COSC343: Artificial Intelligence

Lecture 5: Bayesian Reasoning

Lech Szymanski

Dept. of Computer Science, University of Otago

In today's lecture

- Conditional probability and independence
- Curse of dimensionality
- Bayes' rule
- Naive Bayes Classifier

Recap: a simple medical example

Consider a medical scenario, with 3 Boolean variables

- *cavity* (does the patient have a cavity or not?)
- *toothache* (does the patient have a toothache or not?)
- *catch* (does the dentist's probe catch on the patient's tooth?)

Here's an example probability model: the joint probability distribution $p(Toothache, Cavity, Catch)$

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Conditional probabilities

Probability of an event **given** information about another event.

"given" "and"

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}, \text{ if } P(b) \neq 0$$

probability of a given b

Computing conditional probabilities

Assume we're given the following joint probability distribution $p(toothache, cavity, catch)$ and learn that a patient has a toothache

E.g. how to calculate $P(\neg cavity|toothache)$?

$$P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064}$$
$$= 0.4$$

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

Conditional probability of whole distributions

Conditional probabilities can also be computed for whole distributions. $p(Toothache, Cavity, Catch)$

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

For instance, the **conditional probability distribution** of *Cavity* and *Catch* given *toothache* shows all possible values of *Cavity* and *Catch* given *toothache*. $p(Cavity, Catch|toothache)$

	catch	*¬ catch*
cavity	.188/.2	.12/.2
¬cavity	.16/.2	.064/.2

Complexity of reasoning from a full joint distribution

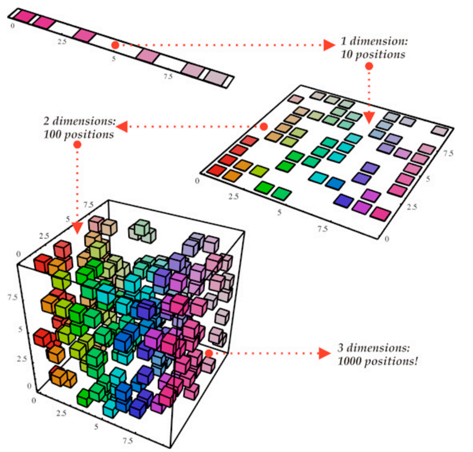
Inference from full joint distributions is all about summing over hidden variables.

- If there are n Boolean variables in the table, it takes $O(2^n)$ time to compute a probability
- Moreover, the space complexity of the table itself is $O(2^n)$
  - This means as n increases, it becomes infeasible even to store the full joint distribution
  - And remember: someone has to build the full joint distribution from empirical data!

In practice, probabilistic reasoning tasks often involve thousands of random variables. Obviously, a more efficient reasoning technique is necessary.

Curse of dimensionality

When the dimensionality of problem increases, the volume of the space becomes so large, you need exponentially more points to sample the space at the same density. This phenomenon is sometimes referred as the **curse of dimensionality**.



https://haifengl.wordpress.com/2016/02/29/there-is-no-big-data-in-machine-learning/

Independence

Huge reductions in complexity can be made if we know that some variables in our domain are **independent** of others

- If $P(A|B) = P(A)$ then it is said that A and B are independent

If A and B are independent, then it follows that:

$$P(A,B) = P(A)P(B)$$

and

$$P(B|A) = P(B)$$

An example of independence

Say we add a variable *Weather* to our dentist model, with four values: *sunny*, *cloudy*, *rainy*, *snowy*.

- Weather and teeth don't affect each other
- To compute the probability of a proposition involving weather and teeth, we can just multiply the results found in the two distributions

Given:

$$P(toothache|Weather = cloudy) = P(toothache)$$

it follows that:

$$P(toothache \wedge Weather = cloudy) = P(toothache)P(Weather = cloudy)$$

Independence

If we wanted to create a full joint distribution, we could do so by multiplying the appropriate probabilities from the two distributions

$$p(Toothache, Cavity, Catch, Weather)$$
$$= p(Toothache, Cavity, Catch)p(Weather)$$

- Instead of 4 dimensions of 2x2x2x4=32 samples we have two spaces of 3 and 1 dimensions with total of 2x2x2+4=12 samples

	toothache		*toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

sunny	.6
cloudy	.029
rainy	.1
snowy	.001

Conditional independence

Absolute independence is powerful, but rare.

- Can we factor out $p(Toothache, Cavity, Catch)$ further?
- For instance, in dentistry, most variables are related, but not directly.

What do do?

A very useful concept is **conditional independence**

- For instance, having a cavity
  - changes the probability of a toothache
  - and changes the probability of a catch
  - but these two effects are *separate*!

Conditional independence

Knowing the value of *Toothache* certainly gives us information about the value of *Catch*.

- But only indirectly, as they are both linked to the variable Cavity.

If I already knew the value of *Cavity*, then learning about *Catch* won't tell me anything else about the value of *Toothache*.

$$p(Toothache|Cavity, Catch) = P(Toothache|Cavity)$$

Therefore:

Conditional probability

$$\begin{aligned} P(Toothache, Catch, Cavity) &= P(Toothache|Catch, Cavity)P(Catch, Cavity) \\ &= P(Toothache|Catch, Cavity)P(Catch|Cavity)P(Cavity) \\ &= P(Toothache|Cavity)P(Catch|Cavity)P(Cavity) \end{aligned}$$

Conditional independence

Bayes' rule

From the definition of conditional probability

$$P(b|a) = \frac{P(a,b)}{P(a)} \quad \text{and} \quad P(a|b) = \frac{P(a,b)}{P(b)}$$

it follows that

$$P(a,b) = P(b|a)P(a) = P(a|b)P(b)$$

and so

Bayes' Rule

$$\boxed{P(b|a) = \frac{P(a|b)P(b)}{P(a)}}$$

A problem

Information from previous patients:

effects

cause

Chills (C)	Running nose (R)	Headache (H)	Fever (Fr)	Flu
y	n	mild	y	n
y	y	no	n	y
y	n	strong	y	y
n	y	mild	y	y
n	n	no	n	n
n	y	strong	y	y
n	y	strong	n	n
y	y	mild	n	y

How to diagnose a person with the following symptoms?

y	n	mild	n	?

Compute

$$P(Flu = y|C = y, R = n, H = mild, Fr = n)$$

and

$$P(Flu = n|C = y, R = n, H = mild, Fr = n)$$

and see which one is more likely.

What's more likely?

cause

effects

Larger or smaller?

$$P(Flu = y|C, H, R, Fr) <?> P(Flu = n|C, H, R, Fr)$$

It's hard to estimate $P(cause|effect)$

- You have to have information on many groups of people – one for each possible combination of symptoms – and test each person in each group for flu

It's far easier to estimate $P(effect|cause)$:

- You take a sample of only two groups of people – those with flu and those without, and test them for symptoms

From Bayes' rule

$$P(Flu|C, R, H, Fr) = \frac{P(C, R, H, Fr|Flu)P(Flu)}{P(C, R, H, Fr)}$$

Invert the problem

cause

effects

$$P(Flu=y|C,H,R,Fr) <?> P(Flu=n|C,H,R,Fr)$$

Bayes

cause

effects

$$\frac{P(C,R,H,Fr|Flu=y)P(Flu=y)}{\cancel{P(C,R,H,Fr)}} <?> \frac{P(C,R,H,Fr|Flu=n)P(Flu=n)}{\cancel{P(C,R,H,Fr)}}$$

Normalising denominator is the same on both sides, and so plays no part in the comparison

$$P(C,R,H,Fr|Flu=y)P(Flu=y) <?> P(C,R,H,Fr|Flu=n)P(Flu=n)$$

Need probability distributions $p(Flu)$ and $p(C,R,H,Fr|Flu)$

Gathering prior probabilities from training set

Chills (C)	Running nose (R)	Headache (H)	Fever (Fr)	Flu
y	n	mild	y	n
y	y	no	n	y
y	n	strong	y	y
n	y	mild	y	y
n	n	no	n	n
n	y	strong	y	y
n	y	strong	n	n
y	y	mild	n	y

$p(Flu)$

y	n
5/8	3/8

Gathering conditional probabilities from training set

Chills (C)	Running nose (R)	Headache (H)	Fever (Fr)	Flu
y	n	mild	y	n
y	y	no	n	y
y	n	strong	y	y
n	y	mild	y	y
n	n	no	n	n
n	y	strong	y	y
n	y	strong	n	n
y	y	mild	n	y

- joint probability distribution $p(C,R,H,Fr|Flu)$ has 4 dimensions and 24 combinations of symptoms
- $p(C|Flu)$ has 1 dimension and 2 combinations of symptoms
- $p(R|Flu)$ has 1 dimension and 2 combinations of symptoms
- $p(H|Flu)$ has 1 dimension and 2 combinations of symptoms
- $p(Fr|Flu)$ has 1 dimension and 3 combinations of symptoms

- If the symptoms were independent events, then from independence we would have

$$p(C,R,H,Fr|Flu) = p(C|Flu)p(R|Flu)p(H|Flu)p(Fr|Flu)$$

Let's just assume the symptoms are independent! Naive Bayes

Gathering conditional probabilities from training set

Chills (C)	Running nose (R)	Headache (H)	Fever (Fr)	Flu
y	n	mild	y	n
y	y	no	n	y
y	n	strong	y	y
n	y	mild	y	y
n	n	no	n	n
n	y	strong	y	y
n	y	strong	n	n
y	y	mild	n	y

$p(Flu)$

y	n
5/8	3/8

$p(C|Flu=y)$

y	n
3/5	2/5

$p(C|Flu=n)$

y	n
1/3	2/3

$p(R|Flu=y)$

y	n
4/5	1/5

$p(R|Flu=n)$

y	n
1/3	2/3

$p(H|Flu=y)$

mild	no	strong
2/5	1/5	2/5

$p(H|Flu=n)$

mild	no	strong
1/3	1/3	1/3

$p(Fr|Flu=y)$

y	n
3/5	2/5

$p(Fr|Flu=n)$

y	n
1/3	2/3

Naive Bayes Classifier

Probability of flu given symptoms:

$$P(Flu|C,R,H,Fr) \sim P(C,R,H,Fr|Flu)$$ - From Bayes' rule

$$= P(C|Flu)P(R|Flu)P(H|Flu)P(Fr|Flu)$$ - From the assumption of independence

And so, for the person with the symptoms:

C	R	H	Fr	Flu
y	n	mild	n	?

$$P(Flu{=}\mathrm{y}|C=y,R=n,H=mild,Fr=n) \sim P(C=y|Flu=y)P(R=n|Flu=y)P(H=mild|Flu=y)P(Fr=n|Flu=y)P(Flu=y)$$
$$=\left(\frac{3}{5}\right)\left(\frac{1}{5}\right)\left(\frac{2}{5}\right)\left(\frac{2}{5}\right)\left(\frac{5}{8}\right)$$
$$=0.12$$

- From the probability distributions derived from the training data

$$P(Flu{=}\mathrm{n}|C=y,R=n,H=mild,Fr=n) \sim P(C=y|Flu=y)P(R=n|Flu=n)P(H=mild|Flu=n)P(Fr=n|Flu=n)P(Flu=n)$$
$$=\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{3}{8}\right)$$
$$=0.19$$

Naive Bayes Classifier deems this person more likely not to have a flu!

Naive Bayes Classifier

flu

$$y = f(\mathbf{x},\mathbf{w}) = d\Big(f_1(\mathbf{x},\mathbf{w}), f_2(\mathbf{x},\mathbf{w})\Big) = \begin{cases} 1 & f_1(\mathbf{x},\mathbf{w}) > f_2(\mathbf{x},\mathbf{w}) \\ 0 & \text{otherwise}, \end{cases}$$

no flu

where

$$f_1(\mathbf{x},\mathbf{w}) = \Big(x_1 w_1 + (1-x_1)w_2\Big)\Big(x_2 w_3 + (1-x_2)w_4\Big)\Big(\delta(x_3)w_5 + \delta(x_3-1)w_6 + \delta(x_3-2)w_7\Big)\Big(x_4 w_8 + (1-x_4)w_9\Big)w_{10}$$
$$f_2(\mathbf{x},\mathbf{w}) = \Big(x_1 w_{11} + (1-x_1)w_{12}\Big)\Big(x_2 w_{13} + (1-x_2)w_{14}\Big)\Big(\delta(x_3)w_{15} + \delta(x_3-1)w_{16} + \delta(x_3-2)w_{17}\Big)\Big(x_4 w_{18} + (1-x_4)w_{19}\Big)w_{20}$$

and

$$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

y n

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \begin{array}{l} \rightarrow C \in \{0,1\} \\ \rightarrow R \in \{0,1\} \\ \rightarrow H \in \{0,1,2\} \\ \downarrow Fr \in \{0,1\} \end{array}$$

no mild strong

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \\ w_{10} \\ w_{11} \\ w_{12} \\ w_{13} \\ w_{14} \\ w_{15} \\ w_{16} \\ w_{17} \\ w_{18} \\ w_{19} \\ w_{20} \end{bmatrix} \begin{array}{l} \rightarrow p(C=y|Flu=y) \\ \rightarrow p(C=n|Flu=y) \\ \rightarrow p(R=y|Flu=y) \\ \rightarrow p(R=n|Flu=y) \\ \rightarrow p(H=no|Flu=y) \\ \rightarrow p(H=mild|Flu=y) \\ \rightarrow p(H=strong|Flu=y) \\ \rightarrow p(Fr=y|Flu=y) \\ \rightarrow p(Fr=n|Flu=y) \\ \rightarrow p(Flu=y) \\ \rightarrow p(C=y|Flu=n) \\ \rightarrow p(C=n|Flu=n) \\ \rightarrow p(R=y|Flu=n) \\ \rightarrow p(R=n|Flu=n) \\ \rightarrow p(H=no|Flu=n) \\ \rightarrow p(H=mild|Flu=n) \\ \rightarrow p(H=strong|Flu=n) \\ \rightarrow p(Fr=y|Flu=n) \\ \rightarrow p(Fr=n|Flu=n) \\ \rightarrow p(Flu=n) \end{array}$$

Summary and reading

- Conditional probabilities model changes in uncertainty given some information
- Independence simplifies computation
- Bayes' rule: possible to compute $p(effect|causes)$ from $p(causes|effect)$
- Naive Bayes: assume *causes* input variables are independent, so that

$$p(causes|effect) = \prod_i p(cause_i|effect)$$

Reading for the lecture: AIMA Chapter 13 Sections 3-5

Reading for next lecture: AIMA Chapter 18 Section 3