Refresher on Probability Theory

Brandon Malone

Much of this material is adapted from Chapters 2 and 3 of Darwiche's book

January 16, 2014

Primitives

The following assumes we have variables *Earthquake*(E), *Burglary* (B) and *Alarm* (A). All variables are binary.

Atoms. $E = e_1$, $E = e_2$, $A = a_2$, $\ldots$.

Operators. $\neg$, $\wedge$, $\vee$ ($\implies$, $\iff$)

Sentences or **Events**. An atom is an event.
If α and β are events, then the following are also events.

- $\neg\alpha$
- $\alpha \wedge \beta$
- $\alpha \vee \beta$

Definitions

Instantiations. An assignment of (unique) values to some variables. $E = e_1$, $A = a_2$.

Worlds, ω_i. An instantiation which includes all variables. $E = e_1$, $B = b_1$, $A = a_2$.

The set of all worlds (*i.e.*, the set of complete, unique instantiations) is denoted by Ω.

If event α is $\texttt{true}$ in ω_i, then $\omega_i \models \alpha$.

$\text{Models}(\alpha) := \{\omega_i : \omega_i \models \alpha\}$

Definitions and Identities

Consistent. $\text{Models}(\alpha) \neq \emptyset$

Valid. $\text{Models}(\alpha) = \Omega$

$$\text{Models}(\alpha \wedge \beta) = \text{Models}(\alpha) \cap \text{Models}(\beta)$$

$$\text{Models}(\alpha \vee \beta) = \text{Models}(\alpha) \cup \text{Models}(\beta)$$

$$\text{Models}(\neg\alpha) = \overline{\text{Models}(\alpha)}$$

Degrees of belief

We attach a probability to each ω_i such that

$$\sum_{\omega_i \in \Omega} Pr(\omega_i) = 1.$$

Then, our belief in event α is

$$Pr(\alpha) := \sum_{\omega_i \models \alpha} Pr(\omega_i).$$

Degrees of belief - Simple example

world	Earthquake	Burglary	Alarm	$Pr(\cdot)$
ω_1	T	T	T	0.0190
ω_2	T	T	F	0.0010
ω_3	T	F	T	0.0560
ω_4	T	F	F	0.0240
ω_5	F	T	T	0.1620
ω_6	F	T	F	0.0180
ω_7	F	F	T	0.0072
ω_8	F	F	F	0.7128

What is $Pr(Alarm = \text{T})$?

What is $Pr(Earthquake = \text{T}, Alarm = \text{F})$?

This is called a **joint probability distribution**.

Updating beliefs

Belief updates give a natural method for handling evidence. This is called **conditional probability**.

Say we know that β is `true`.

Then we say $Pr(\beta|\beta) = 1$ and $Pr(\neg\beta|\beta) = 0$.

The "|" means "given that". The notation $Pr(\alpha|\beta)$ means "The probability that α is `true` given that we know β is `true`."

Updating beliefs

Since we know $Pr(\neg\beta|\beta) = 0$, we will also insist that

$$Pr(\omega_i|\beta) = 0 \quad \text{for all } \omega_i \models \neg\beta.$$

Furthermore, all probability distributions must sum to one, so we know

$$\sum_{\omega_i \models \beta} Pr(\omega_i|\beta).$$

So for a given $\omega_i \models \beta$, what is $Pr(\omega_i|\beta)$?

Updating beliefs

Since we know $Pr(\neg\beta|\beta) = 0$, we will also insist that

$$Pr(\omega_i|\beta) = 0 \quad \text{for all } \omega_i \models \neg\beta.$$

Furthermore, all probability distributions must sum to one, so we know

$$\sum_{\omega_i \models \beta} Pr(\omega_i|\beta).$$

So for a given $\omega_i \models \beta$, what is $Pr(\omega_i|\beta)$?

How about $Pr(\omega_i|\beta) := \frac{Pr(\omega_i)}{Pr(\beta)}$?

Bayes' conditioning

Given some evidence β, must we explicitly compute $Pr(\omega_i|\beta)$ for every ω_i to say something about $Pr(\alpha|\beta)$?

$$Pr(\alpha|\beta) = \sum_{\omega_i \models \alpha} Pr(\omega_i|\beta)$$

Bayes' conditioning

Given some evidence β, must we explicitly compute $Pr(\omega_i|\beta)$ for every ω_i to say something about $Pr(\alpha|\beta)$?

$$
\begin{aligned}
Pr(\alpha|\beta) &= \sum_{\omega_i \models \alpha} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) + \sum_{\omega_i \models \alpha, \neg\beta} Pr(\omega_i|\beta)
\end{aligned}
$$

Bayes' conditioning

Given some evidence β, must we explicitly compute $Pr(\omega_i|\beta)$ for every ω_i to say something about $Pr(\alpha|\beta)$?

$$\begin{aligned} Pr(\alpha|\beta) &= \sum_{\omega_i \models \alpha} Pr(\omega_i|\beta) \\ &= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) + \sum_{\omega_i \models \alpha, \neg\beta} Pr(\omega_i|\beta) \\ &= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) \end{aligned}$$

Bayes' conditioning

Given some evidence β, must we explicitly compute $Pr(\omega_i|\beta)$ for every ω_i to say something about $Pr(\alpha|\beta)$?

$$
\begin{aligned}
Pr(\alpha|\beta) &= \sum_{\omega_i \models \alpha} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) + \sum_{\omega_i \models \alpha, \neg\beta} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i)/Pr(\beta)
\end{aligned}
$$

Bayes' conditioning

Given some evidence β, must we explicitly compute $Pr(\omega_i|\beta)$ for every ω_i to say something about $Pr(\alpha|\beta)$?

$$
\begin{aligned}
Pr(\alpha|\beta) &= \sum_{\omega_i \models \alpha} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) + \sum_{\omega_i \models \alpha, \neg\beta} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i)/Pr(\beta) \\
&= \frac{1}{Pr(\beta)} \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i)
\end{aligned}
$$

Bayes' conditioning

Given some evidence β, must we explicitly compute $Pr(\omega_i|\beta)$ for every ω_i to say something about $Pr(\alpha|\beta)$?

$$
\begin{aligned}
Pr(\alpha|\beta) &= \sum_{\omega_i \models \alpha} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) + \sum_{\omega_i \models \alpha, \neg\beta} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i|\beta) \\
&= \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i)/Pr(\beta) \\
&= \frac{1}{Pr(\beta)} \sum_{\omega_i \models \alpha, \beta} Pr(\omega_i) \\
Pr(\alpha|\beta) &= \frac{Pr(\alpha, \beta)}{Pr(\beta)}
\end{aligned}
$$

Bayes' conditioning class work

world	Earthquake	Burglary	Alarm	$Pr(\cdot)$
ω_1	T	T	T	0.0190
ω_2	T	T	F	0.0010
ω_3	T	F	T	0.0560
ω_4	T	F	F	0.0240
ω_5	F	T	T	0.1620
ω_6	F	T	F	0.0180
ω_7	F	F	T	0.0072
ω_8	F	F	F	0.7128

Calculate the following probabilities.

- $Pr(Alarm = \mathrm{T})$
- $Pr(Earthquake = \mathrm{T})$
- $Pr(Burglary = \mathrm{T})$
- $Pr(Burglary = \mathrm{T}, Earthquake = \mathrm{T})$
- $Pr(Burglary = \mathrm{T}, Alarm = \mathrm{T})$
- $Pr(Alarm = \mathrm{T}, Earthquake = \mathrm{T})$
- $Pr(Alarm = \mathrm{T} | Earthquake = \mathrm{T})$
- $Pr(Alarm = \mathrm{T} | Burglary = \mathrm{T})$
- $Pr(Earthquake = \mathrm{T} | Burglary = \mathrm{T})$
- $Pr(Earthquake = \mathrm{T} | Alarm = \mathrm{T})$
- $Pr(Burglary = \mathrm{T} | Alarm = \mathrm{T})$
- $Pr(Burglary = \mathrm{T} | Earthquake = \mathrm{T})$
- $Pr(Burglary = \mathrm{T} | Alarm = \mathrm{T}, Earthquake = \mathrm{T})$
- $Pr(Burglary = \mathrm{T} | Alarm = \mathrm{T}, Earthquake = \mathrm{F})$

Independence

What did knowing that *Burglary* $=$ T tell us about *Earthquake*?

Independence

What did knowing that $Burglary = \mathsf{T}$ tell us about $Earthquake$?

Nothing.

$$Pr(Earthquake = \mathsf{T}) = Pr(Earthquake = \mathsf{T} | Burglary = \mathsf{T}) = 0.1$$

So we say that $Earthquake$ and $Burglary$ are independent.

Independence defined

Events α and β are **independent** if

$$Pr(\alpha \wedge \beta) = Pr(\alpha) \cdot Pr(\beta).$$

Equivalently, α and β are independent if

$$Pr(\alpha|\beta) = Pr(\alpha).$$

Conditional independence

Are independent events *always* independent?

Conditional independence

Are independent events *always* independent?

$$Pr(Burglary = \mathsf{T}) = ?$$
$$Pr(Burglary = \mathsf{T} | Earthquake = \mathsf{T}) = ?$$
$$Pr(Burglary = \mathsf{T} | Alarm = \mathsf{T}) = ?$$
$$Pr(Burglary = \mathsf{T} | Earthquake = \mathsf{T}, Alarm = \mathsf{T}) = ?$$

Conditional independence

Are independent events *always* independent?

$$
\begin{aligned}
Pr(Burglary = \mathrm{T}) &= ? \\
Pr(Burglary = \mathrm{T} | Earthquake = \mathrm{T}) &= ? \\
Pr(Burglary = \mathrm{T} | Alarm = \mathrm{T}) &= ? \\
Pr(Burglary = \mathrm{T} | Earthquake = \mathrm{T}, Alarm = \mathrm{T}) &= ?
\end{aligned}
$$

So, no.

Note how this naturally handles the non-monotonicity problem.

Conditional independence - simple example

world	Temp	Sensor1	Sensor2	$Pr(\cdot)$
ω_1	normal	normal	normal	0.576
ω_2	normal	normal	extreme	0.144
ω_3	normal	extreme	normal	0.064
ω_4	normal	extreme	extreme	0.016
ω_5	extreme	normal	normal	0.008
ω_6	extreme	normal	extreme	0.032
ω_7	extreme	extreme	normal	0.032
ω_8	extreme	extreme	extreme	0.128

Calculate the following probabilities.

- $Pr(Sensor2 = \text{normal})$
- $Pr(Sensor2 = \text{normal} | Sensor1 = \text{normal})$
- $Pr(Sensor2 = \text{normal} | Temp = \text{normal})$
- $Pr(Sensor2 = \text{normal} | Temp = \text{normal}, Sensor1 = \text{normal})$

Conditional independence - simple example

world	Temp	Sensor1	Sensor2	$Pr(\cdot)$
ω_1	normal	normal	normal	0.576
ω_2	normal	normal	extreme	0.144
ω_3	normal	extreme	normal	0.064
ω_4	normal	extreme	extreme	0.016
ω_5	extreme	normal	normal	0.008
ω_6	extreme	normal	extreme	0.032
ω_7	extreme	extreme	normal	0.032
ω_8	extreme	extreme	extreme	0.128

Calculate the following probabilities.

- $Pr(Sensor2 = \text{normal})$
- $Pr(Sensor2 = \text{normal} | Sensor1 = \text{normal})$
- $Pr(Sensor2 = \text{normal} | Temp = \text{normal})$
- $Pr(Sensor2 = \text{normal} | Temp = \text{normal}, Sensor1 = \text{normal})$

*Sensor*1 and *Sensor*2 began dependent.

Once we **conditioned** on *Temp*, they became independent.

Conditional independence defined

Events α and β are **conditionally independent** given evidence γ if

$$Pr(\alpha, \beta | \gamma) = Pr(\alpha | \gamma) \cdot Pr(\beta | \gamma)$$

Equivalently, α and β are conditionally independent given γ if

$$Pr(\alpha | \beta, \gamma) = Pr(\alpha | \gamma)$$

We always assume the evidence γ has non-zero probability.

Conditional independence notation

Suppose we have disjoint variable sets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$.

The notation $I(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ means that $\mathbf{x}$ is independent of $\mathbf{y}$ given $\mathbf{z}$ for all instantiations of $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$.

The notation $\mathbf{X} \perp \mathbf{Y}$ means that $\mathbf{X}$ is (unconditionally) independent of $\mathbf{Y}$.

The notation $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$ means that $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$.

Chain rule

Suppose we have a large joint probability distribution, $Pr(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Can we rewrite this in some more manageable way?

Chain rule

Suppose we have a large joint probability distribution, $Pr(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Can we rewrite this in some more manageable way?

$$Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) = Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2, \ldots, \alpha_n)$$

Chain rule

Suppose we have a large joint probability distribution, $Pr(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Can we rewrite this in some more manageable way?

$$\begin{aligned} Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2, \ldots, \alpha_n) \\ &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2 | \alpha_3, \ldots, \alpha_n) \end{aligned}$$

Chain rule

Suppose we have a large joint probability distribution, $Pr(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Can we rewrite this in some more manageable way?

$$\begin{aligned}
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2, \ldots, \alpha_n) \\
&= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2 | \alpha_3, \ldots, \alpha_n) \\
&= \ldots \\
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2 | \alpha_3 \ldots, \alpha_n) Pr(\alpha_3 | \alpha_4, \ldots, \alpha_n) \ldots Pr(\alpha_{n-1} | \alpha_n) Pr(\alpha_n)
\end{aligned}$$

Chain rule

Suppose we have a large joint probability distribution, $Pr(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Can we rewrite this in some more manageable way?

$$
\begin{aligned}
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2, \ldots, \alpha_n) \\
&= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2 | \alpha_3, \ldots, \alpha_n) \\
&= \ldots \\
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2 | \alpha_3 \ldots, \alpha_n) Pr(\alpha_3 | \alpha_4, \ldots, \alpha_n) \ldots Pr(\alpha_{n-1} | \alpha_n) Pr(\alpha_n)
\end{aligned}
$$

This is called the **chain rule**.

Chain rule

Suppose we have a large joint probability distribution, $Pr(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Can we rewrite this in some more manageable way?

$$
\begin{aligned}
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2, \ldots, \alpha_n) \\
&= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2 | \alpha_3, \ldots, \alpha_n) \\
&= \ldots \\
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1 | \alpha_2, \ldots, \alpha_n) Pr(\alpha_2 | \alpha_3 \ldots, \alpha_n) Pr(\alpha_3 | \alpha_4, \ldots, \alpha_n) \ldots Pr(\alpha_{n-1} | \alpha_n) Pr(\alpha_n)
\end{aligned}
$$

This is called the **chain rule**.

What if $I(\alpha_1, \{\alpha_2\}, \{\alpha_3, \ldots \alpha_n\})$?

Can we rearrange the order of the αs?

Chain rule

Suppose we have a large joint probability distribution, $Pr(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Can we rewrite this in some more manageable way?

$$
\begin{aligned}
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1|\alpha_2, \ldots, \alpha_n)Pr(\alpha_2, \ldots, \alpha_n) \\
&= Pr(\alpha_1|\alpha_2, \ldots, \alpha_n)Pr(\alpha_2|\alpha_3, \ldots, \alpha_n) \\
&= \ldots \\
Pr(\alpha_1, \alpha_2, \ldots, \alpha_n) &= Pr(\alpha_1|\alpha_2, \ldots, \alpha_n)Pr(\alpha_2|\alpha_3 \ldots, \alpha_n)Pr(\alpha_3|\alpha_4, \ldots, \alpha_n) \ldots Pr(\alpha_{n-1}|\alpha_n)Pr(\alpha_n)
\end{aligned}
$$

This is called the **chain rule**.

What if $I(\alpha_1, \{\alpha_2\}, \{\alpha_3, \ldots \alpha_n\})$?

Can we rearrange the order of the αs?

Efficient inference in Bayesian networks stems from these operations.

Marginalization

How did we calculate $Pr(Alarm = \mathrm{T})$?

Marginalization

How did we calculate $Pr(Alarm = \mathrm{T})$?

Implicitly, we summed over all instantiations of the other variables.

This is called **marginalization**.

$$Pr(\alpha) = \sum_{i=1}^{n} Pr(\alpha|\beta_i)Pr(\beta_i), \text{ where } \beta \text{ has } n \text{ distinct instantiations}$$

Marginalization

How did we calculate $Pr(Alarm = \mathrm{T})$?

Implicitly, we summed over all instantiations of the other variables.

This is called **marginalization**.

$$Pr(\alpha) = \sum_{i=1}^{n} Pr(\alpha|\beta_i)Pr(\beta_i), \text{ where } \beta \text{ has } n \text{ distinct instantiations}$$

Among other things, this will be useful for handling hidden variables.

If β is a continuous variable, we can replace the sum with an integral.

Bayes' rule

Suppose α is a disease and β is the result of a test. Given the result of the test, what is the probability a person has the disease?

Bayes' rule

Suppose α is a disease and β is the result of a test. Given the result of the test, what is the probability a person has the disease?

$$Pr(\alpha|\beta) = Pr(\alpha, \beta)/Pr(\beta)$$
$$Pr(\alpha|\beta) = Pr(\beta|\alpha)Pr(\alpha)/Pr(\beta)$$

This is called **Bayes' rule**. It forms the basis for reasoning about causes given their effects.

Class work

Suppose we have a patient who was just tested for a particular disease and the test came out positive. We know that one in every thousand people has this disease. We also know that the test is not perfect. It has a false positive rate of 2% and a false negative rate of 5%. That is, the test result is positive when the patient does not have the disease 2% of the time, and the result is negative when the patient has the disease 5% of the time. What is the probability that the patient with the positive test result actually has the disease?

Let D stand for "the patient has the disease," and T stand for "the test result." That is, what is $P(D = \mathrm{T} | T = \mathrm{T})$?

Recap

During this class, we discussed

- Basic terminology and definitions for discussing propositional events and reasoning about them probabilistically
- Fundamental properties of joint probability distributions
- Rigorous methods to incorporate evidence and construct conditional probability distributions
- Independence and conditional independence
- Chain rule, marginalization and Bayes' rule

Next time, in probabilistic models...

- A formal introduction to Bayesian networks
- Graphical structures comprising Bayesian networks
- Independence assertions based on the BN structure
- Equivalence among BN structures
- Factorized joint probability distributions