

DATA605

Week7 and Week8 Lecture

Review Interactive Website

Review the idea of likelihood:

<http://students.brown.edu/seeing-theory/basic-probability/index.html#first>

I'll point out in the lecture notes where this website will be useful

Probability Mass Functions

Remember that random variables can come in two flavors, continuous and discrete. Lets handle the discrete case first. A random variable that is discrete is associated with a “Probability Mass Function” which describes the various probabilities for values. Domain = values that random variable can take;
Co-domain = probability / likelihood of that value

Example:

Probability Density Functions

A probability density function is the same idea as a PMF, just for the continuous case. However, we have an issue here to fix: for a discrete set of values, it makes sense to talk about the likelihood of the random variable returning / being one particular value. This doesn't work for continuous values since the set of real numbers is infinite (i.e. is uncountably infinite, where a discrete set is typically finite).

We either:

1. Talk about the probability of a range of values, or
2. Understand that the PDF of a single value means the probability of being within an infinitesimal distance to the value

Rule: Summation of PMF / PDF = 1

By the rules of probability, when you sum probabilities across every element in the sample space, you should get 1. This means that for a PMF:

$$\sum_i P(Y == i) = 1$$

For a PDF, we need to either integrate over the whole sample space, assuming that the sample space is [a,b]:

$$\int_a^b p(x) dx = 1$$

Expectation of a Random Variable

In future courses, you will learn more about ‘summary statistics’. These are ways of summarizing a distribution of a random variable, and one of them is the notion of an ‘expected value’.

The expected value, or mean, of a random variable is a summary statistic that characterizes what ‘average’ value you expect from the variable after many trials / observations of the random variable.

Sidenote: think about Week7’s part 2 question about how to maintain the mean and variance of a stream of values where you cannot store all the values in memory.

Computing Expectation Values

The expected value of a random variable X can be computed, depending on whether or not the variable X is discrete or continuous:

$$E(x) = \sum_i P(x == i) \times i$$

$$E(x) = \int_a^b xp(x)dx$$

Expectation Example

<http://students.brown.edu/seeing-theory/basic-probability/index.html#first>

In the link above, you can simulate trials from a random variable representing the value of a die when being rolled. Play with the values at the bottom by dragging the bars, or click the buttons to run more trials

Non-Analytical PMF / PDF

What if we are dealing with a random variable that we don't really know the analytical form of? Meaning, what if we really don't know how to write the PDF / PMF mathematically? Sampling to the rescue!

You can sample from the distribution (however that is done in the example) and estimate the expected value. The more you sample, the better the estimate.

Check the Estimation tab on this site

<http://students.brown.edu/seeing-theory/basic-probability/index.html#first>

Expectation of a Function

To tidy up expectation values, do note that we can find the expectation value of a random variable as well as a function of a random variable. Its written the same way by instead of X you have $f(X)$

Things to remember for the future: the expected value of the random variable raised to n th power, $E(x^n)$ is termed the n^{th} moment of x .

Variance and Standard Deviation

To capture a measure of how "spread" is a random variable around its expected value, we can use variance or standard deviation. This is another summary statistic in that it tries to summarize a property of a distribution of a random variable into a single number.

$$Var(x) = E((x - E(x))^2)$$

$$Var(x) = E(x^2) - E(x)^2$$

The second equation is useful for part 2 of Week7's HW, since it allows us to calculate the variance without knowing values of X , as all we need is expected values of X and X^2 .

Standard Deviation

One issue with variance is that the units associated with it are whatever unit is associated with x , but squared. When we want to know the spread, it's useful to have it in units of x .

Therefore, we define the standard deviation of a random variable x to simply be the square root of $\text{Var}(x)$. It measures the scale of x and is expressed in the same units as x .

Conditional Probability

Now that we understand simple events, to be more useful we need to think about how events relate and happen in relation to one another. We already talked a little about this, in that independent events are events that basically have no causal link or otherwise don't affect one another.

The next step is to think about Conditional probability. This is defined as the probability of an event occurring **given** the knowledge of some related event occurring.

Conditional Probability Defined

$P(A | B)$ is read as the probability of event A occurring **given** the knowledge that an event B has already occurred.

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

When the two events are independent of each other, $P(A | B)$ is simply $P(A)$. If we know there is no connection between the two, then knowing that B occurred does not give us any more information about the situation.

Conditional probability can be thought of as ‘restricting the sample space’ to only B.

Inference

From wikipedia:

- Let A the event of interest be in the [sample space](#), say (X, P) .
- The occurrence of the event A knowing that event B has or will have occurred, means the occurrence of A as it is restricted to B ,
- Without the knowledge of the occurrence of B , the information about the occurrence of A would simply be $P(A)$
- The probability of A knowing that event B has or will have occurred, will be the probability of $P(A | B)$ compared with $P(B)$, the probability B has occurred.
- This results in $P(A|B) = P(A \cap B) / P(B)$

Notice there is a causal element in that we know A happened, now we want to know the probability of B for the future where A has occurred.

Trees

Example from <https://www.youtube.com/watch?v=WmcoWd8Uv-0> will be done interactively in the office hours.

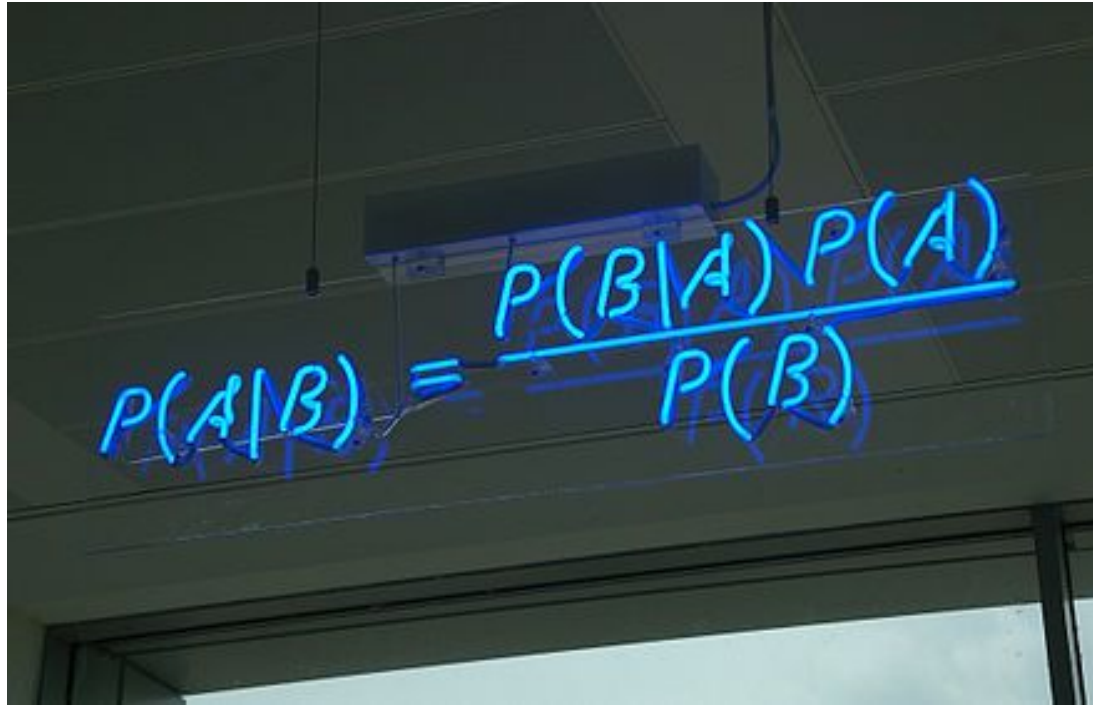
Also note there is an example in Wikipedia which is a simple table way of looking at this: https://en.wikipedia.org/wiki/Conditional_probability#Example

Visualization

The conditional probability tab shown here I think is an interesting way to view this:

<http://students.brown.edu/seeing-theory/compound-probability/index.html#third>

Bayes Rule



A photograph of a blue neon sign mounted on a dark ceiling. The sign displays the Bayes Rule formula in a handwritten style. The formula is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The sign is illuminated, and the background is dark.

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

Bayes Rule Cont'd

Bayes rule is a re-formulation of the conditional probability formula

Bayes' theorem can be derived from the multiplication law

$$P(X \cap Y) = P(Y)P(X|Y)$$

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

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

Bayes' Theorem can also be written in different forms

$$\begin{aligned} P(X|Y) &= \frac{P(X \cap Y)}{P(Y)} \\ &= \frac{P(X)P(Y|X)}{P(Y)} \\ &= \frac{P(X)P(Y|X)}{P(Y|X) + P(Y|X')} \end{aligned}$$

from : <http://www.onlinemathlearning.com/bayes-theorem.html>

Terminology

Likelihood

How probable is the evidence
given that our hypothesis is true?

Prior

How probable was our hypothesis
before observing the evidence?

$$P(H | e) = \frac{P(e | H) P(H)}{P(e)}$$

Posterior

How probable is our hypothesis
given the observed evidence?
(Not directly computable)

Marginal

How probable is the new evidence
under all possible hypotheses?
 $P(e) = \sum P(e | H_i) P(H_i)$

You'll likely see $P(e)$ broken down into $P(e|H) + P(e|\sim H)$, since either H did or did not occur (see formula in red section)

Terminology pt2

Given the 'prior' distribution and a new piece of evidence e , we are taking the prior $P(H)$ and multiplying it by something to get a 'posterior' distribution $P(H | e)$, which is defined as 'what is the new probability of H given that e has now just occurred'.

The something we multiply by is the ratio of the 'likelihood' of the evidence given H and the likelihood of the evidence overall ('marginal')

The posterior is now our updated probabilities (i.e. beliefs) about what events can occur in the future now that we know e happened

Bayesian Links

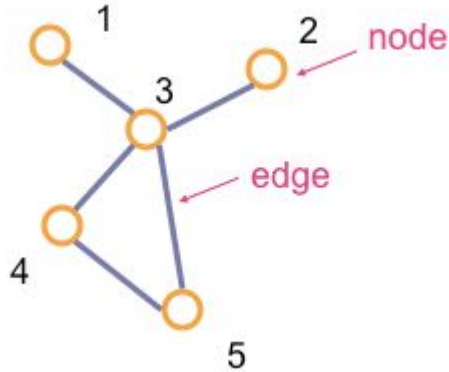
1. <https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/>
2. <https://www.khanacademy.org/math/ap-statistics/probability-ap/stats-conditional-probability/v/bayes-theorem-visualized>

When to use Bayes?

My rule of thumb is that if you can think of the situation as figuring out how you should update your beliefs about a situation, then Bayes fits well. Another way of saying it is that if you can formulate the question in terms of seeing evidence about an event, and that evidence changes your belief about the state of the world, then bayes makes sense.

Bayesian Networks

Probabilistic networks / graphs are a big part of ML research. There is a new class in them given by CUNY, IIRC. A network or graph is a mathematical structure made up of nodes that connect to one another via edges. This is useful for social media analysis, where each node is a person and edges represent friendship / following relationships

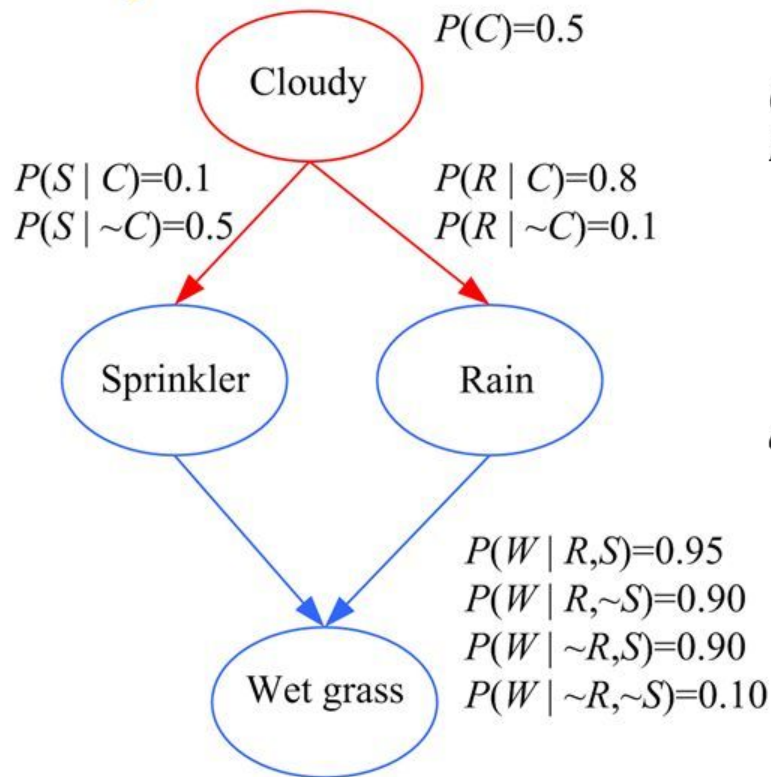


Bayesian Networks

In Bayesian networks, we use graph theory and incorporate probability theory to create a directed acyclic graph (DAG), where nodes represent variables and edges represent some kind of relationship between the variables indicated by a conditional probability table.

In some sense, it is an extension of the trees we have seen, since a tree is a kind of DAG.

Bayesian Networks: Causes



Causal inference:

$$P(W|C) =$$

$$P(W|R,S) P(R,S|\text{C}) + \\ P(W|\sim R,S) P(\sim R,S|\text{C}) + \\ P(W|R,\sim S) P(R,\sim S|\text{C}) + \\ P(W|\sim R,\sim S) P(\sim R,\sim S|\text{C})$$

and use the fact that

$$P(R,S|C) = P(R|C) P(S|C)$$

Diagnostic: $P(C|W) = ?$

Bayesian vs Frequentist

There is an unsolved issue in the philosophy of mathematics world, and it revolves, shockingly, around the exact definition of probability.

From the article: “A lot of the choice between frequentist and Bayesian statistics comes down to whether you think science should comprise statements about the world, or statements about our beliefs”

More resources:

- <http://stats.stackexchange.com/questions/22/bayesian-and-frequentist-reasoning-in-plain-english#56>
- <https://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-statisticians>