Admin

Lab 4 is due tonight.

Lab 3 grades and feedback will be back by Wednesday.

Midterm 1 will also be handed out during the time, and we have three hours to complete it.

Tues + Weds will be review sessions, as opposed to normal classes. We will work on the midterm during that time instead of another lab/

About the Midterm

Handed out this Wednesday and due next Wednesday. Timed 3hrs, but we have one double sided review/cheat sheet to look from. Outside of the study sheet, no phones, no calculators, or notes.

All work on the exam must be our own.

Going over Lab 2

Looking at the residuals, we see more distribution for the Degree 2 residuals. Thus, we are seeing the model is fitting the data better.

We can also see a *slight* pattern in the degree 1 residuals, which tells us that the model is not doing the best.

On the regression dataset, we can see that the model gets 'weird' as we increase the number of degrees. We can also see in the elbow plot that the best fitting models are 9/10, but they are probably overfit. If we try to train this on new test data, it will probably do a bad job.

In summary, these higher polynomial models are bad for two reasons

1. they are expensive
2. They overfit
   We go up up to degree 3, which is the last 'big drop', but at 4 we see no change, just more resources and overfitting. As a result, we want to big the degree 3 polynomial.

Intro to Probability

We live in a world of uncertainty, and things have a chance of happening.

We can consider a *frequentist* approach, where we look at the possible of something happening, or a *futurist* perspective, trying to predict the future based on the past.

Die toss example

Consider tossing a six sided due:

What is the probability of getting a one? $1/6$

Other die toss example

Now let's say we have a die with 2 1s, 2 2s, and 2 3s, what is the probability of getting a one?

$$P(e=1) = \frac{count(1)}{count(1) + count(2) + count(2)} = 1/3$$

More about outcomes (coin toss)

The set of all probabilities is called the probability distribution. I.e, each coin toss is an independent event, (bernouli trial). We can visualize this as a bar chart with equal probabilities for $[H, T]$

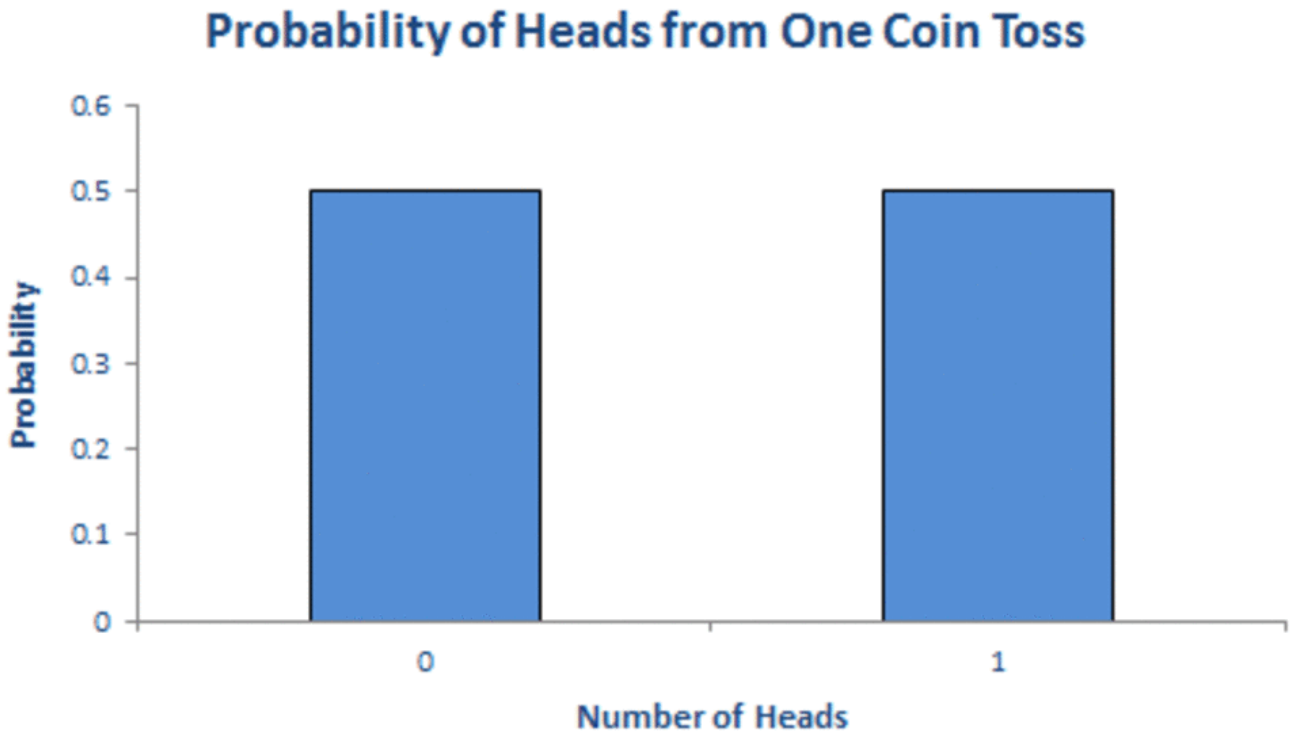


image from google search
Fair coin tosses are **independent**, meaning that the toss of one coin has no impact on the other

Another coin example

Which is greater?
5 heads

$$P(HHHHH)$$

Or 2 H one T and two H

$$P(HHTHH)$$

Since these events are independent, these two have the same probability, and they are just arbitrary values of 5 flips in a row.

Probability axioms

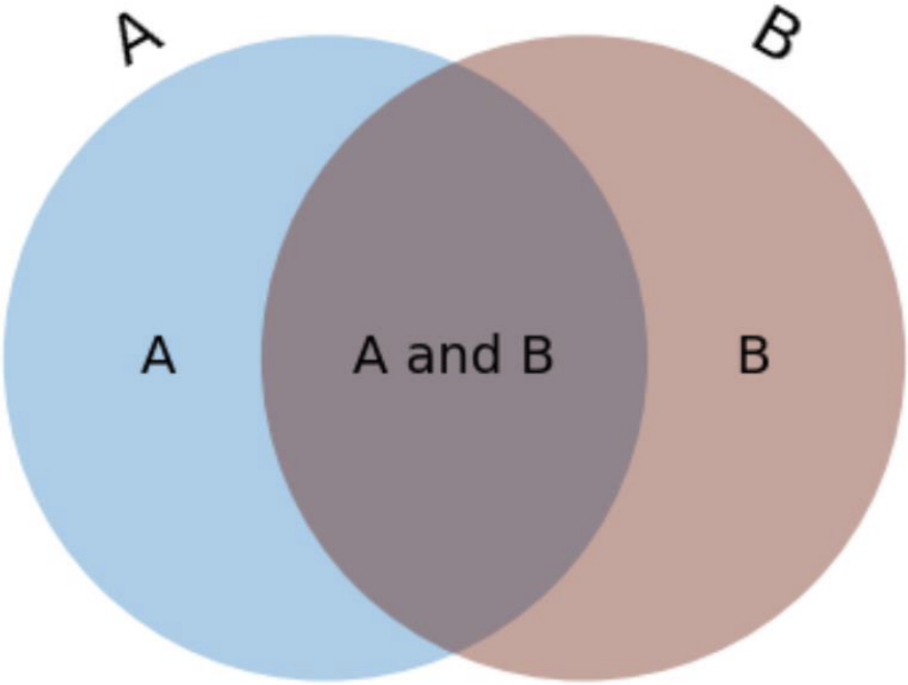
All probabilities much be larger than 0, or

$$P(e) \geq 0$$

Sum of all probabilities is one, or

$$\sum_{e \in E} P(e) = 1$$

Joint Probability



The probability of two events happening together, or

$$P(e_1 \wedge e_2)$$

If these events are independent, then this is this product, or

$$P(e_1 \wedge e_2) = P(e_1)P(e_2)$$

If the intersection of these two events is empty, or

$$P(e_1 \wedge e_2) = \emptyset \rightarrow P(e_1 \wedge e_2) = 0$$

You can think of each probability here as a scaling factor, if event 2 has some probability e_2 and we want to find the probability $P(e_1 \wedge e_2)$ we are basically scaling e_2 with respect to e_1.

If $P(e_1) = 1/2, P(e_2) = 1/3$

$$P(e_1 \wedge e_2) = 1/2 * 1/3 = \frac{1}{6}$$

Looking at the graph, we can see the probability spaces corresponding to the areas in the graph.

$$P(A \wedge B)$$

can be seen as the area occupied by both probability spaces.

Conditional Probability

The probability of one event given another

$$P(e_2|e_1)$$

means P_2 *given* P_1, or P_2 assuming P_1 happened already.
The difference between conditional and joint probability is that both look at the intersection of two things happening, but just with respect to A and B, not the entire universe, like in Joint Probability. This will be explained better in the section on *Bayes Rule.*

For instance, if we are calculating the probability of positive bell shaped mushrooms, we are only considering examples with bell shaped, and not all mushrooms. We don't consider those other examples.

Example

Lets say we have two events R for rain, U for umbrella
The probability of rain is 20% or

$$P(R) = .2$$

The probability of rain and us brining and umbrella is

$$P(R \wedge U) = .15$$

These events are clearly not indepedent, since we are much more likely to bring an umbrella if it is raining. What is $P(U|R)$?
We can find it with this formula

$$P(U|R) = \frac{P(R|U)}{P(R)}$$

$$= \frac{.15}{.2}$$

$$= \frac{3}{4}$$

This formula

$$P(U|R) = \frac{P(R,U)}{P(R)}$$

is known as Bayes Rule.

Bayes Rule

Starting with

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

Consider the first one: We can divide both sides by the $P(B)$ to find

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

or we can start with $P(A, B) = P(B|A)P(A)$ and divide by $P(A)$ to get the other way around.
This is also only true when the events are independent, which is defined when

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

for all outcomes.
This then imples

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

since there is no impact of B on A.

Conditional Independce

There is a much more useful form of independence
Two events are independent give another event happening

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

For A is thunder, B is rain and C is lighting,

Thunder is independent of rain given lighting.

Marginal Probability

If we are given $P(X, Y)$ and we want to find $P(X)$ what can we do?
We can do a few things, we can 'marginalize out' the y by summing over all $y \in Y$, or all the possible values that y can have. In the discrete case we do a simple sum, and in the continous an integral.
Discrete

$$p(x) = \sum_{y \in Y} P(x, y)$$

Continous

$$p(x) = \int p(x, y) dy$$

Ex.

We can have some event rain, $P(R)$ or no rain, $P(\bar{R})$
We can say that the sum of all possible events should be one, or

$$P(R) + P(\bar{R}) = 1$$

Lets say we then have the probability of tennis given weather, or

$$P(T, weather) = \alpha$$

We can find $P(T)$ by summing across all weather scenarios, or

$$P(T) = P(T, R) + P(T, \bar{R})$$

Ex 2

Lets say we got an email, and it is represented by the words in the email and we are trying to find

$$P(spam|words)$$

Using bayes rule we can write as

$$P(spam|words) = \frac{P(spam, words)}{P(words)}$$

This might be very difficult to figure out, as there are a lot of words we could be checking under, so let's rewrite as

$$P(spam|words) = \frac{P(spam, words)}{P(words, spam) + P(words, sp\bar{a}m)}$$

Or further

$$P(A) = \sum_{b \in vals(B)} (A, B = b)$$

Let's apply bayes rule in the numerator and then denominator

$$P(spam|words) = \frac{p(spam)p(words|spam)}{p(spam)p(words|spam) + p(sp\bar{a}m)p(words|sp\bar{a}m)}$$

This let's us reduce the number of things we need down to much more measurable things, as we can find $p(spam)$ and $p(words|spam)$ easier than $p(spam, words)$

There are cool names for all this as well

$$P(spam|words)$$

is called the posterior

The denominator:

$$p(spam)p(words|spam) + p(sp\bar{a}m)p(words|sp\bar{a}m)$$

is the evidence and
The numerator

$$p(words|spam)$$

is the likelihood and

$$p(spam)$$

is the prior.