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)
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:

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

Note however that B and C are not independent as

$$P(B \land C) \neq P(B) \times P(C)$$ in either picture.

Also note that

$$P(A \land B|C) = P(A|C) \times P(B|C)$$ but A and B are NOT conditionally independent given membership in the set $\leftarrow C$, as:

$$P(A \land B| \leftarrow C) \neq P(A| \leftarrow C) \times P(B| \leftarrow C)$$
Say that $B_1, B_2, \ldots, B_k$ partition of the universe $U$ and say that each $B_i$ is defined by a particular value being assigned to a variable ($V_2 = b_i$).

$B_i \cap B_j = \emptyset$, where $i \neq j$ (mutually exclusive)

$B_1 \cup B_2 \cup B_3 \ldots \cup B_k = U$ (exhaustive)

In probabilities:

$P(B_i \cap B_j) = 0$  $P(B_1 \cup B_2 \cup B_3 \ldots \cup B_k) = 1$
Given another set of events $A$ we know that
\[ P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_k) \]
We can write this as conditional probabilities:
\[ P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_k)P(B_k) \]
\[ P(A|B_i)P(B_i) = P(A \cap B_i)/P(B_i) \times P(B_i) = Pr(A \cap B_i) \]
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.

\[ P(A_1 \land A_2 \land \ldots \land A_n) = \]
\[ P(A_1|A_2 \land \ldots \land A_n) \times P(A_2|A_3 \land \ldots \land A_n) \times \ldots \times P(A_{n-1}|A_n) \times P(A_n) \]
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.

\[
P(A_1 \land A_2 \land \ldots \land A_n) = \\
P(A_1 | A_2 \land \ldots \land A_n) \times P(A_2 | A_3 \land \ldots \land A_n) \times \ldots \times P(A_{n-1} | A_n) \times P(A_n)
\]

Proof:

\[
P(A_1 | A_2 \land \ldots \land A_n) \times P(A_2 | A_3 \land \ldots \land A_n) \times \ldots \times P(A_{n-1} | A_n) \times P(A_n)
= \frac{P(A_1 \land A_2 \land \ldots \land A_n)}{P(A_2 \land \ldots \land A_n)} \times \frac{P(A_2 \land \ldots \land A_n)}{P(A_3 \land \ldots \land A_n)} \times \ldots \times \frac{P(A_{n-1} \land A_n)}{P(A_n)} \times P(A_n)
\]
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:

\[
Pr(Y|X) = \frac{Pr(Y \land X)}{Pr(X)}
= \frac{Pr(Y \land X)/Pr(X) \ast P(Y)/P(Y)}{Pr(X)}
= \frac{Pr(Y \land X)/Pr(Y) \ast Pr(Y)/Pr(X)}{Pr(X)}
= Pr(X|Y)Pr(Y)/Pr(X)
\]
Diseases ∈ 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:

$$P(malaria|fever) = P(fever|malaria) \times P(malaria)/P(fever)$$
Using Bayes’ rule - Malaria example

- \( \Pr(\text{malaria}|\text{fever}) = \Pr(\text{fever}|\text{malaria})\Pr(\text{malaria})/\Pr(\text{fever}) \)

- \( \Pr(\text{malaria})? \)
  - 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.

- \( \Pr(\text{fever}|\text{malaria})? \)
  - 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

- \( Pr(\text{fever})? \)
  - This is typically not known, but it can be computed!
  - We eventually have to divide by this probability to get the final answer:
    \[
    Pr(\text{malaria}|\text{fever}) = Pr(\text{fever}|\text{malaria})Pr(\text{malaria})/Pr(\text{fever})
    \]

- 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.
  - \( Pr(\text{fever} \mid \neg \text{malaria} \land \neg \text{cold} \land \neg \text{flu}) = 0 \)
    Fever can't happen with one of these causes.
  - \( Pr(\text{malaria} \land \text{cold}) = Pr(\text{malaria} \land \text{flu}) = Pr(\text{cold} \land \text{flu}) = 0 \)
    these causes can't happen together. (Note that our example is not very realistic!)

- Second, we compute: \( Pr(\text{fever} | \text{malaria})Pr(\text{malaria}), \)
  \( Pr(\text{fever} | \text{cold})Pr(\text{cold}), \) \( Pr(\text{fever} | \text{flu})Pr(\text{flu}). \)
  - We know \( Pr(\text{fever} | \text{cold}) \) and \( Pr(\text{fever} | \text{flu}), \) along with \( Pr(\text{cold}) \) and \( Pr(\text{flu}) \) from the same sources as \( Pr(\text{fever} | \text{malaria}) \) and \( Pr(\text{malaria}). \)
Using Bayes’ rule - Malaria example

- Since flu, cold and malaria are exclusive, flu, cold, malaria, ¬ malaria ∧ ¬cold ∧ ¬flu forms a partition of the universe. So:
  \[ Pr(\text{fever}) = Pr(\text{fever} | \text{malaria}) \times Pr(\text{malaria}) + Pr(\text{fever} | \text{cold}) \times Pr(\text{cold}) + Pr(\text{fever} | \text{flu}) \times Pr(\text{flu}) + Pr(\text{fever} | ¬ \text{malaria} \land ¬ \text{cold} \land ¬ \text{flu}) \times Pr(¬ \text{malaria} \land ¬ \text{cold} \land ¬ \text{flu}) \]

- 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, \( Pr(\text{malaria} | \text{fever}) \), \( Pr(\text{cold} | \text{fever}) \), \( Pr(\text{flu} | \text{fever}) \), we compute the trio of numbers \( Pr(\text{fever} | \text{malaria}) \times Pr(\text{malaria}) \), \( Pr(\text{fever} | \text{cold}) \times Pr(\text{cold}) \), \( Pr(\text{fever} | \text{flu}) \times Pr(\text{flu}) \)

- And then we divide these three numbers by \( Pr(\text{fever}) \).
  - That is we divide these three numbers by their sum: This is called normalizing the numbers.

- Thus we never need actually compute \( Pr(\text{fever}) \) (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:
  \[ \frac{3 + 4 + 2.5 + 1 + 10 + 21.5}{42} = 42 \]

- Normalized vector:
  \[ \langle \frac{3}{42}, \frac{4}{42}, \frac{2.5}{42}, \frac{1}{42}, \frac{10}{42}, \frac{21.5}{42} \rangle = \langle 0.071, 0.095, 0.060, 0.024, 0.238, 0.512 \rangle \]

- After normalizing the vector of numbers sums to 1

- Exactly what is needed for these numbers to specify a probability distribution.
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

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Venn diagram showing the probability distribution of events A, B, and C.
Using the joint distribution

Using the Joint

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

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
</tbody>
</table>

\[ P(\text{Poor Male}) = 0.4654 \]

\[ P(E) = \sum_{\text{rows matching } E} P(\text{row}) \]
Using the joint distribution

Using the Joint

\[ P(Poor) = 0.7604 \]

\[ P(E) = \sum \text{P(row)} \text{ rows matching } E \]
Using the joint distribution

Inference with the Joint

\[ P(E_1 \mid E_2) = \frac{P(E_1 \land E_2)}{P(E_2)} = \frac{\sum P(\text{row})}{\sum P(\text{rows matching } E_2)} \]
Using the joint distribution

Inference with the Joint

\[
P(E_1 \mid E_2) = \frac{P(E_1 \land E_2)}{P(E_2)} = \frac{\sum \text{P(row) matching } E_1 \text{ and } E_2}{\sum \text{P(row) matching } E_2}
\]

\[P(\text{Male} \mid \text{Poor}) = \frac{0.4654}{0.7604} = 0.612\]
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 $P(X|Y \land Z) = P(X|Y)$

Then: $P(Z|Y \land X) = P(Z|Y)$

Proof:

$P(Z|X \land Y) = P(X \land Y|Z) \ast P(Z)/P(X \land Y)$ (Bayes Rule)

$= P(X|Y \land Z) \ast P(Y|Z) \ast P(Z)/P(X|Y) \ast P(Y)$ (Chain Rule)

$= P(X|Y) \ast P(Y|Z) \ast P(Z)/P(X|Y) \ast P(Y)$ (By Assumption)

$= P(Y|Z) \ast P(Z)/P(Y) = P(Z|Y)$ (Bayes Rule)