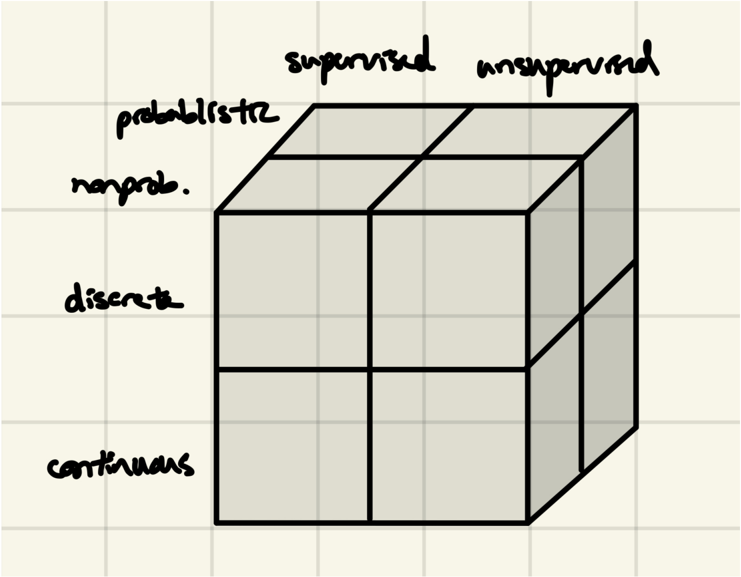
CS 181 LECTURE 3/28/24

Scribe Notes

Contents

1 Where We Are in the Course

We have finished up the cube!



So, now we will be thinking about the structure in distributions. In this part of the semester, when it comes to the probabilistic models, we have been writing down distributions. We have implicitly been saying that probabilistic unsupervised models is about making a distribution. Now, we are thinking about what the structure of that distribution is. Today's content can be reviewed in Chapter 8 of the textbook.

After that, we will talk about decision making.

2 Story of the Day

Does smoking cause lung cancer?

Should big tobacco be held accountable for the health risks that they are potentially posing on the population?

If you think about that question, the argument that various scientists were making is:

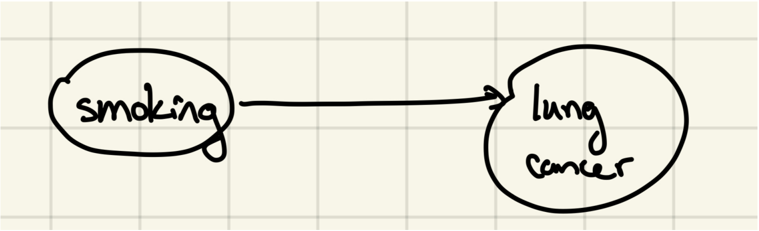


Figure 1: Scientists' argument.

"Smoking" is a variable that has impact on the variable "lung cancer".

The scientists argue that there is a causal link so companies are in the wrong for advertising smoking.

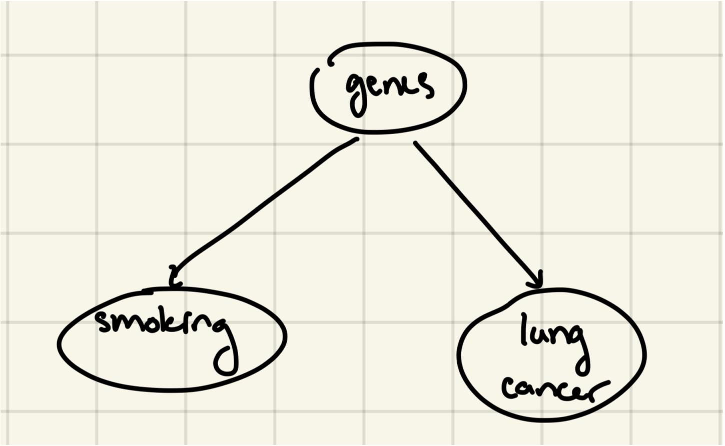


Figure 2: Companies' argument.

But the companies argue something different.

The implication of this graph is that genes cause lung cancer so if you stop someone from smoking, that would not prevent someone from getting lung cancer.

3 Idea of Independence

Let us think about why drawing these graphical models are helpful.

3.1 Mixture Models

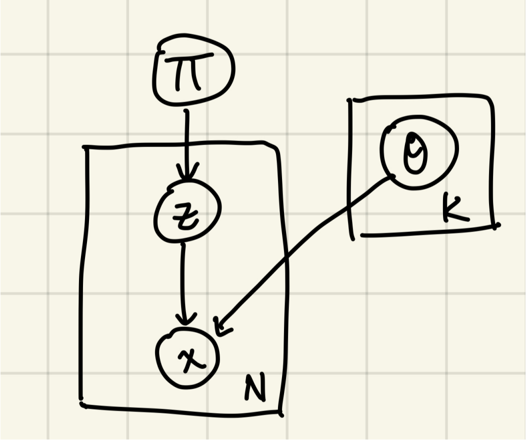


Figure 3: Mixture model.

Here we used plate notation to show that there are multiple copies of the z and x pairs.

$$z_n \sim Cat(\pi)$$
$$x_n \sim N(\mu_{z_n}, \sigma^2 z_n)$$

Recall that in Factor Analysis, we had:

$$z_n \sim N(0, \mathbf{I}_k \sigma^2)$$
$$x_n \sim N(W z_n, \Sigma)$$

There is something that generates the local variables and there is something global. If we look at the structure of the distribution, we see that the structure is the same - even when the specific distributions that we use are different.

To do inference, we used EM. The core ideal behind using EM is that we have two phases: 1) we did inference on local variables z_n, 2) we did inference on the global parameters θ.

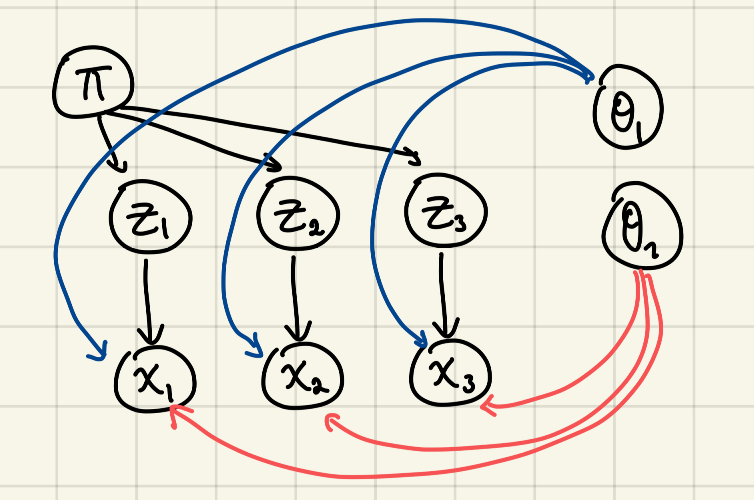


Figure 4: Mixture model, with three variables.

In Figure 4, we do not use the plate notation, so we have shown the relationships between the z's and n's for three variables.

By fixing the z's, π, θ_1 andθ_2 become independent. This is why we can do our two step inference. Our inferences result because fixing certain variables lets us use the independence of other variables.

3.2 Bayesian Networks

A Bayesian netowrk is a directed acyclic graph (DAG) that defines a joint distribution.

For example, consider the following

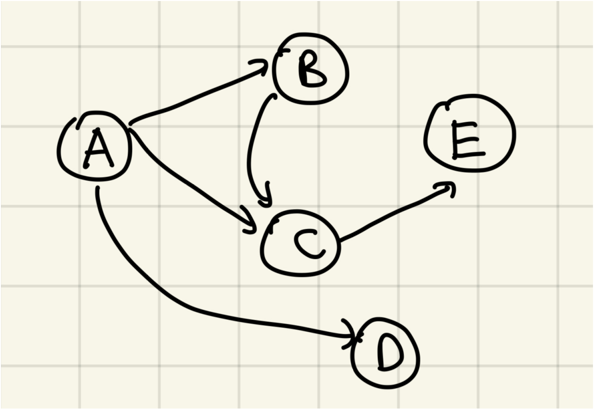


Figure 5: Example DAG.

The graph has directed edges and no cycles, so it is a DAG. In this graph, we can say that A is independent of the other variables. So, we can write the probability of all the variables as:

$$p(A, B, C, D, E) = p(A)p(D|A)p(B|A)p(C|A, B)p(E|C)$$

We can always factorize the entirety.

$$p(A, B, C, D, E) = p(A)p(B|A)p(C|A, B)p(D|A, B, C)p(E|A, B, C, D)$$

But we see from our first equation that we can actually simplify this factorization. For example, for E we only actually need to consider C instead of A, B, C, D.

3.3 D-separation

A and B are considered d-separated if every undirected path between them is blocked.

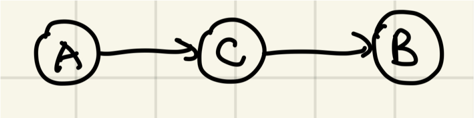


Figure 6: Path between A and B, C blocks the path.

In this graph, there is a path between A and B. If C is observed, then the path is blocked. Let us consider the analogy:

A = Raining outside C = the clothes are wet B = Leaving puddles

If we know that our clothes are wet, then the event of rain outside and leaving puddles become independent.

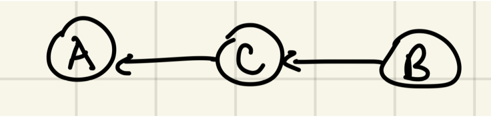


Figure 7: Path between A and B, C blocks the path.

The same thing happens again in Figure 7, where observing the middle event in the path leads to independence between A and B.

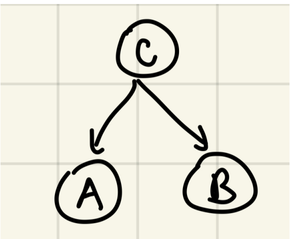


Figure 8: Path between A and B, C blocks the path.

The same thing happens again in Figure 8, where observing the middle event in the path leads to independence between A and B.

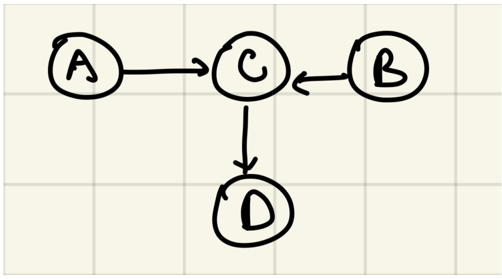


Figure 9: Not observing C blocks the path.

In Figure 9, not observing C (or its descendents) blocks the path. For example,

A = how much you study B = how many pre-requisites C = a student doing well in a class

Blocking only happens if you do not observe C. Here, the observation creates dependence. Now, suppose we have the following graph:

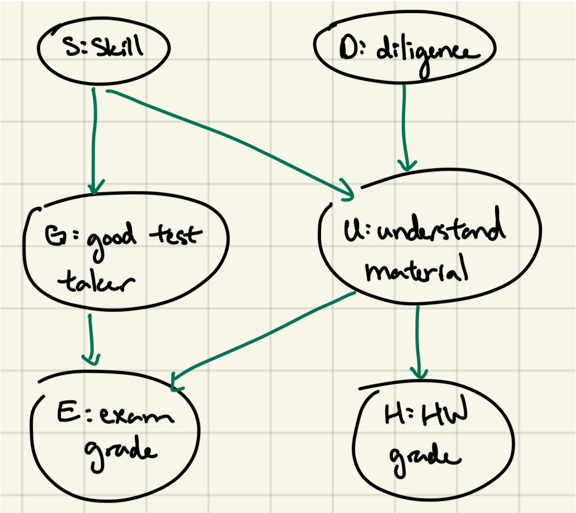


Figure 10: Real-world example.

Given G, U: is E independent of H? of S?

So we have observed that the person is a good test taker and understands the material. We now want to know if $p(E|G, U)$ is equal to $p(E|G, U, H)$. Let us think about the paths from E to H. The path from H to U to E is blocked because we have observed U. Therefore, E and H are independent.

Similarly, we can observe a path from S to G to E. And a path from S to U to E. If G and U are given, then every undirected path is blocked, so E is independent of S.

Given E, are G and U independent? No, because we get the last scenario where not observing the middle event blocks the path.

Therefore,

$$p(S, D.G, U, E, H) = p(S)p(D)p(G|S)p(U|S, D)p(E|G, U)p(H|U)$$

But we could also write the factorization differently. For example,

$$p(E)p(H|E)p(U|E, H)p(S|U, E)p(G|U, S, E)p(U, S)$$

Now, if we build a graph in this ordering, then we get the following

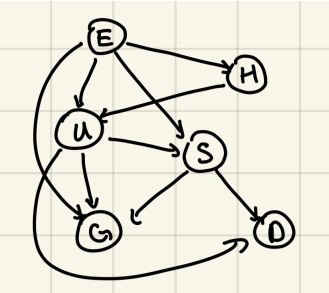


Figure 11: Graph based on re-ordering.

So, we see that we have more parameters than the optimal factorization. Based on fig

10, Since G depends on S, then we have two parameters. Since U depends on S and D, we have 4 parameters. E has two parents, so it also has 4 parameters. H has one parent so it has one parameter. S and D each need one parameter. So there is a total of 14 parameters.

Now, looking at fig 11, we get a total of 23 parameters. E needs 1 parameter, U needs 4, H needs 2, G needs 8, S needs 4, and D needs 4.

Thus, we see that while the relationships we have found in fig 11 are not incorrect, using this model would make learning take a longer time because there are more parameters to learn. Our goal is then to make a simplified graph of any possible model.

4 Aside: Two Other Common Graphical Models

4.1 Undirected Graph

This can help describe certain things more easily, like in the field of vision it has been used to describe pixels. But the problem with undirected graphs is that it is not going to be normalized.

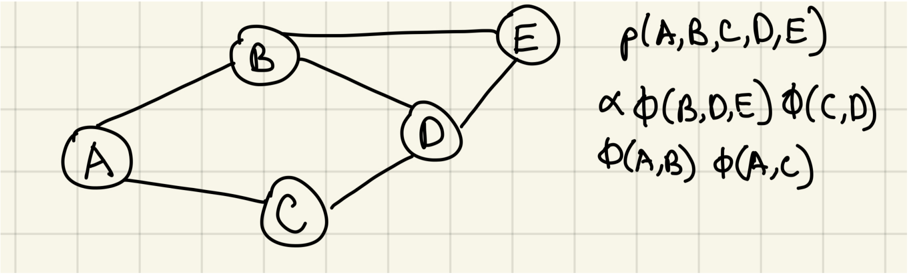


Figure 12: Undirected Graph.

4.2 Factor Graphs

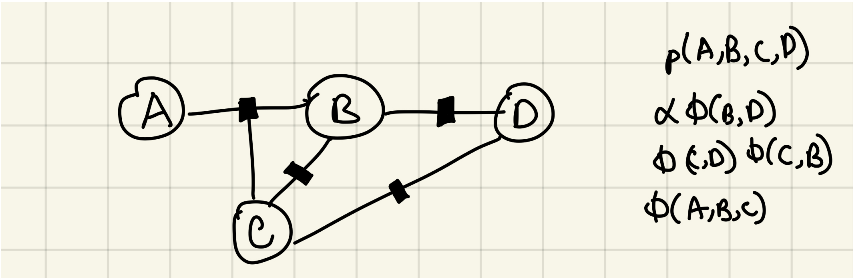


Figure 13: Factor Graph.

These two types of graphs are good to know but not necessary for 181.

5 Concept Check

How many parameters are needed to describe the model if:

1. All the variables are discrete

2. The variables are linear Gaussian, 1D vectors: $x_A \sim N(\mu_A, \sigma^2), x_B \sim N(w_A x_A, \sigma^2)$
3. The variables are linear Gaussian, 4D vectors: $x_A \sim N(\mu_A, I\sigma^2), x_B \sim N(w_A x_A, I\sigma^2)$

6 Concept Check Solutions

1. A will have 3 parameters. B will have 12 parameters. C will have 12 parameters. D will have 48 parameters. E will have 12 parameters. Summing this all up gives us 87 parameters.
2. A needs one parameter to describe its mean. D needs two parameters because it is a linear combination of B and C. Including the other variables, we have a total of 5 parameters.
3. There are 84 parameters total.

What is going on? In the two second cases, we are making a linear assumption. But with the discrete setting, we made no assumptions, so the discrete setting may seem more complicated, giving us a greater number of parameters ($87 > 84$).