



CSC380: Principles of Data Science

Alon Efrat

Today:

- Numpy package
- Conditional probability
- Independence

Credit:

- Jason Pacheco,
- Kwang-Sung Jun,
- Chicheng Zhang
- Xincheng yu

1

1

Review

2

- What is probability?
- Axioms
- Event = set \Rightarrow use set theory!
- Set theory + axiom 3 is quite useful
- Draw diagrams
- Lots of jargons

- Make your own cheatsheet.

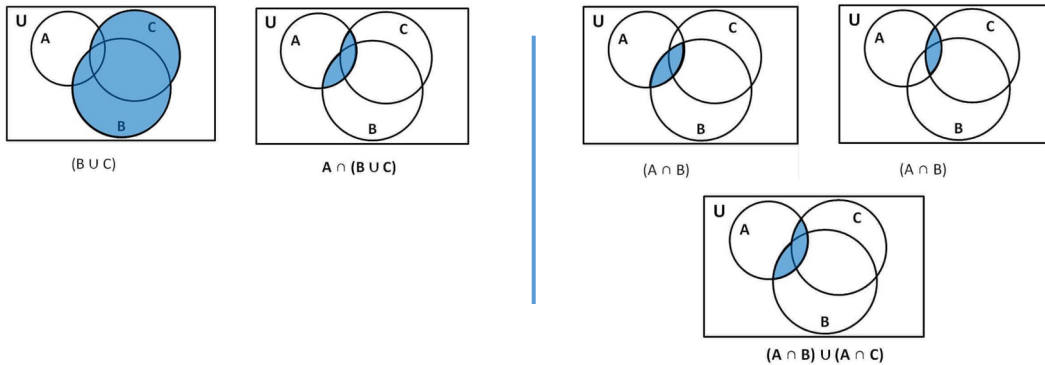
2

Review

3

$$\bullet A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

distributive law by Venn diagram

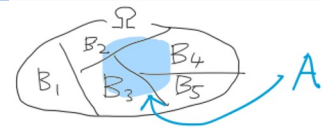


3

Review

4

$$\begin{aligned} \bullet A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ \bullet A &= A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i) \\ &= A \cap (B_1 \cup B_2 \cup B_3 \dots \cup B_n) \\ &= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \dots \cup (A \cap B_n) \end{aligned}$$



Law of total probability: Let A be an event. For any events B_1, B_2, \dots that partitions Ω , we have

$$P(A) = \sum_i P(A \cap B_i)$$

4

Review

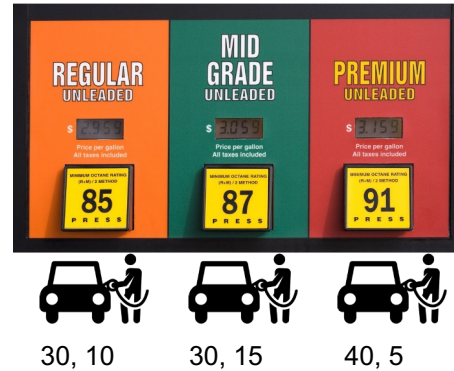
5

$$P(A) = \sum_i P(A \cap B_i)$$

A: the customer (100)

B: fill gas

- B_1 : unleaded (30)
- B_2 : mid grade (30)
- B_3 : premium (40)



$P(A = \text{student})$

$= P(A = \text{student}, B = B_1) + P(A = \text{student}, B = B_2) + P(A = \text{student}, B = B_3)$

$= P(A = \text{student} | B = B_1)P(B = B_1) + P(A = \text{student} | B = B_2)P(B = B_2) + P(A = \text{student} | B = B_3)P(B = B_3)$

5

Overview

6

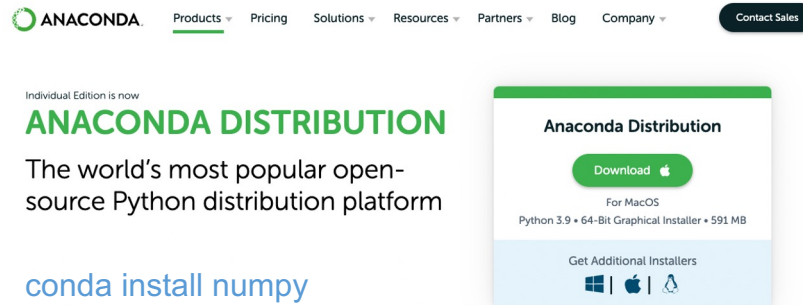
- Numpy package
- Conditional probability
- Independence

6

Numpy Library

7

Package containing many useful numerical functions...



`conda install numpy`

If you use pip:

`pip install numpy`

...we are interested in `numpy.random` at the moment

7

numpy.random

8

`numpy.random.randint`

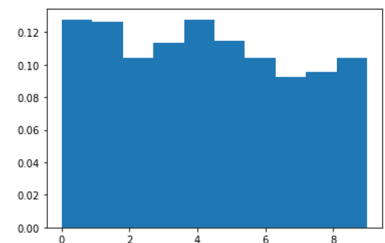
`numpy.random.randint(low, high=None, size=None, dtype='i')`

Return random integers from *low* (inclusive) to *high* (exclusive).

Return random integers from the "discrete uniform" distribution of the specified dtype in the "half-open" interval $[low, high)$. If *high* is None (the default), then results are from $[0, low)$.

Sample a discrete uniform random variable,

```
import matplotlib.pyplot as plt
X = np.random.randint(0,10,1000)
count, bins, ignored = plt.hist(X, 10, density=True)
plt.show()
```



- **Caution** Interval is $[low, high)$ and upper bound is **exclusive**
- **Size** argument accepts tuples for sampling ndarrays (multidimensional arrays)

8

numpy.random

9

Allows sampling from many common distributions

Set (global) random seed as,

```
import numpy as np
seed = 12345
np.random.seed(seed)
```

- 😊 easier to debug (otherwise, you may have 'stochastic' bug)
- 😞 can be risky

E.g., buy into the result based on a particular seed, publish a report.
... turns out, you get a widely different result if you use a different seed!

Recommendation: change the seed every now and then

9

Conditional Probability

10

10

Conditional Probability

11

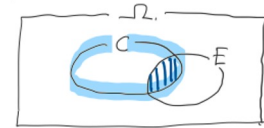
Two fair dice example:

- Suppose I roll two dice secretly and tell you that one of the dice is 2. C
- In this situation, find the probability of two dice summing to 6. E

```
import numpy as np
for n in [10,100,1000,10_000,100_000, 1_000_000]:
    res_dice1 = np.random.randint(6,size=n) + 1
    res_dice2 = np.random.randint(6,size=n) + 1
    res = [(res_dice1[i], res_dice2[i]) for i in range(len(res_dice1))]
```

```
conditioned = list(filter(lambda x: x[0] == 2 or x[1] == 2, res))
n_eff = len(conditioned)
```

```
cnt = len(list(filter(lambda x: x[0] + x[1] == 6, conditioned)))
print("n=%9d, n_eff=%9d, result: %.4f " % (n, n_eff, cnt/n_eff))
```



compare:
without conditioning,
it was 0.138.

```
n= 10, n_eff= 4, result: 0.0000
n= 100, n_eff= 32, result: 0.2500
n= 1000, n_eff= 300, result: 0.1733
n= 10000, n_eff= 3002, result: 0.1742
n= 100000, n_eff= 30590, result: 0.1823
n= 1000000, n_eff= 305616, result: 0.1818
```

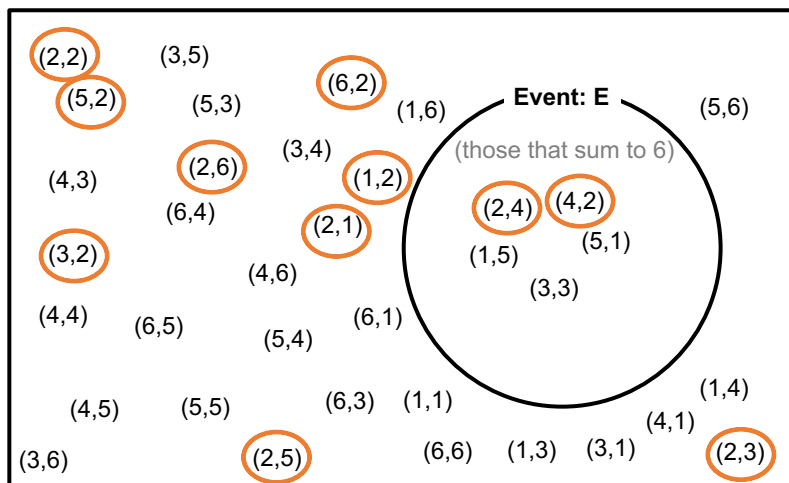
```
n= 10, n_eff= 3, result: 0.3333
n= 100, n_eff= 32, result: 0.0625
n= 1000, n_eff= 343, result: 0.2245
n= 10000, n_eff= 3062, result: 0.1897
n= 100000, n_eff= 30651, result: 0.1811
n= 1000000, n_eff= 305580, result: 0.1808
```

11

Random Events and Probability

12

What is the probability of having two numbers sum to 6 given one of dice is 2?



Each outcome is equally likely.
by the **independence**
(will learn this concept later)
 $\Rightarrow 1/36$

sum to 6:
 $\Rightarrow 5$

one of dice is 2:
 $\Rightarrow 11$

sum to 6 and one of dice
is 2:
 $\Rightarrow 2$

answer:
 $2/11 = 0.181818...$

12

Conditional Probability

13

Two fair dice example

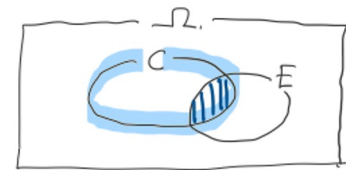


- Find the probability of **one of the dice is 2 (event C)** and **two dice summing to 6 (E)**

$$P(E \cap C)$$

- I secretly tell you **one of the dice is 2**, find the probability of **two dice summing to 6**.

$$\frac{P(E \cap C)}{P(C)}$$



13

Conditional Probability

14

- Two fair dice example:

- Suppose I roll two dice and secretly tell you that **one of the dice is 2**. C
- In this situation**, find the probability of **two dice summing to 6**. E

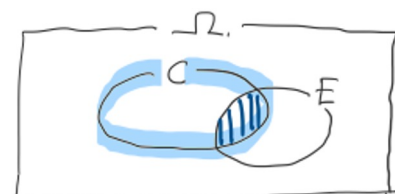
- Turns out, such a probability can be computed by $\frac{P(E \cap C)}{P(C)}$

- It's like "zooming in" to the condition.

- This happens a lot in practice, so let's give it a notation:

$$P(E|C) := \frac{P(E \cap C)}{P(C)}$$

Say: probability of " E given C ", " E conditioned on C "



"it's the ratio"

14

Conditional Probability

15

Q: Conditional probability $P(A|B)$ could be undefined. When?

- A: The denominator can be 0 already. In this case, numerator is also 0!

Note $P(A|B) \neq P(B|A)$ in general!

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

E.g., throw a fair die. $X :=$ outcome. $A = \{X=4\}$, $B = \{X \text{ is even}\}$

Question: $P(A|B) = P(B|A)$?

- $P(A) = 1/6$
- $P(B) = 1/2$
- $P(A \cap B) = 1/6$
- Therefore, $P(A|B) = 1/3$, $P(B|A) = 1$

15

Conditional Probability

16

Chain rule

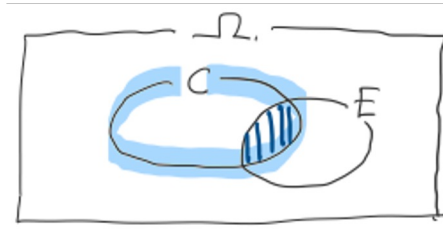
- $P(A \cap B) = P(A|B)P(B)$ ←just a rearrangement of definition: $P(A|B) := \frac{P(A \cap B)}{P(B)}$
- $P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C)$
- $P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1) \prod_{i=2}^n P(E_i | \cap_{j=1}^{i-1} E_j)$ valid for any ordering!

16

Conditional Probability

17

- $P(E \cap C) = P(E|C)P(C) = P(C|E)P(E)$



"it's the ratio"

17

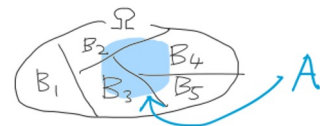
Conditional Probability

18

Recall: let A be an event. For events B_1, B_2, \dots that partitions Ω , we have

$$\begin{aligned} P(E \cap C) &= P(E|C)P(C) \\ &= P(C|E)P(E) \end{aligned}$$

$$P(A) = \sum_i P(A \cap B_i)$$



$$A = A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i)$$

Check axiom 3 & distributive law!

Law of total probability: If $A \in \mathcal{F}$ and $\{B_i \in \mathcal{F}\}_i$ partitions Ω , then

$$P(A) = \sum_i P(A, B_i) = \sum_i P(B_i)P(A|B_i)$$

Shortcut:
 $P(A, B) := P(A \cap B)$

$$= \sum_i P(A)P(B_i|A) \quad (\text{by definition})$$

18

Review

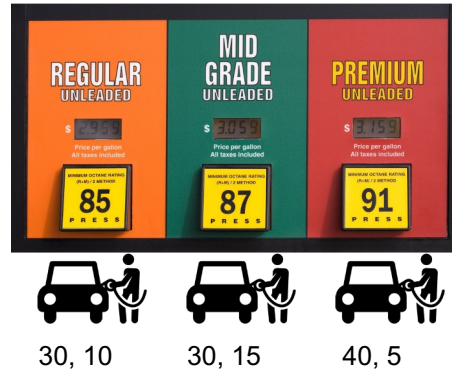
19

$$P(A) = \sum_i P(A \cap B_i)$$

A: the customer (100)

B: fill gas

- B_1 : unleaded (30)
- B_2 : mid grade (30)
- B_3 : premium (40)



$P(A = \text{student})$

$$= P(A = \text{student}, B = B_1) + P(A = \text{student}, B = B_2) + P(A = \text{student}, B = B_3)$$

$$= P(A = \text{student} | B = B_1)P(B = B_1) + P(A = \text{student} | B = B_2)P(B = B_2) + P(A = \text{student} | B = B_3)P(B = B_3)$$

19

A customer A picks type of gas.
What is the prob that A is a student?

Review

20

$$P(A) = \sum_i P(A, B_i) = \sum_i P(B_i)P(A|B_i)$$

$P(A = \text{student})$

$$= P(A = \text{student} | B = B_1)P(B = B_1) + P(A = \text{student} | B = B_2)P(B = B_2) + P(A = \text{student} | B = B_3)P(B = B_3)$$

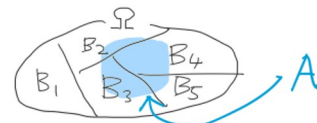
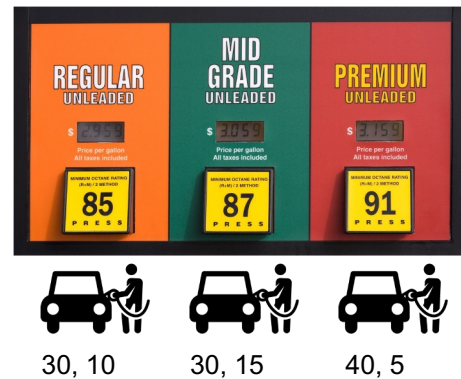
$P(A = \text{student})$

$$= 10/30 \times 30/100 + 15/30 \times 30/100 + 5/40 \times 40/100$$

$$\sum_i P(B_i | A) = 1$$

$$P(B_1 | A = \text{student}) + P(B_2 | A = \text{student}) + P(B_3 | A = \text{student})$$

$$= \frac{10}{10+15+5} + \frac{15}{10+15+5} + \frac{5}{10+15+5} = 1$$



20

Conditional Probability

21

The Public Health Department gives us the following information:

- A test for the disease yields a positive result (+) 90% of the time when the disease is present (Y) $P(+ | Y) = 0.9$
- A test for the disease yields a positive result 1% of the time when the disease is not present (N) $P(+ | N) = 0.01$
- One person in 1,000 has the disease. $P(Y) = 0.001$

Q: What is the probability that a person with positive test has the disease? $P(Y | +)$?

Pick a person **uniformly at random** from the population. Apply the test. When test=+, what is the probability of this person having the disease (Y) ?

21

Conditional Probability

22

What we know:

Positive result Patient positive

$$P(+ | Y) = 0.9$$

$$P(+ | N) = 0.01$$

$$P(Y) = 0.001$$

 \Rightarrow

$$P(- | Y) = 0.1$$

$$P(- | N) = 0.99$$

$$P(N) = 0.999$$

Question: $P(Y | +)$

$$= \frac{P(+, Y)}{P(+)}$$

$$P(+)=P(+, Y)+P(+, N)$$

$$P(+, Y)=P(+|Y)P(Y)$$

$$P(+, N)=P(+|N)P(N)$$

Law of total probability

$$P(A)=\sum_i P(A, B_i)=\sum_i P(B_i)P(A|B_i)$$

The answer is 0.0826...

- A test yields a positive result (+) **90%** of the time when the disease is **present** (Y)
- A yields a positive result **1%** when disease is **not** present (N)
- One person in **1,000** has the disease.

Q: What is the probability that a person with positive test has the disease?

22

Terminology

23

When we have two events A and B...

- Conditional probability: $P(A|B)$, $P(A^c|B)$, $P(B|A)$ etc.
- Joint probability: $P(A, B)$ or $P(A^c, B)$ or ...
- Marginal probability: $P(A)$ or $P(A^c)$

23

Conditional Probability

24

Tip: Make a table of **joint probabilities**

$$P(+ | Y) = 0.9$$

$$P(+ | N) = 0.01$$

$$P(Y) = 0.001$$

Each cell is $P(\text{column event} \cap \text{row event}) = P(T=t \cap D=d) = P(T=t | D=d) P(D=d)$

	Test = +	Test = -	
Disease=Y			0.001
Disease=N			0.999
	0.01089	0.98911	

Workflow:

- make a table, then fill in the cells.
- write down the target $P(A|B)$ all in terms of joint probabilities and marginal probabilities.

$P(\text{test} = +)$

24

Conditional Probability

25

We can directly calculate:

$$P(Y|+) = \frac{P(Y,+)}{P(+)} = \frac{P(+|Y)P(Y)}{P(+)}$$

Bayes rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{proof: definition and definition!}$$

⇒ particularly useful in practice: infer $P(A|B)$ given $P(B|A)$!

$P(A)$: **prior** probability e.g., A ='dice sum to 6', B ='one of the die is 2'

$P(A|B)$: **posterior** probability e.g., A ='disease=Y', B ='test=+'