Probability Tips and Tricks

A quick and dirty intro

Why do we need probability? (Obligatory motivation slide)

- The world is full of random shit
- Our job as professional nerds is to quantify that
- Also, you’re sitting some kind of exam and want to pass, that’s probably enough motivation

Random variables

- Some things occur randomly: these have probability distributions
- A probability distribution describes how these random variables behave
- In practice does this actually work?

Properties of probability distributions

- The probability of anything happening has to lie between 0 and 1.
- Probability =0 <- will not happen, no chance
- Probability =1 <- certainty, no chance
- Anything in between? <- randomness

Types of Probability Distributions

- There are lots of types:
  - Probability Density Function
  - Cumulative Density Function

  - Marginal Probability
  - Joint Probability
  - Conditional Probability

Random Variables come in two kids

- Discrete/categorical
  - Typical example: gender <- flawed example
  - Typical example: colour of your car <- boring example
  - Will my kids listen to me and turn off youtube at the end of the next minecraft video? Yes/no
- Continuous
  - Typical example: height/weight/income
  - Favourite example: how many Cadbury Crème eggs can I eat? <- continuous because I’m going to count them in fractional amounts, e.g. 2.5 eggs. It would be categorical/discrete if I only counted them in whole eggs

PDFs work differently for the two kinds of random variables

- Fundamentally: it's all about the area under the curve
  - But for discrete random variables, the curve is not smooth and cannot be integrated, so we add instead
- Continuous random variables <- integrals
  - The CDF is the integral of the probability density function up to the break point
  - The probability of a continuous random variable lying between two bounds is the integral between them
- Discrete random variables <- summations
  - The CDF is the summation up to the break point
  - The probability of a discrete RV lying between two bounds is the summation between: beware start and end points!

Python walk through with Allen Downey

- http://allendowney.blogspot.com.au/2016/06/what-is-distribution.html

- Allen is a great data scientist and does a really good notebook walk through.

Conditional Probability Visual

- http://setosa.io/conditional/?utm_content=buffereb70e&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer
- By Victor Powell- so what is a conditional probability?

Bayes' Rule: The Linkage Between these Probabilities (and the closest thing stats nerds have to a turf war)

$$P\left(F \middle| D\right) = \frac{P(F \text{ and } D)}{P(D)}$$

- Conditional = Joint/Marginal – we can manipulate this
- Conditional * Marginal = Joint
- Marginal = Joint/Conditional
- The trick with Bayes' Rule is manipulating it to give you what you need.

Question 4.45 from Groebner, Shannon and Fry: *Business Statistics,* 9th Edition, Pearson 2014.

4-45. Parts and materials for the skis made by the Downhill Adventures Company are supplied by two suppliers. Supplier A's materials make up 30% of what is used, with Supplier B providing the rest. Past records indicate that 15% of Supplier A's materials are defective and 10% of B's are defective. Since it is impossible to tell which supplier the materials came from once they are in inventory, the manager wants to know which supplier more likely supplied the defectiv materials the foreman has brought to his attention. Provide the manager this information.

Joint, Marginal and Conditional Probabilities

- P(Supplier A) = 0.3 – marginal probability
- P(Supplier B) = 0.7 – marginal probability
  - P(<insert something all by itself>) is usually the marginal probability
- P(Defective|Supplier A) = 0.15 – conditional probability of defective **given** supplier A
- P(Defective|Supplier B) = 0.10 – conditional probability of defective **given** supplier B
  - P(<something>|<something else>) is the conditional probability
- P(Defective and Supplier A) is the joint probability- the probability of both things happening at the same time

Step (1) What do we want?

- P(Supplier A|Defective)
- P(Supplier B|Defective)
- And to find out which is higher

Step (2) What do we have?

- P(Supplier A) = 0.3
- P(Supplier B) = 0.7
- P(Defective|Supplier A) = 0.15
- P(Defective|Supplier B) = 0.10

We want to reverse these!

Step (3): Panic

- OK don't really: remember Bayes' Rule
- P(Supplier A|Defective)= P(Supplier A and Defective)/P(Defective)
- But we don't have either P(Supplier A and Defective) OR P(Defective)!!

Step (4): We just nailed Bayes' Rule we can do this!

- P(Defective|Supplier A)=P(Defective and Supplier A)/P(Supplier A)

We have these….!!!!

- So…
- P(Defective and Supplier A)=P(Defective|Supplier A)*P(Supplier A)=0.15*0.3

Now we just need P(Defective)

P(Defective) = P(Defective and Supplier A)+ P(Defective And Supplier B)

- We are summing up both options

P(Defective) =P(Defective|Supplier A)P(Supplier A) + P(Defective|Supplier B)P(Supplier B)

(Bayes' Rule again!)

= (0.15)(0.3) + (0.10)(0.7) = 0.115

Step (5): Plug these results into Step 3

- P(Defective and Supplier A)=0.45
- P(Defective)=0.115
- P(Supplier A|Defective)= P(Supplier A and Defective)/P(Defective)

  =0.45/0.115

  =0.3913

Same for Supplier B

- P(Supplier B|Defective) = (0.10)(0.7)/0.115 = 0.6087
- Supplier B is the most likely to have supplied the defective parts.