Reasoning Under Uncertainty

Uncertainty material is covered in chapters 13 and 14.

Chapter 13 gives some basic background on probability from the point of view of A.I.

Chapter 14 talks about Bayesian Networks, exact reasoning in Bayes Nets as well as approximate reasoning, which will be main topics for us.

*Note: Slides in this section draw **from** Faheim Bacchus, Craig Boutillier, Andrew Moore, Sheila McIlraith ...*

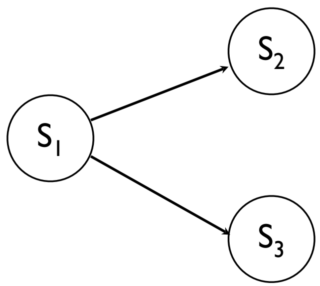
Reasoning Under Uncertainty

1. Assignment 3 is due March 17.
2. Drop deadline is March 15.
3. Assignment 4 will cover uncertainty and be posted March 16.
4. Assignment 4 will be due on April 3rd.

Reasoning Under Uncertainty

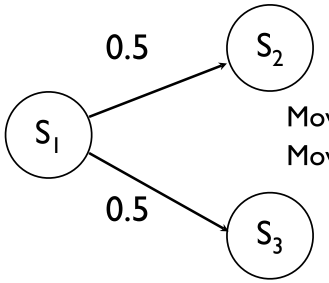
- The world is a very uncertain place.
- As of this point, we've basically danced around that fact. We've assumed that what we see in the world is really there, what we do in the world has predictable outcomes, etc.
    - i.e., if you are in state S_1 and you execute action A you arrive at state S_2.

Example: Sokoban



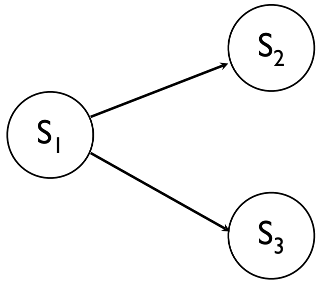
move(S_1,'N') = S_2
move(S_1,'S') = S_3

Probabilistic Sokoban is a Very Different Game



Move(S_1,'N') = S_2 50% of the time
Move(S_1,'N') = S_3 50% of the time

Probabilistic Sokoban is a Very Different Game



Based on what we can see, there's a 30% chance we're in cell S_1, 30% in S_s and 40% in S_3....

Life in an Uncertain World

We might not know the effects of an action

- The action might have a random component, like rolling dice.
- We might not know the long term effects of a drug.
- We might not know the status of a road when we choose to drive down it.

We may not know exactly what state we are in

- E.g., we can't see our opponents cards in a poker game.
- We don't know what a patient's ailment is.

We may still need to act, but we can't act solely on the basis of facts. We have to "gamble".

Uncertainty

But how do we gamble rationally?

- If we must arrive at the airport at 9pm on a week night we could "safely" leave for the airport ½ hour before.

  Some probability of the trip taking longer, but the probability is low.

- If we must arrive at the airport at 4:30pm on Friday we most likely need 1 hour or more to get to the airport.

  Relatively high probability of it taking 1.5 hours.

- Acting rationally under uncertainty typically corresponds to maximizing one's **expected utility**. There are various reason for doing this.

Expected Utility

You may not know what state arises from your actions due to uncertainty. But if you know (or can estimate) the probability you are in each of these different states (i.e., if you have a probability distribution) you can compute the expected utility and take the actions that lead to a distribution with highest expected utility.

Expected Utility Example

- Probability distribution over outcomes (also called a "joint distribution")

Event	Go to Bloor St.	Go to Queen Street
Find Ice Cream	0.5	0.2
Find donuts	0.4	0.1
Find live music	0.1	0.7

- Utilities of outcomes

Event	Utility
Ice Cream	10
Donuts	5
Music	20

Expected Utility Example

- Maximum Expected Utility?

Event	Go to Bloor St.	Go to Queen Street
Ice Cream	0.5 * 10	0.2 *10
Donuts	0.4 * 5	0.1 * 5
Music	0.1 * 20	0.7 * 20
Utility	**9.0**	**16.5**

- Here, it's "Go to Queen Street"
- If the utility of Donuts of Ice Cream had been higher, however, it might have been "Go to Bloor Street".

Maximizing Utility

So, to maximize utilities, we will need:

- Probability Distributions and tools to reason about probabilities
- Mechanisms to discover utilities or preferences. This is an active area of research.

Review: Probability Distributions over Finite Sets

A probability is a function defined over a set of atomic events U.

U represents the universe of all possible events.

Review: Probability over Finite Sets

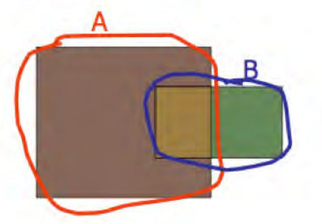
Given **U** (a universe of events), a probability function is a function defined over subsets of **U** that maps each subset onto the real numbers and that satisfies the Axioms of Probability. These are:

1. **P(U) = 1**
2. **P(A) ∈ [0,1]**
3. **P({}) = 0**
4. **P(A ∪ B) = P(A) + P(B) – P(A ∩ B)**

NB: if A ∩ B = {} then P(A ∪ B) = P(A) + P(B)

Review: Probability over Finite Sets

P(A ∪ B) = P(A) + P(B) – P(A ∩ B)



Notation: Properties and Sets

We often write

A∨B: to represent the set of events with either property A or B, i.e. the set A∪B

A∧B: to represent the set of events both property A and B, i.e. the set A∩B

¬A: to represent the set of events that do not have property A: the set U-A (i.e., the complement of A w.r.t. the universe of events U)

Review: Probability over Feature Vectors

As we move forward, ee will model sets of events in our universe as vectors of feature values.

Like CSPs, we have

1. a set of variables $V_1, V_2, \ldots, V_n$
2. a finite domain of values for each variable, $Dom[V_1], Dom[V_2], \ldots, Dom[V_n]$.

The universe of events U is the set of all vectors of values for the variables

$$\langle d_1, d_2, \ldots, d_n \rangle: d_i \in Dom[V_i]$$

When we write P(A=a, B=b), we will mean the probability that variable A has been assigned value 'a' **and** variable B has been assigned value 'b'. Note that here, sets of events are induced by a given value assignment. So, P(A=a) represents a set of events in which A holds the value 'a'.

Review: Probability over Feature Vectors

Our event space has size $\prod_i |Dom[V_i]|$, i.e., the product of the domain sizes. If $|Dom[V_i]| = 2$, we have 2^n distinct atomic events.

Note the size of possible event outcomes (or variable assignments) grows **exponentially** with the number of variables.

Review: Probability over Feature Vectors

We often want to look at subsets of U defined by value assignments to particular variables.

E.g.

$\{V_1 = a\}$ is the set of all events where $V_1 = a$

$\{V_1 = a, V_3 = d\}$ is the set of all events where $V_1 = a$ and $V_3 = d$.

Note that

$$P(\{V_1 = a\}) = \sum_{x \in Dom[V_3]} P(\{V_1 = a, V_3 = x\}).$$

Review: Probability over Feature Vectors

If we have probability of every atomic event (wherein every event is a full instantiation of the variables) we can compute the probability of any other set of events.

E.g.

$\{V_1 = a\}$ is the set of all events where $V_1 = a$

$$P(\{V_1 = a\}) = \sum_{x_2 \in Dom[V_2]} \sum_{x_3 \in Dom[V_3]} \sum_{x_4 \in Dom[V_4]} \ldots \sum_{x_n \in Dom[V_n]} P(\{V_1 = a, V_2 = x_2, V_3 = x_3, V_4 = x_4 \ldots, V_n = x_n\}).$$

Review: Probability over Feature Vectors

Example:

$$P(\{V_1 = 1\}) = \sum_{x_2 \in Dom[V_2]}, \sum_{x_3 \in Dom[V_3]} P(\{V_1 = 1, V_2 = x_2, V_3 = x_3\}).$$

(V1 = 1, V2 = 1, V3 = 1)	(V1 = 2, V2 = 1, V3 = 1)	(V1 = 3, V2 = 1, V3 = 1)
(V1 = 1, V2 = 1, V3 = 2)	(V1 = 2, V2 = 1, V3 = 2)	(V1 = 3, V2 = 1, V3 = 2)
(V1 = 1, V2 = 1, V3 = 3)	(V1 = 2, V2 = 1, V3 = 3)	(V1 = 3, V2 = 1, V3 = 3)
(V1 = 1, V2 = 2, V3 = 1)	(V1 = 2, V2 = 2, V3 = 1)	(V1 = 3, V2 = 2, V3 = 1)
(V1 = 1, V2 = 2, V3 = 2)	(V1 = 2, V2 = 2, V3 = 2)	(V1 = 3, V2 = 2, V3 = 2)
(V1 = 1, V2 = 2, V3 = 3)	(V1 = 2, V2 = 2, V3 = 3)	(V1 = 3, V2 = 2, V3 = 3)
(V1 = 1, V2 = 3, V3 = 1)	(V1 = 2, V2 = 3, V3 = 1)	(V1 = 3, V2 = 3, V3 = 1)
(V1 = 1, V2 = 3, V3 = 2)	(V1 = 2, V2 = 3, V3 = 2)	(V1 = 3, V2 = 3, V3 = 2)
(V1 = 1, V2 = 3, V3 = 3)	(V1 = 2, V2 = 3, V3 = 3)	(V1 = 3, V2 = 3, V3 = 3)

Review: Probability over Feature Vectors

Example:

$$P(\{V_1 = 1, V_3 = 2\}) = \sum_{x_2 \in Dom[V_2]} P(\{V_1 = 1, V_2 = x_2, V_3 = 2\}).$$

(V1 = 1, V2 = 1, V3 = 1)	(V1 = 2, V2 = 1, V3 = 1)	(V1 = 3, V2 = 1, V3 = 1)
(V1 = 1, V2 = 1, V3 = 2)	(V1 = 2, V2 = 1, V3 = 2)	(V1 = 3, V2 = 1, V3 = 2)
(V1 = 1, V2 = 1, V3 = 3)	(V1 = 2, V2 = 1, V3 = 3)	(V1 = 3, V2 = 1, V3 = 3)
(V1 = 1, V2 = 2, V3 = 1)	(V1 = 2, V2 = 2, V3 = 1)	(V1 = 3, V2 = 2, V3 = 1)
(V1 = 1, V2 = 2, V3 = 2)	(V1 = 2, V2 = 2, V3 = 2)	(V1 = 3, V2 = 2, V3 = 2)
(V1 = 1, V2 = 2, V3 = 3)	(V1 = 2, V2 = 2, V3 = 3)	(V1 = 3, V2 = 2, V3 = 3)
(V1 = 1, V2 = 3, V3 = 1)	(V1 = 2, V2 = 3, V3 = 1)	(V1 = 3, V2 = 3, V3 = 1)
(V1 = 1, V2 = 3, V3 = 2)	(V1 = 2, V2 = 3, V3 = 2)	(V1 = 3, V2 = 3, V3 = 2)
(V1 = 1, V2 = 3, V3 = 3)	(V1 = 2, V2 = 3, V3 = 3)	(V1 = 3, V2 = 3, V3 = 3)

In these examples we are "summing out" some variables, which is also known as "marginalizing" our distribution

Review: Probability over Feature Vectors

Problem:

There is an exponential number of atomic probabilities to specify.

Requires summing up an exponential number of items.

To evaluate the probability of sets containing a particular subset of variable assignments we can do much better. Improvements come from the use of:

1. **probabilistic independence, especially conditional independence.**
2. **approximation techniques, many of which depend on distributions structured by independence.**

Review: Conditional Probability

- Before we get to conditional independence, we need to define the meaning of conditional probabilities.
- These capture conditional information, i.e. information about the influence of any one variable's value on the probability of others'.
- Conditional probabilities are essential for both representing and reasoning with probabilistic information.

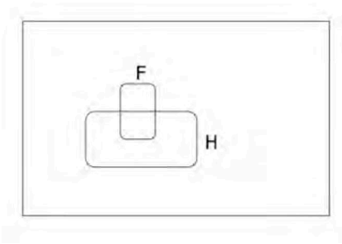
Review: Conditional Probability

- Say that A is a set of events such that P(A=a) > 0.
- Then one can define a conditional probability w.r.t. the probability that A=a:

$$P(B=b|A=a) = P(B=b,A=a)/P(A=a)$$

Review: Conditional Probability

P(A=a|B=b) refers to the fraction of worlds in which B=b that also have A=a. An example:



P(Headache=true) = 1/10

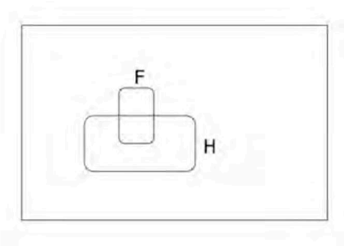
P(Flu=true) = 1/40

P(Headache=true|Flu=true) = 1/2

Headaches are rare and having flu is rarer. But, given flu, there is a 50/50 chance you have a headache.

Review: Conditional Probability

P(Headache=true|Flu=true) represents the fraction of flu-infected worlds in which you have a headache.



= # worlds with flu and headache/#worlds with flu

= area of flu and headache/area of flu

= P(Headache=true,Flu=true)/P(Flu=true)

Review: Conditional Probability

A conditional probability is a also probability function, but now over a *subset* of events in the universe instead of over the entire universe. Similar axioms hold:

P(A|A) = 1

P(B|A) ∈ [0,1]

P(C ∪ B|A) = P(C|A) + P(B|A) – P(C ∩ B|A)

Review: Independence

Probability density is a measure of likelihood. Assume you pick an element at random from U. Density (i.e. the value of P(B) is a measure as to how likely is it to also be in set B.

It could be that the density (i.e. likelihood) of B given A is **identical** to its density (or likelihood) in U.

Alternately, the density of B given A could be very **different** that its density (or likelihood) in U.

In the first case we say that B is **independent** of A. While in the second case B is **dependent** on A.

Review: Independence

A and B are **independent** properties:

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

A and B are **dependent**:

$$P(B|A) \neq P(B)$$

Review: Conditional Independence

Say that we have picked an element from U. Then we find out that this element has property A (i.e., is a member of the set A).

- Does this tell us anything more about how likely it is that the element also has property B?
- If B is independent of A then we have learned nothing new about the likelihood of the element being a member of B.

Review: Conditional Independence

E.g., say we have a feature vector, we don't know which one. We then find out that it contains the feature $V_1 = a$.

- i.e., we know that the vector contains $V_1 = a$ and is therefore a member of the set $\{V_1 = a\}$.
- Does this tell us anything about whether or not $V_2=a$, $V_3=c$, …, etc.?
- This depends on whether or not these features are independent/dependent of $V_1=a$.

Review: Conditional Independence

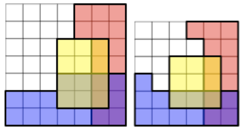
If $P(V_1|V_2=b,V_3=c) = P(V_1|V_2=b)$, we have not gained any additional information about V_1 from knowing $V_3=c$.

In this case we say that V_1 is conditionally independent of V_3 *given* V_2.

That is, once we know V_2, additionally knowing V_3 is irrelevant (it will give us no more information as to the value of V_1).

Note we could have $P(V_1|V_3=c) \neq P(V_1)$. But once we learn $V_2=b$, the value of V_3 becomes irrelevant.

Review: Conditional Independence



These pictures represent the probabilities of event sets 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 C because:

$$P(A \wedge B|\ C) = P(A|C)P(B|C)$$

BUT A and B are NOT conditionally independent given ¬C, as:

$$P(A \wedge B|\neg C) \neq P(A|\neg C)P(B|\neg C)$$

Review: Variable Independence

Note in our class, we generally want to deal with situations where we have *variables* that are conditionally independent (i.e. the variables are independent of one another). This is subtly different than asking if different sets of events are independent.

Variables X and Y are conditionally independent *given variable Z* if and only if $\forall$ x,y,z. $x \in Dom(X) \wedge y \in Dom(Y) \wedge z \in Dom(Z)$:

X=x is conditionally independent of Y=y given Z = z i.e.

$$P(X=x \wedge Y=y|Z=z) = P(X=x|Z=z) * P(Y=y|Z=z)$$

Can apply to sets of more than two variables.

Computational Impact

We will soon see in more detail how independence allows us to speed up computations related to inference. But the fundamental insight is that

If A and B are independent properties then

$$\mathbf{P(A \wedge B) = P(B) * P(A)}$$

Proof:

Computational Impact

We will soon see in more detail how independence allows us to speed up computations related to inference. But the fundamental insight is that

If A and B are independent properties then

$P(A \wedge B) = P(B) * P(A)$

Proof:

$P(B|A) = P(B)$ (def'n of independence)
$P(A \wedge B)/P(A) = P(B)$
$P(A \wedge B) = P(B) * P(A)$

Computational Impact

- Independence property allows us to "break" up the computation of a conjunction "$P(A \wedge B)$" into two separate computations "$P(A)$" and "$P(B)$".
- Dependent on how we express our probabilistic knowledge this can yield great computational savings.

Computational Impact

Similar results hold for conditional independence. If B and C are conditionally independent given A, then

$P(B \wedge C|A) = P(B|A) * P(C|A)$

Proof:

Computational Impact

Similar results hold for conditional independence. If B and C are conditionally independent given A, then

$P(B \wedge C|A) = P(B|A) * P(C|A)$

Proof:
$P(B|C \wedge A) = P(B|A)$ (def'n of conditional independence)
$P(B \wedge C \wedge A)/P(C \wedge A) = P(B \wedge A)/P(A)$
$P(B \wedge C \wedge A)/P(A) = P(C \wedge A)/P(A) * P(B \wedge A)/P(A)$
$P(B \wedge C|A) = P(B|A) * P(C|A)$.

Computational Impact

As with independence, conditional independence allows us to break up our computation onto distinct parts

$P(B \wedge C|A) = P(B|A) * P(C|A)$

It also allows us to ignore certain pieces of information during computations

$P(B|A \wedge C) = P(B|A)$

Review: Chain Rule

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

Proof:

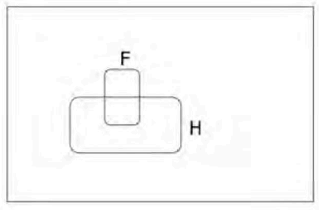
Review: Chain Rule

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

Proof:

$P(A_1| A_2 \wedge \ldots \wedge A_n) * P(A_2| A_3 \wedge \ldots \wedge A_n)$
$* \ldots * P(A_{n-1}| A_n)$
$= P(A_1 \wedge A_2 \wedge \ldots \wedge A_n)/ P(A_2 \wedge \ldots \wedge A_n) *$
$P(A_2 \wedge \ldots \wedge A_n)/ P(A_3 \wedge \ldots \wedge A_n) * \ldots *$
$P(A_{n-1} \wedge A_n)/P(A_n) * P(A_n)$

Back to Flu World



P(Headache=true) = 1/10

P(Flu=true) = 1/40

P(Headache=true|Flu=true) = 1/2

Headaches are rare and having flu is rarer. But, given flu, there is a 50/50 chance you have a headache.

What is P(Flu=true|Headache=true)?

What we just did

We Derived Bayes' Rule.

$$P(Y|X) = P(X|Y)P(Y)/P(X)$$

$$
\begin{aligned}
P(Y|X) &= P(Y \wedge X)/P(X) \\
&= P(Y \wedge X)/P(X) * P(Y)/P(Y) \\
&= P(Y \wedge X)/P(Y) * P(Y)/P(X) \\
&= P(X|Y)P(Y)/P(X)
\end{aligned}
$$

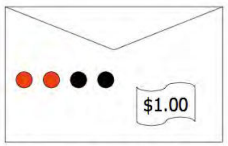
What we just did, more formally

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

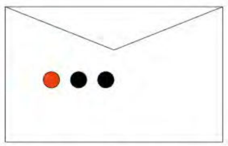
This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

Using Bayes Rule to gamble



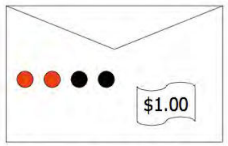
The "Win" envelope has a dollar and four beads in it



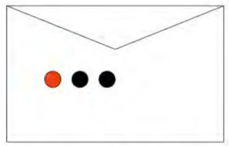
The "Lose" envelope has three beads and no money

Trivial question: Someone picks an envelope and random and asks you to bet as to whether or not it holds a dollar. What are your odds?

Using Bayes Rule to gamble



The "Win" envelope has a dollar and four beads in it



The "Lose" envelope has three beads and no money

Not trivial question: Someone lets you take a bead out of the envelope before you bet. If it is black, what are your odds? If it is red, what are your odds?

Using Bayes Rule

Note that for Bayes Rule to work requires knowledge of several probabilities:

P(Heart Disease | High Cholesterol)
= P(High Cholesterol | Heart Disease)
  * P(Heart Disease)/P(High Cholesterol)

We will return to this later.

Review: Joint Distributions

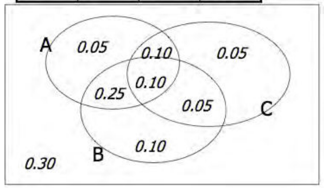
As we discussed, the **joint distribution** records the probabilities that variables will hold particular values.

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.

Normalization involves converting raw counts of data in a table into a legal probability distribution (i.e. into a distribution that sums to 1).

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



Review: Normalizing

To **normalize** a vector of k numbers or a column in our table, e.g., <3, 4, 2.5, 1, 10, 21.5> we must sum them and divide each number by the sum:

3 + 4 + 2.5 +1 +10 + 21.5 = 42

Normalized vector:
= <3/42, 4/42, 2.5/42, 1/42, 10/42, 21.5/42>
= <0.071, 0.095, 0.060, 0.024, 0.238, 0.512>

After normalizing the vector of numbers sums to 1

It therefore can be used to specify a probability 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

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

Note: these probabilities are from the UCI "Adult" Census, which you, too, can fool around with in your leisure ....

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

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

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

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

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

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