene tells you something about your gene.
  - But grandparent's gene isn't useful if you know parent's gene.

D-Separation Case 0 (No Paths and Direct Links)

Are genes in person x independent of the genes in person y?

- No path: x and y are not related (independent),

  We have $x \perp y$: there are no paths to be blocked.

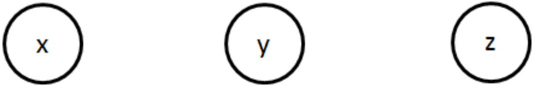
- Direct link: x is the parent of y,

  We have $x \not\perp y$: knowing x tells you about y (direct paths aren't blockable).

D-Separation Case 0 (No Paths and Direct Links)

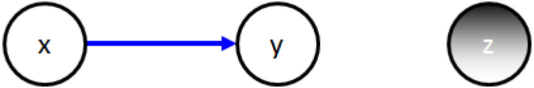
Neither case changes if we have a third independent person z:

- No path: If x and y are independent,



  We have $x \perp y$: adding z doesn't make a path.

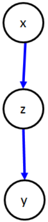
- Direct link: x is the parent of y,



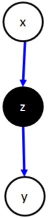
  We have $x \not\perp y \mid z$: adding z doesn't block path.
  - We use **black or shaded** nodes to denote values we condition on (in this case z).
    - We sometimes also call the nodes that we condition on the "observations".

D-Separation Case 1: Chain

- Case 1: x is the grandparent of y.
  - If z is the mother we have:



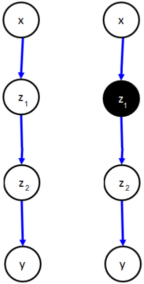
We have $x \not\perp y$: knowing x would give information about y because of z
  - But if z is *observed*:



In this case $x \perp y \mid z$: knowing z "breaks" dependence between x and y.

D-Separation Case 1: Chain

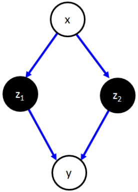
- The same logic holds for great-grandparents:



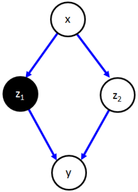
- We have $x \not\perp y$ (left), but $x \perp y \mid z_1$ (right).
  - We also have $x \perp y \mid z_2$ and that $x \perp y \mid z_1, z_2$.
- This case lets you test any independence in Markov chains.
  - "Do observe any value in between the two nodes?"

D-Separation Case 1: Chain

- Consider weird case where parents z_1 and z_2 share parent x:
  - If z_1 and z_2 are observed we have:



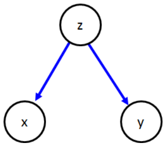
    We have $x \perp y \mid z_1, z_2$: knowing both parents breaks dependency.
  - But if only z_1 is *observed*:



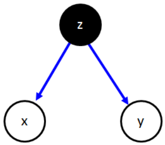
    We have $x \not\perp y \mid z_1$: dependence still "flows" through z_2.

D-Separation Case 2: Common Parent

- Case 2: x and y are sibilings.
  - If z is a common unobserved parent:



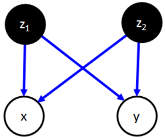
    We have $x \not\perp y$: knowing x would give information about y.
  - But if z is *observed*:



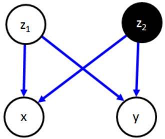
    In this case $x \perp y \mid z$: knowing z "breaks" dependence between x and y.
- This is type of independence used in naive Bayes and "mixture of independent".

D-Separation Case 2: Common Parent

- Case 2: x and y are sibilings.
  - If z_1 and z_2 are common observed parents:



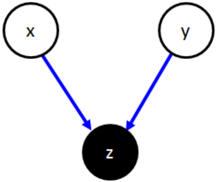
We have $x \perp y \mid z_1, z_2$: knowing z_1 and z_2 breaks dependence between x and y.
  - But if we only observe z_2:



Then we have $x \not\perp y \mid z_2$: dependence still "flows" through z_1.

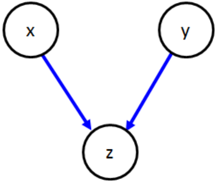
D-Separation Case 3: Common Child

- Case 3: x and y share a child z:
  - If we observe z then we have:



We have $x \not\perp y \mid z$: if we know z, then knowing x gives us information about y.

  - But if z is not observed:



We have $x \perp y$: if you don't observe z then x and y are independent.

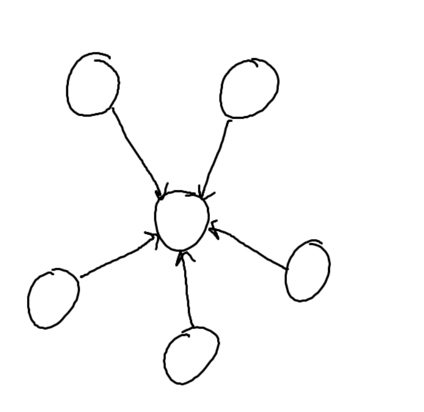
- Different from Case 1 and Case 2: not observing the child blocks path.

Summary

- Joint distribution of models we've discussed can be written as DAG models.
- Conditional independence of A and B given C:
  - Knowing B tells us nothing about A if we already know C.
- D-separation allows us to test conditional independences based on graph.
- Next time: the IID assumption as a DAG?

Conditional Independence in Star Graphs

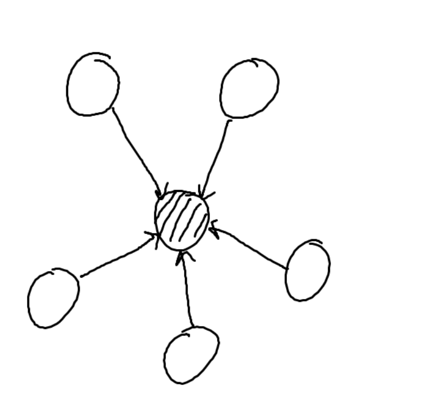
- Consider the following star graph:



- “5 aliens get together and make a baby alien”.
  - Unconditionally, the 5 aliens are independent.

Conditional Independence in Star Graphs

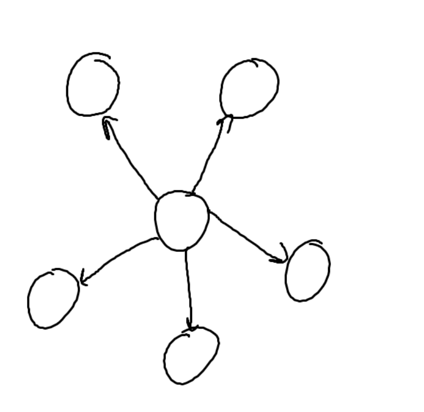
- Consider the following star graph:



- "5 aliens get together and make a baby alien".
  - Conditioned on the baby, the 5 aliens are dependent.

Conditional Independence in Star Graphs

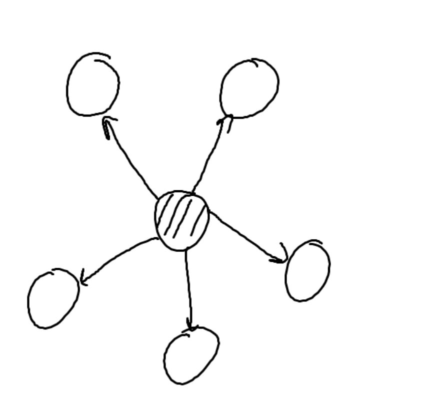
- Consider the following star graph:



- "An organism produces 5 clones".
  - Unconditionally, the 5 clones are dependent.

Conditional Independence in Star Graphs

- Consider the following star graph:



- "An organism produces 5 clones".
  - Conditioned on the original, the 5 clones are independent.