Introduction to Data Science

Probability and Bayes Rule

Gordon Anderson

Probability Theory

- What’s the probability that something will occur?
- Have to define a domain or “universe” that the something is part of.
- For example, let’s say the domain is all values on two 6-sided dice.
- What is the probability that 7 occurs when the dice are rolled?
- Let X be a random variable that represents values that occur on 2 dice.
- $P(X = 7) = \frac{ways\ that\ it\ can\ occur}{all\ possible\ outcomes} = \frac{(1,6)+(6,1)+(3,4)+(4,3)+(2,5)+(5,2)}{6x6=36} = \frac{6}{36}$

- The denominator implies that we consider probability in terms of a defined domain, or “sample space”.

Probability Standard Notation

- A random variable is a variable that is assigned the result of sampling from a random process, or “experiment”.
- Examples of “standard” probability notation:
- $P(X = 7)$ the probability that the random variable “X” equals 7.
- $P(color = green)$ the probability that the random variable “color” equals “green”.
- $P(spam)$ the probability that “spam” occurs.
- $P(grade \geq 50)$ the probability that the random variable “grade” is greater than or equal to 50. This is called a “cumulative probability”.

Probability Theory

- Example 1: a coin with heads and tails.
- Define the sample space of all possible outcomes:

S =

Head	Tail

(the "S" stands for Sample Space)

- $P(Head) = \frac{ways\ that\ it\ can\ occur\text{:}only\ 1\ way}{all\ possible\ outcomes\text{:}there\ are\ 2\text{−}Head\text{,}Tail} = \frac{1}{2}$

Probability Theory

$$P(Head) = \frac{ways\ that\ it\ can\ occur{:}only\ 1\ way}{all\ possible\ outcomes{:}there\ are\ 2{-}Head{,}Tail} = \frac{1}{2}$$

$$P(Tail) = \frac{ways\ that\ it\ can\ occur{:}only\ 1\ way}{all\ possible\ outcomes{:}there\ are\ 2{-}Head{,}Tail} = \frac{1}{2}$$

$$P(Head) + P(Tail) = \frac{1}{2} + \frac{1}{2} = 1$$

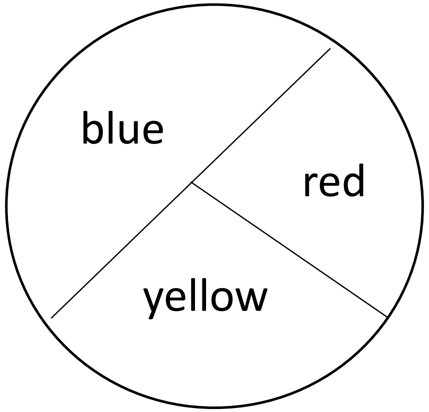
(the sum of the probabilities of all events in a sample space must equal 1)

- Note that: $P(Head) = 1 - P(Tail)$

Probability Theory

- Example 2: 3 red, 5 blue, 6 yellow marbles.
- Define the probability space of all possible outcomes:

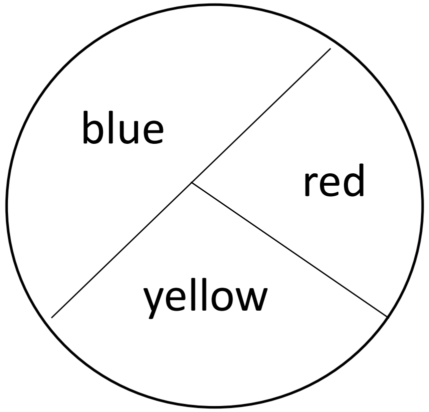
S =



- $P(red) = \frac{ways\ that\ it\ can\ occur}{all\ possible\ outcomes} = \frac{3}{14}$

Probability Theory

- Example: 3 red, 5 blue, 6 yellow marbles.
- $P(red) = \frac{3}{14}$
- $P(blue) = \frac{5}{14}$
- $P(yellow) = \frac{6}{14}$



All must sum to 1:

$$\frac{3}{14} + \frac{5}{14} + \frac{6}{14} = 1$$

Probability Theory

- Example 3: a text document containing words.
- What is the probability of a specific word occurring in the doc?

S = $\boxed{\text{All words in doc}}$

- $P(word = "data") = \dfrac{number\ of\ occurences\ of\ "data"\ in\ doc}{ocurences\ of\ all\ words\ in\ doc}$

Independent/Dependent Events

- Independent events: the occurrence of event A has no bearing on the occurrence of event B.
  - Example: The outcome of a dice roll does not influence the outcome of the next roll.
- Dependent event: If the occurrence of event A affects the probability of the occurrence of event B.
  - Example: Owning a sports car increases the likelihood of getting a speeding ticket.
- Question: is the probability of a word occurring in a document independent of the probability of a different word occurring?

Independent/Dependent Events

- Probability of two independent events happening at the same time:
- $P(A\ and\ B) = P(A) \times P(B)$

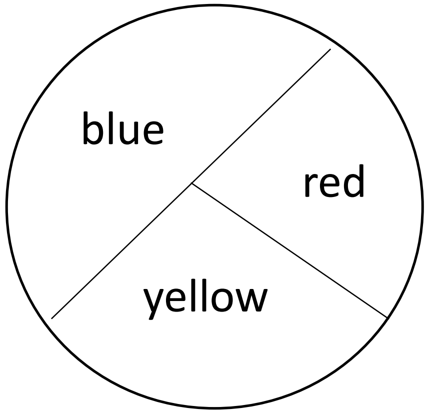
- Probability of two dependent events happening: Well, they can’t happen at the same time- so, say A happens first:
- $P(A\ and\ B) = P(A) \times P(B\ after\ A\ happened)$

Sampling and Dependence

- In probability and statistics we are working with observations obtained (sampled) from the world (sample vs population).
- There are two kinds of sampling: with replacement and without replacement.
- Sample with replacement: the sample observed is returned to the sample space to be (possibly) sampled again.
- Sample without replacement: the sample observed is not returned to the sample space.

Sampling with replacement

- Example: 3 red, 5 blue, 6 yellow marbles.
- $Prob(red) = \frac{3}{14}$
- $Prob(blue) = \frac{5}{14}$
- $Prob(yellow) = \frac{6}{14}$



Probability of picking a red marble:

$$\frac{3}{14}$$

Probability of picking a blue marble replacing the first marble:

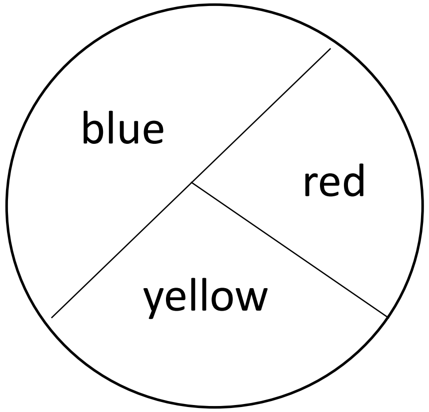
$$\frac{5}{14}$$

Probability of picking a red marble replacing the first and second marbles:

$$\frac{3}{14}$$

Sampling without replacement

- Example: 3 red, 5 blue, 6 yellow marbles.
- $Prob(red) = \frac{3}{14}$
- $Prob(blue) = \frac{5}{14}$
- $Prob(yellow) = \frac{6}{14}$



Probability of picking a red marble:

$$\frac{3}{14}$$

Probability of picking a blue marble without replacing the first marble:

$$\frac{5}{13}$$

Probability of picking a red marble without replacing the first and second marbles:

$$\frac{2}{12}$$

Joint and Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

Table of frequencies (occurrences) of majors vs student home demographics.

We’ll look at using it to calculate joint and conditional probabilities.

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

Joint probability: the probability of two events happening together:

$$P(A\ and\ B) = P(AB) = P(A, B)$$

Probability a student is an informatics major and from a rural location:

$$P(Rural, Informatics) = \frac{Rural\ and\ Informatics}{All\ students} =?$$

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

Joint probability: the probability of two events happening together:

$$P(A\ and\ B) = P(AB) = P(A, B)$$

Probability a student is an informatics major and from a rural location:

$$P(Rural, Informatics) = \frac{Rural\ and\ Informatics}{All\ students} = \frac{1}{75} = .013$$

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an informatics major and from an urban location?

$$P(Urban, Informatics) = ?$$

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an informatics major and from an urban location?

$$P(Urban, Informatics) = \frac{37}{75} = .49$$

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an English major and from an urban location?

$$P(Urban, English) = ?$$

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an English major and from an urban location?

$$P(Urban, English) = \frac{20}{75} = .27$$

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an Informatics major?

$$P(Informatics) =?$$

Joint Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an Informatics major?

$$P(Informatics) = \frac{38}{75} = .51$$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

Conditional probability: the probability of one event happening given a previous event occurred:

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

What is the probability a student is from an urban area given she is an informatics major?

$$P(Urban \mid Informatics) = \frac{Urban\ and\ Informatics}{All\ Informatics\ students} = ?$$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

Conditional probability: the probability of one event happening given a previous event occurred:

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

What is the probability a student is from an urban area given she is an informatics major?

$$P(Urban \mid Informatics) = \frac{Urban\ and\ Informatics}{All\ Informatics\ students} = \frac{37}{38} = .97$$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an informatics major given he is from an urban area?

$$P(Informatics|Urban) = \frac{Urban\ and\ Informatics}{All\ Urban\ students} =?$$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

What is the probability a student is an informatics major given he is from an urban area?

$$P(Informatics|Urban) = \frac{Urban\ and\ Informatics}{All\ Urban\ students} = \frac{37}{57} = .65$$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

Compare- they are not symmetric:

$$P(Informatics|Urban) = \frac{37}{38} = .97$$

$$P(Urban \mid Informatics) = \frac{37}{57} = .65$$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

Calculate:

$$P(Informatics|Rural) = ?$$
$$P(Rural \mid Informatics) = ?$$
$$P(English|Urban) = ?$$
$$P(Urban| \; English) = ?$$
$$P(English|Rural) = ?$$
$$P(Rural| \; English) = ?$$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

$P(Informatics|Rural) =?$

$P(Urban|\ English) =?$

$P(Rural\ |\ Informatics) =?$

$P(English|Rural) =?$

$P(English|Urban) =?$

$P(Rural|\ English) =?$

Conditional Probability

	Informatics	English	Total
Rural	1	17	18
Urban	37	20	57
Total	38	37	75

$$P(Informatics|Rural) = \frac{1}{18} = .06$$

$$P(Urban|\ English) = \frac{20}{37} = .54$$

$$P(Rural \mid Informatics) = \frac{1}{38} = .03$$

$$P(English|Rural) = \frac{17}{18} = .94$$

$$P(English|Urban) = \frac{20}{57} = .35$$

$$P(Rural|\ English) = \frac{17}{37} = .46$$

Conditional Probability

- General notation:

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

- Note that this is generally true: $P(A|B) \neq P(B|A)$
- Joint probabilities *are* symmetrical: $P(A\ and\ B) = P(B\ and\ A)$
- Also, $P(A\ and\ B)$ is also written as $P(A, B)$ or $P(AB)$

Bayes Rule (derivation)

1. Given this conditional probability: $P(A|B) = P(A,B)/P(B)$
2. Multiply by $P(B)$: $P(A|B)P(B) = P(A,B)$
3. Given this conditional probability: $P(B|A) = P(B,A)/P(A)$
4. Multiply by $P(A)$: $P(B|A)P(A) = P(B,A)$

Since the r.h.s. of 2 and 4 are equal:

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

Now divide by $P(B)$:

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

Bayes Rule

General notation:

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

Put this another way, suppose we have a theory T and evidence E:

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

Bayes Rule

Let's say our "theory" is that an email is spam, and the evidence is the occurrence of a keyword, such as "viagra".

$$P(spam|word = "viagra") = \frac{P(word = "viagra"|spam)P(spam)}{P(word = "viagra")}$$

The denominator is calculated (leaving out the "viagra" part for brevity):

$$P(word) = P(word|spam)P(spam) + P(word|\neg spam)P(\neg spam)$$

Bayes Rule

The denominator is not too bad given that we have one theory: an email is either spam or it is not spam, so:

$$P(spam) = 1 - P(\neg spam)$$

$$P(word) = P(word|spam)P(spam) + P(word|\neg spam)P(\neg spam)$$

But what if we had several theories?

Bayes Rule

Suppose we have *n* theories $\vec{T}$ and evidence E:

likelihood

prior probability

posterior probability

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

The above is a Bayesian *probability model*. If we add a selection process, to assign the theory with the highest posterior probability, $\hat{T}$, to the evidence, then we have a Bayesian *classifier*:

$$\hat{T} = ARGMAX_i = P(E|T_i)P(T_i)$$

Bayes Rule

In the Bayes model:

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

We can “disregard” the denominator in the classifier as it does not depend on the theory:

$$\hat{T} = ARGMAX_i = P(E|T_i)P(T_i)$$

Bayes Rule

One advantage of Bayesian analysis is that we can use a “prior” belief to boost the model, thus incorporating knowledge about the domain.

$$\hat{T} = ARGMAX_i = \underbrace{P(E|T_i)}_{\text{“Likelihood” Term}}\underbrace{P(T_i)}_{\text{“Prior” Term}}$$

More on this later on…