Lecture 12: Probability and Bayesian inference

CAB203 Discrete Structures

Matthew McKague

Queensland University of Technology

matthew.mckague@qut.edu.au

Outline

Probability

Conditional Probability

Bayes rule

Bayesian inference

A little bit of decision theory

Readings

Some books if you want to learn a little more:

- *Introduction to probability* Josep Blitzstein, Jessica Hwang
- *Bayesian statistics the fun way* Will Kurt.

Both books (along with the entire O'Reilly catalog) are available via QUT's subscription. Access via QUT library: https://secure.qut.edu.au/library/resources/current/databases/cou/oreilly.php

Outline

Probability

Conditional Probability

Bayes rule

Bayesian inference

A little bit of decision theory

Probability

Probability is a number that we assign to some event that quantifies how likely it is to happen, or the chances of it happening.

- But what does "likely" or "chances" mean?

Frequentist approach

The *frequentist* approach assigns probability by how many times something actually occurs:

- The probability of an event E is written as $P(E)$
- If we do the same process n times (where n is large), and E occurs about m times then $P(E) \approx m/n$.
- E.g. tossing a coin 1000000 times we see heads come up 500000 times, so the probability is $\approx 1/2$

The frequentist approach is problematic for cases where something can only occur once by definition, e.g. what is the probability that some particular political party wins the election in 2050?

Subjectivity of probability

Consider:

- Alice flips a fair coin, looks at it and then covers it.
- Bob didn't see the coin.
- Alice knows the coin is heads up. For her $P(\mathit{Heads}) = 1$.
- Bob knows nothing about the coin. For him $P(\mathit{Heads}) = 1/2$.

Probability is about *information*. In most situations, with enough information, the outcome is certain.

This is not true in quantum physics, where some things are inherently random.

Sample spaces

A *sample space* is the set of all possible outcomes for some observation.

- You can think of it as all possible states of some system that we are investigating
- Example: for a single coin toss the sample space is $\{H, T\}$
- Example: for a single 6-sided die roll the sample space is $\{1, 2, 3, 4, 5, 6\}$

Event

An *event* is a subset of the sample space.

- Example: for 6-sided die toss, the event corresponding to an odd number coming up would be

$$\{1, 3, 5\}$$

We can form events however we like and apply set theoretic operations ($\cap, \cup, \setminus, \ldots$) to combine them.

Probability function

Given a sample space S, a *probability function* or *probability distribution* is a function $P : \mathcal{P}(S) \to \mathbb{R}$ from events to real numbers such that:

- $0 \leq P(E) \leq 1$ for all events $E \subseteq S$
- $P(S) = 1$ and $P(\emptyset) = 0$
- If $E_1, \ldots, E_n$ are all disjoint events (i.e. $E_j \cap E_k = \emptyset$ whenever $j \neq k$) then
$$\sum_{i=1}^{n} P(E_i) = P\left(\bigcup_{i=1}^{n} E_i\right)$$
- For an outcome $s \in S$ we will write $P(s)$ as a shorthand for $P(\{s\})$

You will often see $P(A, B)$ for events A, B which means $P(A \cap B)$.
(Note: using $\mathcal{P}(\mathcal{S})$ for the power set of S here.)

Joint distributions

- Given two state spaces S and T we can form a lager state space $S \times T$
- $S \times T$ contains all possible combinations of outcomes for S and T simultaneously.
- A probability distribution on $S \times T$ is called a *joint distribution*
- Given $E \subseteq S$ we often silently lift it to an event on $S \times T$:

$$\{(s, t) : s \in E, t \in T\}$$

- If $P(E, F) = P(E)P(F)$ for all $E \subseteq S$ and $F \subseteq T$ then we say that P is a *product distribution*

Two coins

Given two coins, we can ask about the joint probabilities for tossing them at the same time.

- $S = \{h, t\}$ state space for first coin
- $T = \{h, t\}$ state space for second coin
- $S \times T = \{(h, h), (h, t), (t, h), (t, t)\}$ is the state space for both coins
- Normal coins do not influence each other, so we would have a product distribution on $S \times T$.

Outline

Conditional probability

The *conditional probability* of event A given even B is given by:

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

This gives a new probability function with B as the state space. Event A is interpreted as $A \cap B$ on this new space.

Interpretations of conditional probability

$P(A|B)$ can be viewed as:

- The probability of A occurring, assuming that B has already occurred
- The credence that I should assign to A after receiving information B

Conditional probability example

Let A be the event that it will rain today. Let B be the event that it will be sunny today. Compare:

- $P(A)$: How likely is it to rain today?
- $P(A|B)$: How likely is it to rain today, given that it is going to be sunny today?
- $P(B|A)$: How likely is it to be sunny today, given that it will rain?
- $P(A, B)$: How likely is it both to be sunny and to rain today?

Uses of probability

Probability forms the basis for many other theories and is used in many applications. Some examples:

- Statistics
- Decision theory
- Game theory
- Economics
- Analysing algorithms
- Data science
- Lotteries, gambling, betting
- Predictions of election outcomes, stock market prices
- Medical decisions
- Scientific processes in general (e.g. evaluating evidence for hypotheses)

Outline

Bayesian approach to probabilities

The Bayesian approach is:

- Probabilities represent extent of belief, likelihood, or credence that an event will happen
- Related to what odds you are willing to take on a bet
- Focus on most rational ways of updating probabilities based on new information using *Bayes rule*

Deriving Bayes' rule

By definition we have:

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

Rearrange to get:

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

Similarly:

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

Right hand sides are the same! Equate the left sides and get...

Deriving Bayes' rule (2)

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

rearrange once more to get Bayes' rule!

Lemma (Bayes' rule)

Let P be a probability distribution and let A, B be events. Then

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

Bayes' rule tells us how the probability of B changes when A is observed.

Bayes' rule example

Suppose that a test for some disease has a *sensitivity* of s and a *false positive rate* of f:

- Let T_+ be the event of a positive test outcome
- Let D be the event of having the disease
- $P(T_+|D) = s$
- $P(T_+|\overline{D}) = f$

What is the probability that you have the disease?

Base rate fallacy

Suppose that the sensitivity is $s = 90\%$, the false positive rate is $f = 2\%$, and your test is positive. What is the probability that you have the disease?

We don't have enough information to say!

- Most people (and many doctors!) guess that the chances of you having the disease is around 90%.
- This is called the *base rate fallacy*: not taking into account *how prevalent the disease is*.

Bayes to the rescue!

We can calculate the probability of having the disease, given a positive test result:

$$
\begin{aligned}
P(D|T_+) &= \frac{P(T_+|D)P(D)}{P(T_+)} \\
&= \frac{sP(D)}{P(T_+)}
\end{aligned}
$$

- $P(D)$ is how prevalent the disease is.
- What is $P(T_+)$?
- We could measure directly, or...

Bayes rescue in progress...

With S the entire sample space:

$$
\begin{aligned}
P(T_+) &= P(T_+ \cap S) \\
&= P(T_+ \cap (D \cup \overline{D}) \\
&= P((T_+ \cap D) \cup (T_+ \cap \overline{D})) \\
&= P(T_+, D) + P(T_+, \overline{D}) \\
&= P(T_+|D)P(D) + P(T_+|\overline{D})(1 - P(D)) \\
&= sP(D) + f(1 - P(D)) \\
&= P(D)(s - f) + f
\end{aligned}
$$

Sub in with s and f:

$$P(D|T_+) = \frac{sP(D)}{P(D)(s - f) + f}$$

Try it with some values

Suppose $P(D) = 0.1$. Then

$$P(D|T_+) = \frac{0.90 \times 0.1}{0.1 \times 0.88 + 0.02} \approx 0.83$$

Suppose $P(D) = 0.001$. Then

$$P(D|T_+) = \frac{0.90 \times 0.001}{0.001 \times 0.88 + 0.02} \approx 0.043$$

For exactly the same test and result, the probability of having the disease could be very likely or very unlikely depending on the prevalence of the disease.

Notice some things...

- We can often control when certain events happen, e.g. we can select who we give a test to.
- Sometimes it is easier to estimate the conditional probability. E.g. you can give the test to a bunch of people with the disease to find $P(T_+|D)$ without learning $P(D)$.
- There may be *other* information that affects $P(D)$ for some particular person without affecting $P(T_+|D)$, e.g. whether they have been exposed to people with the disease. Really we should have $P(D|\mathit{previous\ information})$
- What we are actually getting is $P(D|\mathit{previous\ information}, T_+)$

Outline

Bayseian inference

Bayesian inference is about updating probabilistic models of the world based on new information.
Why should we care about this?

- No knowledge is certain. All of your knowledge has some probability of being true less than 1.
- New information is always incoming: from our senses, other people, science, etc.
- How can we make sure our beliefs most accurately reflect the information that we learn?

An example

Suppose we have some scenario: a particular coin is either fair (50% chance of heads) or biased (70% chance of heads) but you're not sure which. We can write down some information:

- If the coin is unbiased then $P(H) = 0.5$, $P(T) = 0.5$
- If the coin is biased then $P(H) = 0.7$, $P(T) = 0.3$

To capture our lack of knowledge about the coin we can write:

- U means the coin is unbiased, B means the coin is biased
- We think the coin is probably not biased, so maybe $P(U) = 0.9$ $P(B) = 0.1$

Adjusting our notation

U and B are now events/outcomes so we can rewrite our knowledge like so:

- $P(H|U) = 0.5$, $P(T|U) = 0.5$
- $P(H|B) = 0.7$, $P(T|B) = 0.3$
- $P(U) = 0.9$, $P(B) = 0.1$

Here we *start* with the conditional probabilities for H and T! There is a joint distribution, but we don't really need it. We'll just use the conditionals.

New information

So you have a model:

The likelihood that the coin is biased is 0.1

We learn new information:

The coin comes up heads.

How should we update our model?

What is the probability that the coin is biased, given our current model and the observed event?

Biased?

Bayes rule gives us:

$$P(B|H) = \frac{P(H|B)P(B)}{P(H)} = \frac{0.7 \times 0.1}{0.52} \approx 0.13$$

where

$$\begin{aligned} P(H) &= P(H|B)P(B) + P(H|U)P(U) \\ &= 0.7 \times 0.1 + 0.5 \times 0.9 \\ &= 0.52 \end{aligned}$$

This is our new model given the additional information:

The likelihood that the coin is biased is 0.13.

Predicting the future

Given our updated model for the coin, we can calculate the probability of a heads on the next coin toss:

$$\begin{aligned} P(H) &= P(H|B)P(B) + P(H|U)P(U) \\ &= 0.7 \times 0.13 + 0.5 \times (1 - 0.13) \\ &= 0.526 \end{aligned}$$

Formalising Bayesian inference

Suppose that we have a number of disjoint hypotheses: $H_1, \ldots, H_n$.

- The *prior probability* for H_j is our current model (probability distribution) for how likely H_j is: $P(H_j)$.
- For hypothesis H_j the *likelihood function* is $P(E|H_j)$: the probability of E if hypothesis H_j is true
- For some event E the *posterior probability* is $P(H_j|E)$: the probability of H_j given evidence E
- $P(E)$ is the *marginal likelihood* of E

Formulas for formalism

$$\text{posterior probability} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

$$P(H_j|E) = \frac{P(E|H_j)P(H_j)}{P(E)}$$

The marginal likelihood $P(E)$ can be found as

$$P(E) = \sum_{j=1}^{n} P(E|H_j)P(H_j)$$

which is also the prediction for E based on the current model.

We can see $P(H_j|E)$ as the fraction of the probability of E that comes from H_j.

What about the prior?

What if you are starting from scratch? What should the prior be?

- There is no "correct" answer! Only suggested best practices
- Ideally, prior should be informed by scientific plausibility
- Uniform distribution ($1/n$ for n hypotheses) is reasonable if no other information
- No hypothesis should have prior of 0 (but it can be very very low)

The good news: given enough information, the prior doesn't really matter! (*as long as no hypothesis starts at 0*)

Implications for thinking in general

The Bayesian approach can inform less formal ways of thinking about our beliefs in light of new evidence.

- People can start with different priors and draw different conclusions from the same evidence
- We should consider, not only whether a hypothesis predicts an outcome (high $P(E|H_j)$) but whether *other* hypothesis also predict it (is $P(E)$ close to $P(E|H_j)$?)
- Extraordinary claims (with a very low prior $P(H_j)$) require extraordinary evidence (low $P(E)$, high $P(E|H_j)$)
- Unless there is some extraordinary evidence, we should make modest changes to our beliefs

Outline

Normative decision theory

Normative decision theory aims to provide optimal decisions in uncertain situations.

- One framework focuses on maximising *expected utility*
- A *utility function* $u : S \to \mathbb{R}$ assigns some numerical value to all points in the sample space.
- Given some utility function u the *expected utility* is

$$\mathcal{E}_P(u) = \sum_{s \in S} u(s)P(s)$$

- Different choices that you might make give different utility functions
- One decision theory rule says to make the choice that gives the highest expected utility

Betting

To make a bet on an event E,

- You pay some amount of money m (the *stake*) to the bookmaker.
- If E occurs, the bookmaker pays out $m \times odds$ (which includes the original stake).
- The net profit if E occurs is $m \times (odds - 1)$, otherwise the stake is lost
- If the bet is perfectly fair then $odds = 1/P(E)$

Example: making a bet

Suppose you are considering betting on a sports game.

- A means team A wins, B means team B wins
- The odds are 2.1 for A and 1.5 for B.
- Your choices are: bet on A, bet on B, don't bet.
- You estimate $P(A) = 0.40$

Utility functions (assuming stake of 1):

Choice	A	B	Expected utility
Bet on A	1.1	-1	$1.1 \times 0.4 + (-1) \times 0.6 = -0.16$
Bet on B	-1	0.5	$(-1) \times 0.4 + 0.5 \times 0.6 = -0.10$
No bet	0	0	0

The best choice is to make no bets! Any casino or bookmaker is set up to have positive utility *for them* on all bets.

Some nuance

Maximising utility is easy to misuse:

- It is usually not obvious how to measure utility except in simple cases.
- The practical effect of a bet payout or loss for a person may not be proportional to the monetary value.
- In many cases the worst or best outcome is more important than the expected outcome.
- Works best when making many many decisions so that the overall average outcome is similar to the expected outcome (e.g. casinos or bookmakers who take many many bets).