Reasoning under uncertainty

Probability Review II

CSC384

March 16, 2018

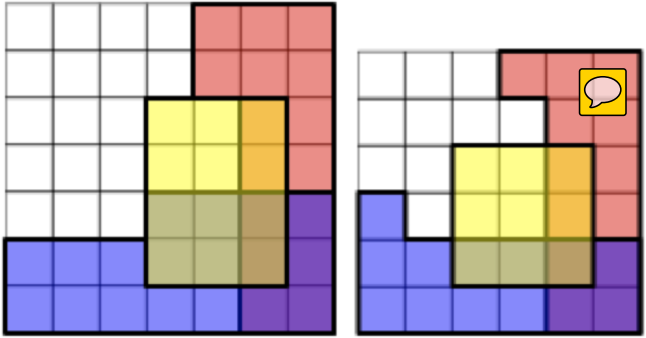
From last class

- Axioms of probability
- Probability over feature vectors
- Independence
- Conditional independence

Outline

1. Conditional independence
2. Chain rule
3. Bayes' Rule
4. Probability distributions

Conditional independence (continued)



Conditional independence (continued)

- The pictures in the previous slide represent the probabilities of events A, B and C by the areas shaded red, blue and yellow respectively with respect to the total area. In both examples A and B are conditionally independent given membership in set C because:
  $P(A|B \wedge C) = P(A|C)$
- Note however that B and C are not independent as $P(B \wedge C) \neq P(B) * P(C)$ in either picture.
- Also note that $P(A \wedge B|C) = P(A|C) * P(B|C)$ but A and B are NOT conditionally independent given membership in the set $\leftarrow C$, as:
  $P(A \wedge B| \leftarrow C) \neq P(A| \leftarrow C) * P(B| \leftarrow C)$

More on summing out variables

- Say that B_1, B_2, ..., B_k partition of the universe U and say that each B_i is defined by a particular value being assigned to a variable $(V_2 = b_i)$.
- $B_i \cap B_j = \{\}$, where $i \neq j$ (mutually exclusive)
- $B_1 \cup B_2 \cup B_3 ... \cup B_k = U$ (exhaustive)
- In probabilities:
  $P(B_i \cap B_j) = 0$ $P(B_1 \cup B_2 \cup B_3 ... \cup B_k) = 1$

More on summing out variables

- Given another set of events A we know that $P(A) = P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_k)$
- We can write this as conditional probabilities:
- $P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ... + P(A|B_k)P(B_k)$
- $P(A|B_i)P(B_i) = P(A \cap Bi)/P(B_i) * P(B_i) = Pr(A \cap Bi)$
- Often we know $P(A|B_i)$, so we can compute P(A) by "summing" across the B_i sets in the equation above (to sum out variable V_2).

The chain rule

- The joint probability is the probability of two events happening together.
- The chain rule allows us to calculate a joint probability using only conditional probabilities.

$$P(A_1 \wedge A_2 \wedge ... \wedge A_n) =$$

$$\mathrm{P}(\mathrm{A}_1|A_2 \wedge ... \wedge A_n) * P(A_2|A_3 \wedge ... \wedge A_n) * ... * P(A_{n-1}|A_n) * P(A_n)$$

The chain rule

- The joint probability is the probability of two events happening together.
- The chain rule allows us to calculate a joint probability using only conditional probabilities.

$$P(A_1 \wedge A_2 \wedge ... \wedge A_n) =$$

$$\mathrm{P(A_1}|A_2 \wedge ... \wedge A_n) * P(A_2|A_3 \wedge ... \wedge A_n) * ... * P(A_{n-1}|A_n) * P(A_n)$$

- Proof:
  $P(A_1|A_2 \wedge ... \wedge A_n) * P(A_2|A_3 \wedge ... \wedge A_n) * ... * P(A_{n-1}|A_n)$
  $= \frac{P(A_1 \wedge A_2 \wedge ... \wedge A_n)}{P(A_2 \wedge ... \wedge A_n)} * \frac{P(A_2 \wedge ... \wedge A_n)}{P(A_3 \wedge ... \wedge A_n)} * ... * \frac{P(A_{n-1} \wedge A_n)}{P(A_n)} * P(A_n)$

Bayes' Rule

- Bayes rule is a simple mathematical fact. But it has great implications w.r.t. how probabilities can be reasoned with.

$$Pr(Y|X) = \frac{Pr(X|Y)Pr(Y)}{Pr(X)}$$

Bayes' Rule

- Bayes rule is a simple mathematical fact. But it has great implications w.r.t. how probabilities can be reasoned with.

$$Pr(Y|X) = \frac{Pr(X|Y)Pr(Y)}{Pr(X)}$$

- Proof:
  $Pr(Y|X) = Pr(Y \wedge X)/Pr(X)$
  $= Pr(Y \wedge X)/Pr(X) * P(Y)/P(Y)$
  $= Pr(Y \wedge X)/Pr(Y) * Pr(Y)/Pr(X)$
  $= Pr(X|Y)Pr(Y)/Pr(X)$

Using Bayes' rule - Malaria example

- Diseases $\in$ malaria, cold, flu; symptoms $=$ fever
  - Must compute $P(disease|fever)$ to prescribe treatment
- Why not assess this quantity directly?
  - $P(malaria|fever)$ is not easy to assess nor does it reflect the underlying causal mechanism (i.e. that malaria causes fever).
  - $P(malaria|fever)$ is not stable: a malaria epidemic may change this quantity (for example)
- Try Bayes rule:
  $P(malaria|fever) = P(fever|malaria) * P(malaria)/P(fever)$

Using Bayes' rule - Malaria example

- $Pr(malaria|fever) = Pr(fever|malaria)Pr(malaria)/Pr(fever)$
- $Pr(malaria)$?
  - This is the prior probability of Malaria, i.e., before you exhibited a fever, and with it we can account for other factors, e.g., a malaria epidemic, or recent travel to a malaria risk zone.
  - E.g., The center for disease control keeps track of the rates of various diseases.
- $Pr(fever|malaria)$?
  - This is the probability a patient with malaria exhibits a fever.
  - Again this kind of information is available from people who study malaria and its effects.

Using Bayes' rule - Malaria example

- $Pr(fever)$?
  - This is typically not known, but it can be computed!
  - We eventually have to divide by this probability to get the final answer: $Pr(malaria|fever) = Pr(fever|malaria)Pr(malaria)/Pr(fever)$
- First, we find a set of mutually exclusive and exhaustive causes for fever:
  - Say that in our example, malaria, cold and flu are only possible causes of fever and they are mutually exclusive.
  - $Pr(fever|\neg malaria \wedge \neg cold \wedge \neg flu) = 0$
    Fever cant happen with one of these causes.
  - $Pr(malaria \wedge cold) = Pr(malaria \wedge flu) = Pr(cold \wedge flu) = 0$
    these causes cant happen together. (Note that our example is not very realistic!)
- Second, we compute: $Pr(fever|malaria)Pr(malaria)$, $Pr(fever|cold)Pr(cold)$, $Pr(fever|flu)Pr(flu)$.
  - We know $Pr(fever|cold)$ and $Pr(fever|flu)$, along with $Pr(cold)$ and $Pr(flu)$ from the same sources as $Pr(fever|malaria)$ and $Pr(malaria)$.

Using Bayes' rule - Malaria example

- Since flu, cold and malaria are exclusive, flu, cold, malaria, $\neg$ malaria $\wedge$ $\neg$cold $\wedge$ $\neg$flu forms a partition of the universe. So:
  $Pr(fever) = Pr(fever|malaria) * Pr(malaria) + Pr(fever|cold) * Pr(cold) + Pr(fever|flu) * Pr(flu) + Pr(fever|\neg malaria \wedge \neg cold \wedge \neg flu) * Pr(\neg malaria \wedge \neg cold \wedge \neg flu)$
- The last term is zero as fever is not possible unless one of malaria, cold, or flu is true.
- So to compute the trio of numbers, $Pr(malaria|fever)$, $Pr(cold|fever)$, $Pr(flu|fever)$, we compute the trio of numbers $Pr(fever|malaria) * Pr(malaria)$, $Pr(fever|cold) * Pr(cold)$, $Pr(fever|flu) * Pr(flu)$
- And then we divide these three numbers by $Pr(fever)$.
  - That is we divide these three numbers by their sum: This is called normalizing the numbers.
- Thus we never need actually compute Pr(fever) (unless we want to).

Normalizing

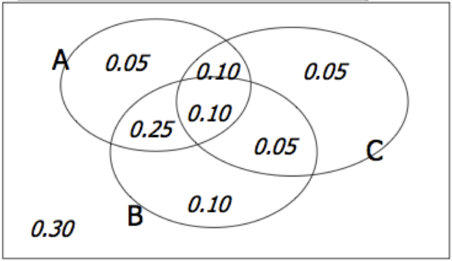
- If we have a vector of k numbers, e.g., $\langle 3, 4, 2.5, 1, 10, 21.5\rangle$ we can normalize these numbers by dividing each number by the sum of the numbers:
  $3 + 4 + 2.5 + 1 + 10 + 21.5 = 42$
- Normalized vector:
  $\langle 3/42, 4/42, 2.5/42, 1/42, 10/42, 21.5/42\rangle$
  $= \langle 0.071, 0.095, 0.060, 0.024, 0.238, 0.512\rangle$
- After normalizing the vector of numbers sums to 1
- Exactly what is needed for these numbers to specify a probability distribution.

Creating probability distributions

- A joint distribution records the probabilities that variables will hold particular values.
- They can be populated using expert knowledge, by using the axioms of probability, or by actual data.
- The sum of all the probabilities MUST be 1 in order to satisfy the axioms of probability.
- We can use normalization to convert raw counts of data into a legal probability distribution (i.e. into a distribution that sums to 1).

Creating probability distributions

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the joint distribution

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the joint distribution

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the joint distribution

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the joint distribution

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Using the joint distribution

Inference with the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Conditional independence is symmetric

- Assume $P(X|Y \wedge Z) = P(X|Y)$
- Then: $P(Z|Y \wedge X) = P(Z|Y)$

Conditional independence is symmetric

- Assume $P(X|Y \wedge Z) = P(X|Y)$
- Then: $P(Z|Y \wedge X) = P(Z|Y)$
- Proof:
  $P(Z|X \wedge Y) = P(X \wedge Y|Z) * P(Z)/P(X \wedge Y)$ (Bayes Rule)
  $= P(X|Y \wedge Z) * P(Y|Z) * P(Z)/P(X|Y) * P(Y)$ (Chain Rule)
  $= P(X|Y) * P(Y|Z) * P(Z)/P(X|Y) * P(Y)$ (By Assumption)
  $= P(Y|Z) * P(Z)/P(Y) = P(Z|Y)$ (Bayes Rule)