DATA ANALYTICS

Unit 5: Advanced Techniques

Swati Pratap Jagdale
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5: Concept of hidden variables

Swati Pratap Jagdale

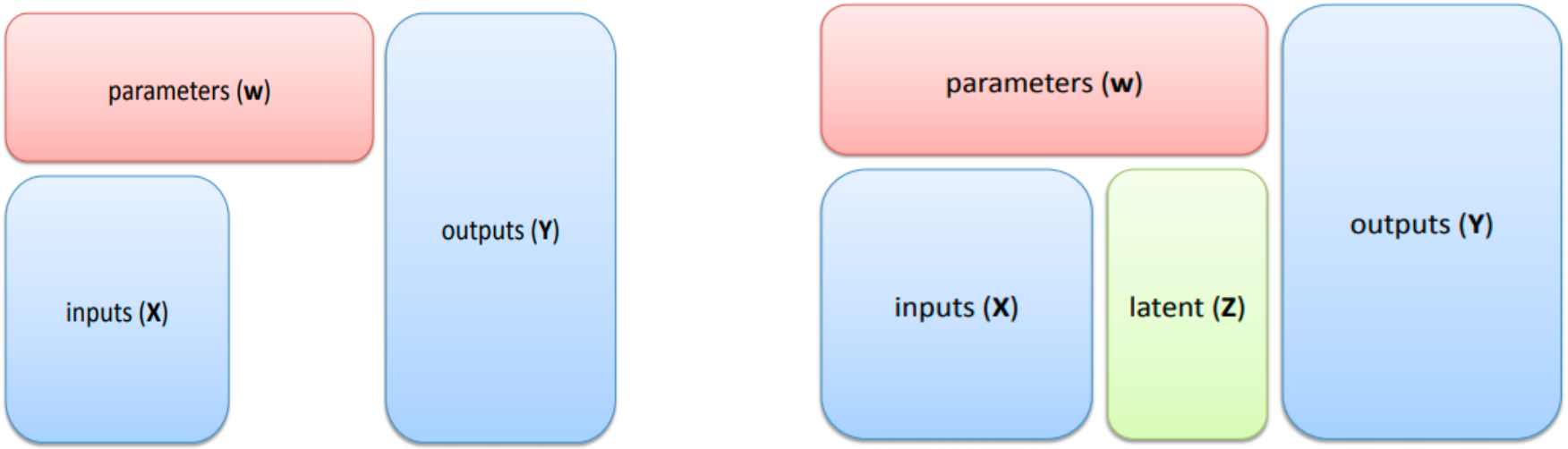
Department of Computer Science and Engineering

Hidden Variables

- Real world problem have hidden variables
- No training data available on hidden variables
- Model cannot be built without training data
- Inference cannot be done without model
- How to learn models with hidden variables

Hidden Variables

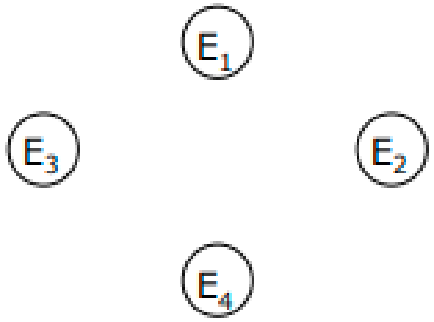
- Random variables in supervised learning



- We use the term “hidden variable” (or “latent variable”) to refer to something we never see.
  - Not even in training.
  - Sometimes we believe they are real.
  - Sometimes we believe they only approximate reality.

Learning With Hidden Variables - Bayesian networks with hidden variables

- Consider a situation in which you can observe a whole bunch of different evidence variables, E1 through En. Maybe they're all the different symptoms that a patient might have. Or maybe they represent different movies and whether someone likes them.

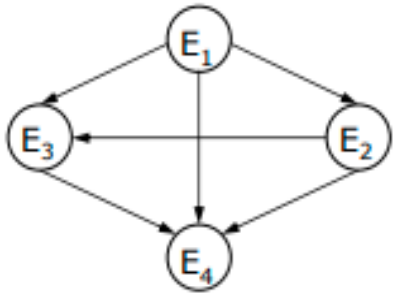


Learning With Hidden Variables

If those variables are all conditionally dependent on one another, then we'd need a highly connected graph that's capable of representing the entire joint distribution between the variables.

Because the last node has n-1 parents, it will take on the order of 2^n parameters to specify the conditional probability tables in this network.

Without the cause, all the evidence is dependent on each other

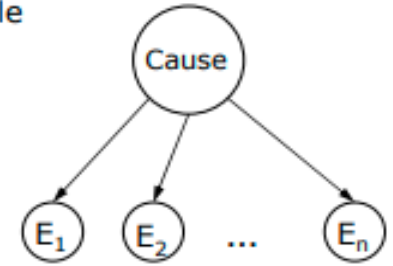


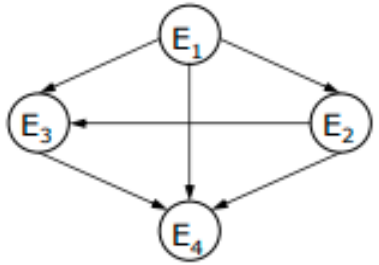
$O(2^n)$ parameters

Learning With Hidden Variables

But, in some cases, we can get a considerably simpler model by introducing an additional “cause” node.
It might represent the underlying disease state that was causing the patients’ symptoms or some division of people into those who like westerns and those who like comedies.

Cause is unobservable





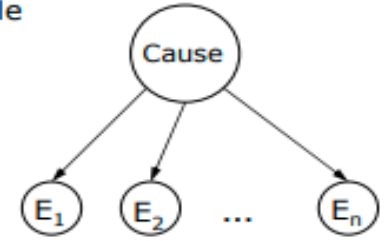
$O(2^n)$ parameters

Without the cause, all the evidence is dependent on each other

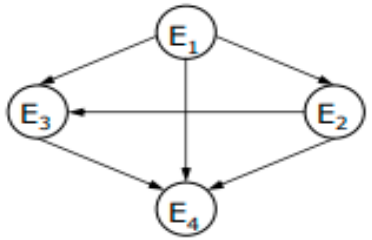
Learning With Hidden Variables

- In the simpler model, the evidence variables are conditionally independent given the causes. That means that it would only require on the order of n parameters to describe all the CPTs in the network, because at each node, we just need a table of size 2 (if the cause is binary; or k if the cause can take on k values), and one (or k-1) parameter to specify the probability of the cause.

Cause is unobservable



O(n) parameters



O(2^n) parameters

Without the cause, all the evidence is dependent on each other

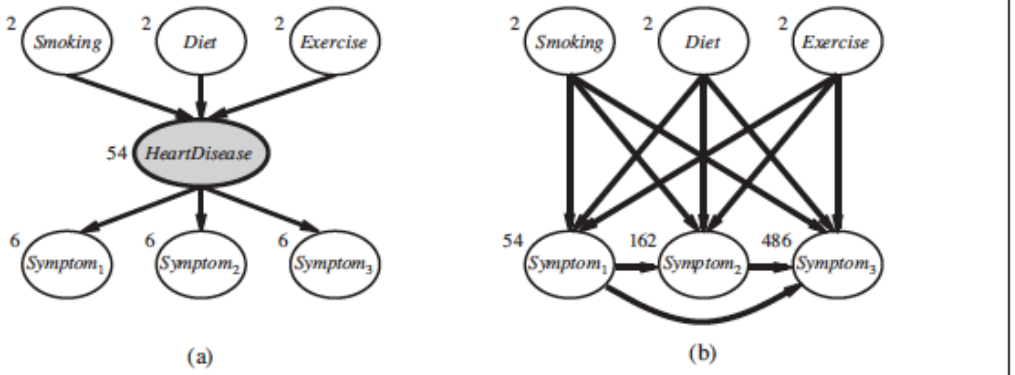


Figure 20.10 (a) A simple diagnostic network for heart disease, which is assumed to be a hidden variable. Each variable has three possible values and is labeled with the number of independent parameters in its conditional distribution; the total number is 78. (b) The equivalent network with *HeartDisease* removed. Note that the symptom variables are no longer conditionally independent given their parents. This network requires 708 parameters.

Learning With Hidden Variables

- In the simpler model, the evidence variables are conditionally independent given the causes. That means that it would only require on the order of n parameters to describe all the CPTs in the network, because at each node, we just need a table of size 2 (if the cause is binary; or k if the cause can take on k values), and one (or k-1) parameter to specify the probability of the cause.

- So, what if you think there's a hidden cause? How can you learn a network with unobservable variables?

Simpson's Paradox

- Edward Hugh Simpson, a statistician and former cryptanalyst at Bletchley Park, described the statistical phenomenon - Simpson's paradox

- The art of data science is seeing beyond the data — using and developing methods and tools to get an idea of what that hidden reality looks like.

- Simpson's paradox showcases the importance of skepticism and interpreting data with respect to the real world, and also the dangers of oversimplifying a more complex truth by trying to see the whole story from a single data-viewpoint.

Simpson's Paradox

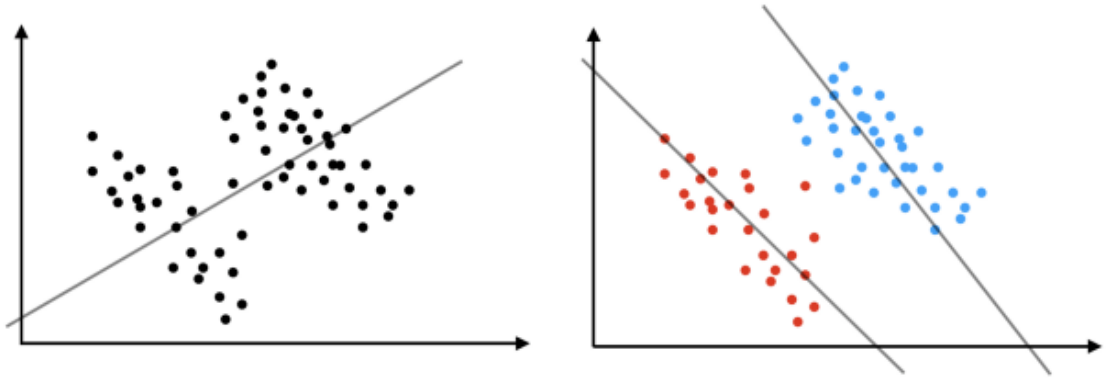
- ***Simpson's Paradox:***

A trend or result that is present when data is put into groups that reverses or disappears when the data is combined.

Example:
UC Berkley's suspected gender-bias.
At the beginning of the academic year in 1973, UC Berkeley's graduate school had admitted roughly 44% of their male applicants and 35% of their female applicants.

- There was a statistically significant gender bias in favour of women for 4 out of the 6 departments, and no significant gender bias in the remaining 2.
- It is discovered that women tended to apply to departments that admitted a smaller percentage of applicants overall, and that this hidden variable affected the marginal values for the percentage of accepted applicants in such a way as to reverse the trend that existed in the data as a whole.

Simpson's Paradox



A visual example: the overall trend reverses when data is grouped by some colour-represented category.

Simpson's Paradox

a simple example in business:

- Suppose we're in the soft drinks industry and we're trying to choose between two new flavours we've produced. We could sample public opinion on the two flavours

Flavour	Sample Size	# Liked Flavour
Sinful Strawberry	1000	800
Passionate Peach	1000	750

- 80% of people enjoyed 'Sinful Strawberry' whereas only 75% of people enjoyed 'Passionate Peach'. So 'Sinful Strawberry' is more likely to be the preferred flavour.

Simpson's Paradox

- Some other information while conducting the survey, such as the sex of the person sampling the drink. What happens if we split our data up by sex?
- 84.4% of men and 40% of women liked 'Sinful Strawberry' whereas 85.7% of men and 50% of women liked 'Passionate Peach'

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

Simpson's Paradox

- *according to our sample data, generally people prefer 'Sinful Strawberry', but both men and women separately prefer 'Passionate Peach'.*
- *This is an example of Simpson's Paradox!*

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

Simpson's Paradox

- **Lurking variables (Hidden Variables)**

- Simpson's paradox arises when there are hidden variables that split data into multiple separate distributions.
- Such a hidden variable is aptly referred to as a **lurking variable**, and they can often be difficult to identify.

- **Consider the lurking variable (sex) and a little bit of probability theory:-**

- P(Liked Strawberry) = P(Liked Strawberry | Man)P(Man) + P(Liked Strawberry | Woman)P(Woman)

- 800/1000 = (760/900)×(900/1000) + (40/100)×(100/1000)

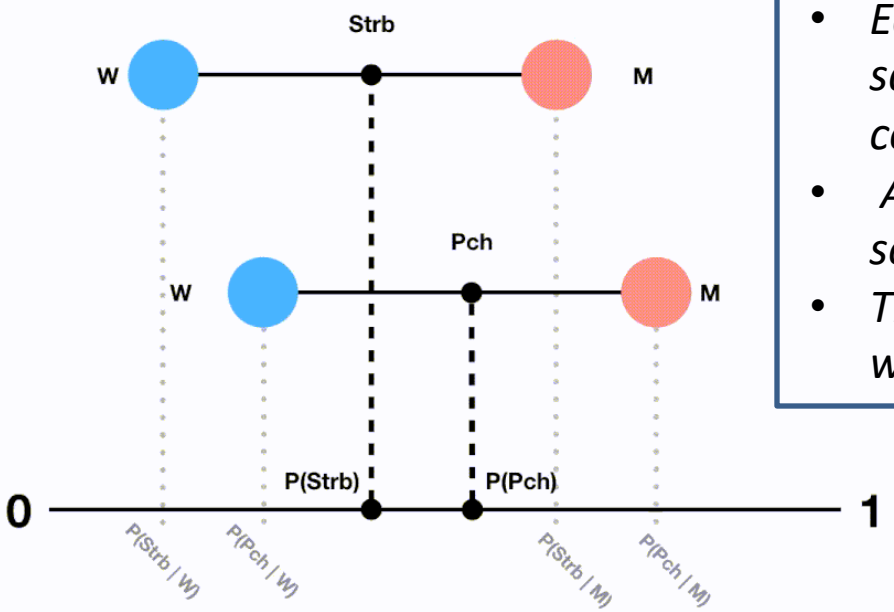
- P(Liked Peach) = P(Liked Peach | Man)P(Man) + P(Liked Peach | Woman)P(Woman)

- 750/1000 = (600/700)×(700/1000) + (150/300)×(300/1000)

Simpson's Paradox

- **Lurking variables (Hidden Variables)**

- We can think of the marginal probabilities of sex (P(Man) and P(Woman)) as weights that, in the case of 'Sinful Strawberry', cause the total probability to be significantly shifted towards the male opinion.



- *Each coloured circle represents either the men or women that sampled each flavour, the position of the centre of each circle corresponds to that group's probability of liking the flavour.*
- *As the circles grow (i.e. sample proportions change) we can see how the marginal probability of liking the flavour changes.*
- *The marginal distributions shift and switch as samples become weighted with respect to the lurking variable (sex).*

References

http://www.cs.cmu.edu/~nasmith/psnlp/lecture5.pdf

https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-825-techniques-in-artificial-intelligence-sma-5504-fall-2002/lecture-notes/Lecture18FinalPart1.pdf

https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765

Russell, Stuart J. (Stuart Jonathan). Artificial intelligence : a modern approach

THANK YOU

Swati Pratap Jagdale
Department of Computer Science
swatigambhire@pes.edu