Computational Cognitive Science (2010-2011)

Notes for Lecture 10: Probability Theory

Sharon Goldwater
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh

1 What is probability and why do we care?

1.1 Probability as counting

- Most widespread (frequentist) view: probability ≈ *counting*.
  - Repeat experiment ∞ times, consider proportions of different outcomes.
    - Ex. P(die comes up 2) = 1/6.
    - Ex. P(coin comes up heads) = 1/2.
    - Ex. P(two dice total 6) = 5/36.
      (This one solved by enumerating (counting!) possible outcomes and adding up their probabilities: P(1,5) + P(2,4) + P(3,3) + P(4,2) + P(5,1), each w/ prob. 1/36, equals 5/36. Note this relies on outcomes being mutually exclusive.)
  - Useful for building intuitions; later we'll consider another (Bayesian) view: probability ≈ *belief*.
    - Ex. P(it will rain tomorrow)
    - Ex. P(the word I heard was "dog")

    These have no associated repeatable experiment (why not?)
- Most probability courses have lots of counting (combinatorics), sample spaces, Venn diagrams, etc.
  - Good theoretical grounding, perhaps useful for intutions, but less practical.
  - If you want to learn/revise these things, the optional Manning & Shütze chapter listed in the course readings covers them.
- Here: hopefully we'll get to interesting and useful things faster. Where *interesting and useful things = inference*.

1.2 Generation and Inference

Probability theory and probabilistic models are useful for two types of problems.

1. **Generation/prediction.** Reasoning from causes to effects: Given a known set of causes and knowledge about how they interact, what are likely/unlikely outcomes?
   - Ex. roll a die ⇒ predict outcome.

   (We call this *generation* because we can think of it as using the probabilistic system to "generate" outcomes. This will become more clear when we get to more complex examples later.)
2. **Inference.** Reasoning from effects to causes: Given knowledge about possible causes and how they interact, as well as some observed outcomes, which causes are likely/unlikely?

- Ex. observe many outcomes of a coin flip $\Rightarrow$ determine if coin is fair.
- Ex. observe features of an object $\Rightarrow$ determine if it is a cat or dog.
- Ex. observe patient's symptoms $\Rightarrow$ determine disease.

Notes:

- In machine learning, causes are often referred to as *hidden* or *latent* variables, and effects as *observed* variables. This is somewhat more general terminology, since it doesn't imply causation in every case.
- In everyday reasoning, we often do *both* inference and prediction. For example, under Anderson's view of categorization, we first infer the category of an object (cat or dog?) in order to predict its future behavior (purr or bark?).

2 Random variables and distributions

- A **random variable** (r.v.) is a variable that represents the outcome of a random experiment.
  - Discrete random variable: possible outcomes can be mapped to integers.
    - Ex. Outcome of fair coin flip is a *binary* r.v. X with $P(X = h) = P(X = t) = 1/2$.
    - Ex. Sum of two dice is a r.v. Y with values in $2 \ldots 12$. $P(Y = 6) = 5/36$.
  - Continuous random variable: possible outcomes can be mapped to real numbers.
    - Ex. Distance a player kicks a ball.
    - Ex. Height of a randomly chosen individual.

    We will leave these aside for now and deal with discrete r.v.'s, which are more intuitive.
- For discrete r.v., its **distribution** tells us the probability of each outcome. If the number of distinct outcomes is finite, the distribution can be represented using a table or a graph.
  - Ex. Suppose I roll a six-sided die whose faces are colored as followed: one side is red (r), two sides are green (g), and three sides are blue (b). Let Y be the color facing up. Then the distribution of Y is

$Y = r$	$Y = g$	$Y = b$
1/6	2/6	3/6

  - Note that a distribution is a *function* from outcomes to real numbers, known as the **probability mass function**.
- Ex. of infinite distribution: r.v. X representing the number of coin flips before getting a head. Must use an equation instead of a table: $P(X = n) = (1/2)^n$. (How did I get this?)
- Note that for all outcomes x_i:

$$0 \leq P(X = x_i) \leq 1 \tag{1}$$

$$\sum_{x_i} P(X = x_i) = 1 \tag{2}$$

where the sum is over all possible values of x_i.

3 Joint distributions

- To do anything interesting, we need to be able to deal with multiple variables simultaneously. The **joint distribution** over X and Y, written $P(X,Y)$, tells us the probability of each pair of outcomes (one for X, one for Y) occurring together.
  - Ex. Suppose we have a device that produces a sound (X) and a light (Y) each time we press a button. There are two different sounds (chime and tone) and three different colored lights (red, green, blue), with the following joint distribution:

	$Y=r$	$Y=g$	$Y=b$
$X=c$	0.1	0.2	0.1
$X=t$	0.1	0.2	0.3

  - Note that now $\sum_{x_i,y_i} P(X=x_i,Y=y_i)=1$. This generalizes to more variables, though it's hard to draw the tables!

- To find the distribution over a single variable, add up the probabilities in the corresponding row or column:

	$Y=r$	$Y=g$	$Y=b$	
$X=c$	0.1	0.2	0.1	0.4
$X=t$	0.1	0.2	0.3	0.6
	0.2	0.4	0.4	

  So, $P(Y=g)=0.4$, for example.
  - We can write this mathematically as the **sum rule**:

$$P(X)=\sum_Y P(X,Y) \tag{3}$$

  - Summing over one variable in a joint distribution is sometimes called *marginalizing over* or *marginalizing out* that variable, because if you represent the distributions as tables (like we are doing), the marginalized distribution appears in the *margin* of the table.

- Often it is useful to know whether two (or more) variables are **independent**. Intuitively, variables are independent if they have no influence on each other, or more generally, if knowing the value of one doesn't help us predict the value of the other. Mathematically, X and Y are independent (by definition) iff

$$P(X,Y)=P(X)P(Y) \tag{4}$$

  That is, if $P(X=x_i,Y=y_i)=P(X=x_i)P(Y=y_i)$ for all values of x_i and y_i.
  - Ex. If I roll two dice, the outcomes of the two are independent.
  - Ex. If X represents whether I stay up late, and Y represents whether I arrive to my 9 a.m. class on time, then X and Y are presumably not independent.
  - Ex. In the joint distribution table above, X and Y are not independent, because multiplying together the numbers in the margins does not yield the numbers inside the table.

4 Conditional probabilities and Bayes' rule

4.1 Definitions

- Sometimes we want to talk about the probability that one variable has a particular outcome, given that the outcome of another variable is already known.
  - Ex. If I plug your ears but you see the red light in my button-pressing machine ($Y = r$), what is the probability that chime went off ($X = c$)?
  - Intuitively, narrow the world down to those cases where $Y = r$, and consider the probability of $X = c$ within them, i.e. consider $P(X = c, Y = r)$ as a proportion of $P(Y = r)$.
- This is known as the **conditional probability** of X given Y, defined as

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{5}$$

  - So, $P(X = c|Y = r) = 0.1/0.2 = 0.5$.
  - Note that for any value of y_i, $\sum_{x_i} P(X = x_i|Y = y_i) = 1$. (Can you see why?)
  - Another way to show that two variables are independent is to show that $P(X|Y) = P(X)$. (Can you prove this?)
- We can rewrite this definition to get the **product rule**:

$$P(X,Y) = P(X|Y)P(Y) \tag{6}$$

- Putting together the product rule and the sum rule gives us the **rule of total probability**:

$$P(X) = \sum_{Y} P(X|Y)P(Y) \tag{7}$$

- Note that the product rule can also be written as $P(X,Y) = P(Y|X)P(X)$, so we know that $P(X|Y)P(Y) = P(Y|X)P(X)$. Rewrite, and we have **Bayes' Rule**:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{8}$$

4.2 Fun and profit

- Now we can actually start to do inference!
  - Ex. Suppose in an experiment on human memory, participants have to memorize two types of items: words (w) and pictures (p). Using X to represent the type of item, the experiment is designed so that $P(X = w) = 0.6$ and $P(X = p) = 0.4$ Participants then try to recall the items, with $R = 0$ representing failure, and $R = 1$ representing success. Results show that $P(R = 1|X = w) = 0.4$, and $P(R = 1|X = p) = 0.7$. Suppose we know that the subject recalled a particular item. What is the probability that this item is a picture?
  - Preliminary note on notation: the notation we have been using is cumbersome. If we stick to the convention of using capital letters for random variables and lowercase letters for values, then we can just use $P(w)$ to mean $P(X = w)$. So let's do that, and also use r to represent $R = 1$, and $\neg r$ to represent $R = 0$. This is often done in the literature, but if you get confused, it may help to reintroduce the more explicit notation.

– Solution. We want to know $P(p|r)$, but only have conditional probabilities in the other direction. Use Bayes' rule! (And also the rule of total probability).

$$\begin{aligned} P(p|r) &= \frac{P(r|p)P(p)}{P(r)} && (9)\\ &= \frac{P(r|p)P(p)}{\sum_X P(r|X)p(X)} && (10)\\ &= \frac{P(r|p)P(p)}{P(r|p)P(p)+P(r|w)P(w)} && (11)\\ &= \frac{(0.7)(0.4)}{(0.7)(0.4)+(0.4)(0.6)} && (12)\\ &= \frac{0.7}{1.3} && (13) \end{aligned}$$

or just over one half.

* $P(p)$ is known as the **prior** probability of a picture occurring, i.e. the probability before we observe anything. $P(p|r)$ is the **posterior** probability, the probability we compute after having observed some data (here, the fact that the subject recalled the item in question).
* Note that although the prior probability of a picture is less than one half, once we know that the subject recalled this item, we are more likely to think that it's a picture than a word, because the subject is much more likely to recall pictures than words.

5 Multiple variables

- Anderson's model of memory that we saw last class (and many other models) invoke the notion of *conditional independence*. Variables A and B are said to be **conditionally independent** given a third variable C iff

$$P(A,B|C) = P(A|C)P(B|C) \quad (14)$$

  – Intuitively: A and B are conditionally independent given C iff once we know the value of C, A and B are independent.
  – Ex. Whether I stay up late and whether I show up on time for my 9 a.m. class are not independent. But, given that I left the house on time, they are independent. That is, they are conditionally independent given the knowledge of whether I left the house on time.

- In general, when multiple conditioning variables are present, all of the above rules can be applied to the two variables immediately adjacent to the '|' sign, with the additional conditioning variables just hanging around on the end.

  – Ex. Bayes' rule with an extra variable:

$$P(X|Y,Z) = \frac{P(Y|X,Z)P(X|Z)}{P(Y|Z)} \quad (15)$$

  – Ex. Rule of total probability with an extra variable:

$$P(X|Z) = \sum_Y P(X|Y,Z)P(Y|Z) \quad (16)$$

- Alternatively, you can treat some or all of the variables on one side of the '|' as a single "compound" variable. For instance, the value of the pair (X,Y) can be thought of as a single random variable A whose outcomes are all of the outcomes in the cross product of X and Y.

– Ex. Another way to use Bayes' rule with an extra variable:

$$P(X|Y,Z) = \frac{P(Y,Z|X)P(Y,Z)}{P(Y,Z)} \tag{17}$$

– Ex. Bayes' rule, treating B,C as a single variable but leaving D as a conditioning variable:

$$P(A|B,C,D) = \frac{P(B,C|A,D)P(A|D)}{P(B,C|D)} \tag{18}$$

- Finally, note that the ordering of variables is irrelevant as long as they stay on the same side of the '|' sign:

$$P(X,Y|Z) = P(Y,X|Z) \tag{19}$$
$$P(X|Y,Z) = P(X|Z,Y) \tag{20}$$
$$P(X|Y,Z) \neq P(X,Y|Z)!! \tag{21}$$

6 Expected values

- Sometimes it's useful to have an idea of the *average* value of a random variable.

  – Ex. I offer you a bet. You roll two dice, and I'll pay you £4 if you get a 6 or 11, but you pay me £1 otherwise. Should you take the bet?

  – Solution: your chance of getting £4 is P(6) + P(11) = 5/36 + 2/36 = 7/36, while your chance of losing £1 is 1 - 7/36 = 29/36. So on average you will make $(4)(7/36) - (1)(29/36) = -1/36$ pounds on this bet, i.e. you'll lose about 3 pence. Not a good bet!

  The average of a random variable is known as its **expected value** (or **expectation**) and is formally defined as

$$E[X] = \sum_x xP(X=x) \tag{22}$$

  That is, the sum of the probability of each value times the value itself. If we ran an experiment n times, we would *expect* the sum of the X values to be about $nE[X]$. (Can you see why? Consider the betting example above.)

- We can generalize to taking the expectation of a function of X, $f(X)$:

$$E[f(X)] = \sum_x f(x)P(X=x) \tag{23}$$

  This gives us the average value of the function.

  – Ex. I offer to let you roll a single die, and will give you a number of pounds equal to the square of the number that comes up. How much would you be willing to pay to play this game?

  – Solution: If X is the number that comes up, the expected value of your winnings is

$$E[X^2] = (1/6)(1^2) + (1/6)(2^2) + (1/6)(3^2) \tag{24}$$
$$+(1/6)(4^2) + (1/6)(5^2) + (1/6)(6^2)$$
$$= 91/6 \tag{25}$$

  So you should be willing to pay as much as £15.16 to play this game (assuming the die is fair!).