CS 331: Artificial Intelligence
Fundamentals of Probability II

Thanks to Andrew Moore for some course material

Full Joint Probability Distributions

Toothache	**Cavity**	**Catch**	**P(Toothache, Cavity, Catch)**
false	false	false	0.576
false	false	true	0.144
false	true	false	0.008
false	true	true	0.072
true	false	false	0.064
true	false	true	0.016
true	true	false	0.012
true	true	true	0.108

"Catch" means the dentist's probe catches in my teeth

This cell means P(Toothache = true, Cavity = true, Catch = true) = 0.108

Full Joint Probability Distributions

Toothache	Cavity	Catch	P(Toothache, Cavity, Catch)
false	false	false	0.576
false	false	true	0.144
false	true	false	0.008
false	true	true	0.072
true	false	false	0.064
true	false	true	0.016
true	true	false	0.012
true	true	true	0.108

The probabilities in the last column sum to 1

3

Joint Probability Distribution

From the full joint probability distribution, we can calculate any probability involving the three random variables in this world e.g.

P(*Toothache* = *true* OR *Cavity* = true) =

P(*Toothache=true, Cavity=false, Catch=false*) +
P(*Toothache=true, Cavity=false, Catch=true*) +
P(*Toothache=false, Cavity=true, Catch=false*) +
P(*Toothache=false, Cavity=true, Catch=true*) +
P(*Toothache=true, Cavity=true, Catch=false*) +
P(*Toothache=true, Cavity = true, Catch=true*) +

= 0.064 + 0.016 + 0.008 + 0.072 + 0.012 + 0.108 = 0.28

Marginalization

We can even calculate marginal probabilities (the probability distribution over a subset of the variables) e.g:

P(*Toothache=true, Cavity=true*) =

P(*Toothache=true, Cavity=true, Catch=true*) +

P(*Toothache=true, Cavity=true, Catch=false*)

= 0.108 + 0.012 = 0.12

Marginalization

Or even:

P(*Cavity=true*) =
P(*Toothache=true*, *Cavity=true*, *Catch=true*) +
P(*Toothache=true*, *Cavity=true*, *Catch=false*) +
P(*Toothache=false*, *Cavity=true*, *Catch=true*) +
P(*Toothache=false*, *Cavity=true*, *Catch=false*)

= 0.108 + 0.012 + 0.072 + 0.008 = 0.2

Marginalization

The general marginalization rule for any **sets** of variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$:

$$\boldsymbol{P}(\boldsymbol{Y}) = \sum_{\mathbf{z}} \boldsymbol{P}(\boldsymbol{Y}, \mathbf{z})$$

z is over all possible combinations of values of $\boldsymbol{Z}$ (remember $\boldsymbol{Z}$ is a set)

or

$$\boldsymbol{P}(\boldsymbol{Y}) = \sum_{\mathbf{z}} \boldsymbol{P}(\boldsymbol{Y} \mid \mathbf{z}) P(\mathbf{z})$$

7

Marginalization

For continuous variables, marginalization involves taking the integral:

$$\boldsymbol{P}(\boldsymbol{Y}) = \int \boldsymbol{P}(\boldsymbol{Y}, \mathbf{z}) d\mathbf{z}$$

8

Normalization

$$P(Cavity = true \mid Toothache = true)$$
$$= \frac{P(Cavity = true, Toothache = true)}{P(Toothache = true)}$$
$$= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

Note that 1/P(*Toothache*=*true*) remains constant in the two equations.

$$P(Cavity = false \mid Toothache = true)$$
$$= \frac{P(Cavity = false, Toothache = true)}{P(Toothache = true)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Normalization

- In fact, 1/P(*toothache*) can be viewed as a normalization constant for ***P***(*Cavity* | *toothache*), ensuring it adds up to 1
- We will refer to normalization constants with the symbol α

$$\mathbf{P}(Cavity \mid toothache) = \alpha \mathbf{P}(Cavity, toothache)$$

Inference

- Suppose you get a query such as P(*Cavity = true* | *Toothache = true*)

Toothache is called the evidence variable because we observe it. More generally, it's a set of variables.

Cavity is called the query variable (we'll assume it's a single variable for now)

There are also unobserved (aka hidden) variables like *Catch*

11

Inference

- We will write the query as $\mathbf{P}(X \mid \mathbf{e})$

This is a probability distribution hence the boldface

X = Query variable (a single variable for now)

$\mathbf{E}$ = Set of evidence variables

$\mathbf{e}$ = the set of observed values for the evidence variables

$\mathbf{Y}$ = Unobserved variables

12

Inference

We will write the query as $\mathbf{P}(X \mid \mathbf{e})$

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$$

Summation is over all possible combinations of values of the unobserved variables $\mathbf{Y}$

X = Query variable (a single variable for now)

$\mathbf{E}$ = Set of evidence variables

$\mathbf{e}$ = the set of observed values for the evidence variables

$\mathbf{Y}$ = Unobserved variables

Inference

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$$

Computing $\mathbf{P}(X \mid \mathbf{e})$ involves going through all possible entries of the full joint probability distribution and adding up probabilities with $X=x_i$, $\mathbf{E}=\mathbf{e}$, and $\mathbf{Y}=\mathbf{y}$

Suppose you have a domain with n Boolean variables. What is the space and time complexity of computing $P(X \mid \mathbf{e})$?

14

Independence

- How do you avoid the exponential space and time complexity of inference?
- Use independence (aka factoring)

15

Independence

Suppose the full joint distribution now consists of four variables:

Toothache = {*true*, *false*}

Catch = {*true*, *false*}

Cavity = {*true*, *false*}

Weather = {*sunny*, *rain*, *cloudy*, *snow*}

There are now 32 entries in the full joint distribution table

16

Independence

Does the weather influence one's dental problems?

Is P(*Weather=cloudy* | *Toothache = toothache*, *Catch = catch*, *Cavity = cavity*) = P(*Weather=cloudy*)?

In other words, is *Weather* independent of *Toothache*, *Catch* and *Cavity*?

17

Independence

We say that variables X and Y are independent if any of the following hold: (note that they are all equivalent)

$$\boldsymbol{P}(X \mid Y) = \boldsymbol{P}(X) \quad \text{or}$$

$$\boldsymbol{P}(Y \mid X) = \boldsymbol{P}(Y) \quad \text{or}$$

$$\boldsymbol{P}(X, Y) = \boldsymbol{P}(X)\boldsymbol{P}(Y)$$

18

Why is independence useful?

Assume that Weather is independent of toothache, catch, cavity i.e.

P(*Weather=cloudy* | *Toothache = toothache*, *Catch = catch*, *Cavity = cavity*) = P(*Weather=cloudy*)

Now we can calculate:

P(*Weather=cloudy, Toothache = toothache*, *Catch = catch*, *Cavity = cavity)*

= P(*Weather=cloudy* | *Toothache = toothache*, *Catch = catch*, *Cavity = cavity*) * P(*toothache*, *catch*, *cavity*)

= P(*Weather=cloudy*) * P(*Toothache = toothache*, *Catch = catch*, *Cavity = cavity*)

Why is independence useful?

P(*Weather=cloudy*, *Toothache = toothache*, *Catch = catch*, *Cavity = cavity)*

$$= \underbrace{P(Weather=cloudy)}_{\text{This table has 4 values}} * \underbrace{P(Toothache = toothache, Catch = catch, Cavity = cavity)}_{\text{This table has 8 values}}$$

- You now need to store 12 values to calculate ***P***(*Weather*, *Toothache*, *Catch*, *Cavity*)
- If *Weather* was not independent of *Toothache*, *Catch*, and *Cavity* then you would have needed 32 values

Independence

Another example:

- Suppose you have n coin flips and you want to calculate the joint distribution $\boldsymbol{P}(C_1, \ldots, C_n)$
- If the coin flips are not independent, you need 2^n values in the table
- If the coin flips are independent, then

$$P(C_1, \ldots, C_n) = \prod_{i=1}^{n} P(C_i)$$

Each P(C_i) table has 2 entries and there are *n* of them for a total of 2*n* values

Independence

- Independence is powerful!
- It required extra domain knowledge. A different kind of knowledge than numerical probabilities. It needed an understanding of causation.

Bayes' Rule

The product rule can be written in two ways:

P(A, B) = P(A | B)P(B)

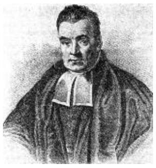
P(A, B) = P(B | A)P(A)

You can combine the equations above to get:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

23

Bayes' Rule



More generally, the following is known as Bayes' Rule:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)}$$

Note that these are distributions

Sometimes, you can treat ***P***(B) as a normalization constant α

$$\mathbf{P}(A|B) = \alpha \mathbf{P}(B|A)\mathbf{P}(A)$$

24

More General Forms of Bayes Rule

If A takes 2 values:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

If A takes n_A values:

$$P(A=v_i|B) = \frac{P(B|A=v_i)P(A=v_i)}{\sum_{k=1}^{n_A} P(B|A=v_k)P(A=v_k)}$$

25

Bayes Rule Example

Meningitis causes stiff necks with probability 0.5. The prior probability of having meningitis is 0.00002. The prior probability of having a stiff neck is 0.05. What is the probability of having meningitis given that you have a stiff neck?

Let m = patient has meningitis

Let s = patient has stiff neck

P(s | m) = 0.5

P(m) = 0.00002

P(s) = 0.05

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{(0.5)(0.00002)}{0.05} = 0.0002$$

Bayes Rule Example

Meningitis causes stiff necks with probability 0.5. The prior probability of having meningitis is 0.00002. The prior probability of having a stiff neck is 0.05. What is the probability of having meningitis given that you have a stiff neck?

Let m = patient has meningitis

Let s = patient has stiff neck

$P(s \mid m) = 0.5$

$P(m) = 0.00002$

$P(s) = 0.05$

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{(0.5)(0.00002)}{0.05} = 0.0002$$

Note: Even though P(s|m) = 0.5, P(m|s) = 0.0002

When is Bayes Rule Useful?

Sometimes it's easier to get P(X|Y) than P(Y|X).

Information is typically available in the form P(effect | cause) rather than P(cause | effect)

For example, P(symptom | disease) is easy to measure empirically but obtaining P(disease | symptom) is harder

How is Bayes Rule Used

In machine learning, we use Bayes rule in the following way:

$$\underbrace{\boldsymbol{P}(h \mid D)}_{\text{Posterior probability}} = \frac{\overbrace{\boldsymbol{P}(D \mid h)}^{\text{Likelihood of the data}} \overbrace{\boldsymbol{P}(h)}^{\text{Prior probability}}}{\boldsymbol{P}(D)}$$

h = hypothesis

D = data

Bayes Rule With More Than One Piece of Evidence

Suppose you now have 2 evidence variables *Toothache* = *true* and *Catch* = *catch* (note that Cavity is uninstantiated below)

$\boldsymbol{P}(Cavity \mid Toothache = true, Catch = true) = \alpha$
$\boldsymbol{P}(Toothache = true, Catch = true \mid Cavity)\ \boldsymbol{P}(Cavity)$

In order to calculate $\boldsymbol{P}(Toothache = true, Catch = true \mid Cavity)$, you need a table of 4 probability values. With N evidence variables, you need 2^N probability values.

Conditional Independence

Are *Toothache* and *Catch* independent?

No – if probe catches in the tooth, it likely has a cavity which causes the toothache.

But given the presence or absence of the cavity, they are independent (since they are directly caused by the cavity but don't have a direct effect on each other)

Conditional independence:

P(*Toothache* = *true*, *Catch* = *true* | *Cavity*) =

P(*Toothache* = *true* | *Cavity*) * ***P***(*Catch* = *true* | *Cavity*)

31

Conditional Independence

General form:

$$\boldsymbol{P}(A, B \mid C) = \boldsymbol{P}(A \mid C)\boldsymbol{P}(B \mid C)$$

Or equivalently:

$$\boldsymbol{P}(A \mid B, C) = \boldsymbol{P}(A \mid C)$$ and

$$\boldsymbol{P}(B \mid A, C) = \boldsymbol{P}(B \mid C)$$

How to think about conditional independence:

In P(*A* | *B*, *C*) = P(*A* | *C*): if knowing *C* tells me everything about A, I don't gain anything by knowing *B*

32

Conditional Independence

$$\overbrace{\boldsymbol{P}(Toothache, Catch, Cavity)}^{\text{7 independent values in table (have to sum to 1)}}$$

$$= \boldsymbol{P}(Toothache, Catch \mid Cavity)\ \boldsymbol{P}(Cavity)$$

$$= \underbrace{\boldsymbol{P}(Toothache \mid Cavity)}_{\text{2 independent values in table}}\ \underbrace{\boldsymbol{P}(Catch \mid Cavity)}_{\text{2 independent values in table}}\ \underbrace{\boldsymbol{P}(Cavity)}_{\text{1 independent value in table}}$$

Conditional independence permits probabilistic systems to scale up!

33

What You Should Know

- How to do inference in joint probability distributions
- How to use Bayes Rule
- Why independence and conditional independence is useful

34