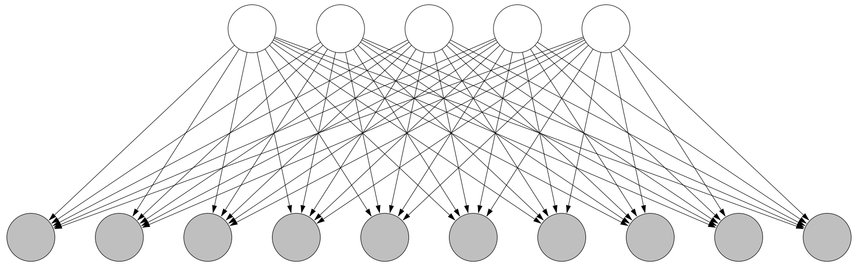
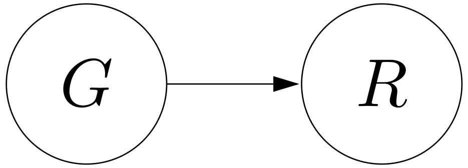
Bayesian networks: smoothing



- In this module, I'll talk about how Laplace smoothing for guarding against overfitting.

Review: maximum likelihood



$$\mathbb{P}(G = g, R = r) = p_G(g)p_R(r \mid g)$$

$$\mathcal{D}_{\text{train}} = \{(\mathsf{d}, 4), (\mathsf{d}, 4), (\mathsf{d}, 5), (\mathsf{c}, 1), (\mathsf{c}, 5)\}$$

θ:

g	$\text{count}_G(g)$	$p_G(g)$
d	3	3/5
c	2	2/5

g	r	$\text{count}_R(g, r)$	$p_R(r \mid g)$
d	4	2	2/3
d	5	1	1/3
c	1	1	1/2
c	5	1	1/2

Do we really believe that $p_R(r = 2 \mid g = \mathsf{c}) = 0$?

Overfitting!

- Suppose we have a two-variable Bayesian network whose parameters (local conditional distributions) we don't know.
- Instead, we obtain training data, where each example includes a full assignment.
- Recall that maximum likelihood estimation in a Bayesian network is given by a simple count + normalize algorithm.
- But is this a reasonable thing to do? Consider the probability of a 2 rating given comedy? It's hard to believe that there is zero chance of this happening. That would be very closed-minded.
- This is a case where maximum likelihood has overfit to the training data!

Laplace smoothing example

Idea: just add $\lambda = 1$ to each count

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 4), (\text{d}, 5), (\text{c}, 1), (\text{c}, 5)\}$$

θ:

g	$\text{count}_G(g)$	$p_G(g)$
d	1+3	4/7
c	1+2	3/7

g	r	$\text{count}_R(g, r)$	$p_R(g, r)$
d	1	1	1/8
d	2	1	1/8
d	3	1	1/8
d	4	1+2	3/8
d	5	1+1	2/8
c	1	1+1	2/7
c	2	1	1/7
c	3	1	1/7
c	4	1	1/7
c	5	1+1	2/7

Now $p_R(r = 2 \mid g = \text{c}) = \frac{1}{7} > 0$

- There is a very simple patch to this form of overfitting called **Laplace smoothing**: just add some small constant λ (called a **pseudocount** or virtual count) for each possible value, regardless of whether it was observed or not.
- As a concrete example, let's revisit the two-variable model from before.
- We preload all the counts (now we have to write down all the possible assignments to g and r) with λ. Then we add the counts from the training data and normalize all the counts.
- Note that many values which were never observed in the data have positive probability as desired.

Laplace smoothing

Key idea: maximum likelihood with Laplace smoothing

For each distribution d and partial assignment $(x_{\text{Parents}(i)}, x_i)$:

Add λ to $\text{count}_d(x_{\text{Parents}(i)}, x_i)$.

Further increment counts $\{\text{count}_d\}$ based on $\mathcal{D}_{\text{train}}$.

Hallucinate λ occurrences of each local assignment

- More formally, when we do maximum likelihood with Laplace smoothing with smoothing parameter $\lambda > 0$, we add λ to the count for each distribution d and local assignment $(x_{\text{Parents}(i)}, x_i)$. Then we increment the counts based on the training data $\mathcal{D}_{\text{train}}$.
- Advanced: Laplace smoothing can be interpreted as using a Dirichlet prior over probabilities and doing maximum a posteriori (MAP) estimation.

Interplay between smoothing and data

Larger $\lambda \Rightarrow$ more smoothing $\Rightarrow$ probabilities closer to uniform

g	$\text{count}_G(g)$	$p_G(g)$
d	$1/2$+1	$3/4$
c	$1/2$	$1/4$

g	$\text{count}_G(g)$	$p_G(g)$
d	1+1	$2/3$
c	1	$1/3$

Data wins out in the end (suppose only see $g =$ d):

g	$\text{count}_G(g)$	$p_G(g)$
d	1+1	$2/3$
c	1	$1/3$

g	$\text{count}_G(g)$	$p_G(g)$
d	1+998	0.999
c	1	0.001

- By varying λ, we can control how much we are smoothing. The larger the λ, the stronger the smoothing, and the closer the resulting probability estimates become to the uniform distribution.
- However, no matter what the value of λ is, as we get more and more data, the effect of λ will diminish. This is desirable, since if we have a lot of data, we should be able to trust our data more and more.

Summary

g	$\text{count}_G(g)$	$p_G(g)$
d	$\lambda + 1$	$\frac{1+\lambda}{1+2\lambda}$
c	λ	$\frac{\lambda}{1+2\lambda}$

- Pull distribution closer to uniform distribution

- Smoothing gets washed out with more data

- In conclusion, Laplace smoothing provides a simple way to avoid overfitting by adding a smoothing parameter λ to all the counts, pulling the final probability estimates away from any zeros and towards the uniform distribution.
- But with more amounts of data, then the effect of smoothing wanes.