DATA 133 - Introduction to Data Science I

Instructor: Renzhi Cao
Computer Science Department
Pacific Lutheran University

Announcements

- Check hint on course website for numpy example
- Today we are going to learn probabilities briefly

Reference book

- Data Science from Scratch - First Principles with Python. O'Reilly Media, 2015.

- Reading (Data Science from Scratch):

  - Read chapter 4: Linear Algebra
  - Read chapter 5: Statistics
  - Read chapter 6: Probability

Pictures from: https://www.google.com/search?

Introduction

- Probability

The laws of probability, so true in general, so fallacious in particular.
—Edward Gibbon

Probability

- It’s hard to do data science without probability
- Probability is used to quantify the uncertainty associated with events chosen from a universe of events.

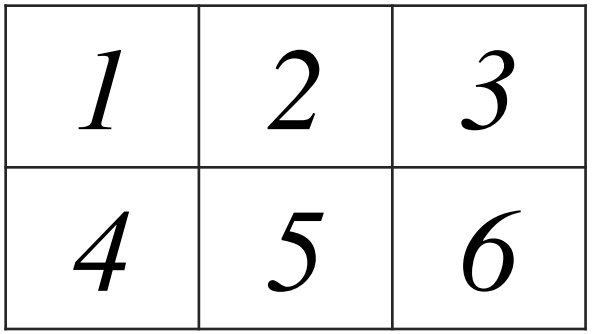
Universe: All possible outcomes
Event: A subset of those outcomes

Probability space

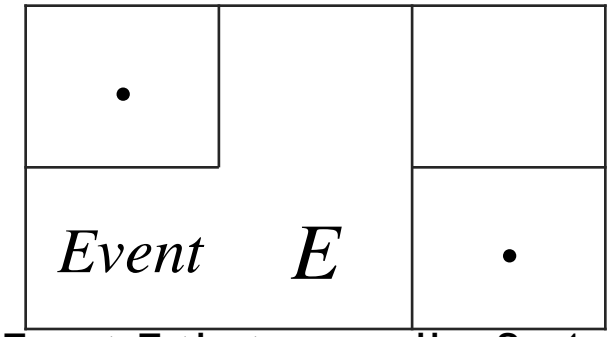
Finite set of points, whereby each of them represents a possible outcome of a specific experiment

- Each point (outcome) has a probability associated with it
- Probabilities are always positive!!!
- The sum of all probabilities is always 1
- Assume an equal probability distribution if not otherwise stated
  - e.g. 1/6 for a specific number of a die throw (unless die is not fair)

Probability space example



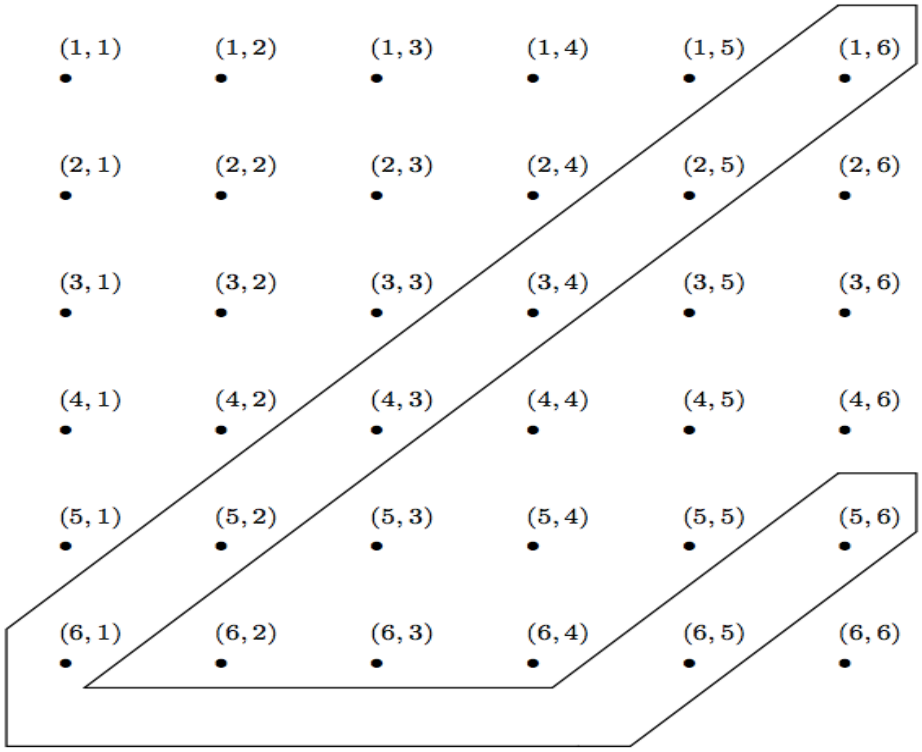
Probability space



Event E that you roll a 2, 4, or 5

Imagine you throw a dart randomly at the box. You will hit the area of E 50% of the time. P(E) = 0.5

Probability space example



Throw 2 dice and calculate the probability of obtaining a total of 7 or 11.

36 possible outcomes

The event contains 8 points.

$p = 8/36 \approx 22\%$

Dependence and Independence

The probability of an event E: P(E)
What about two events?

Events P and E are dependent if knowing something about whether E happens gives information about whether F happens.

Independent: The opposite
Tossing a coin two times: Dependent or independent?

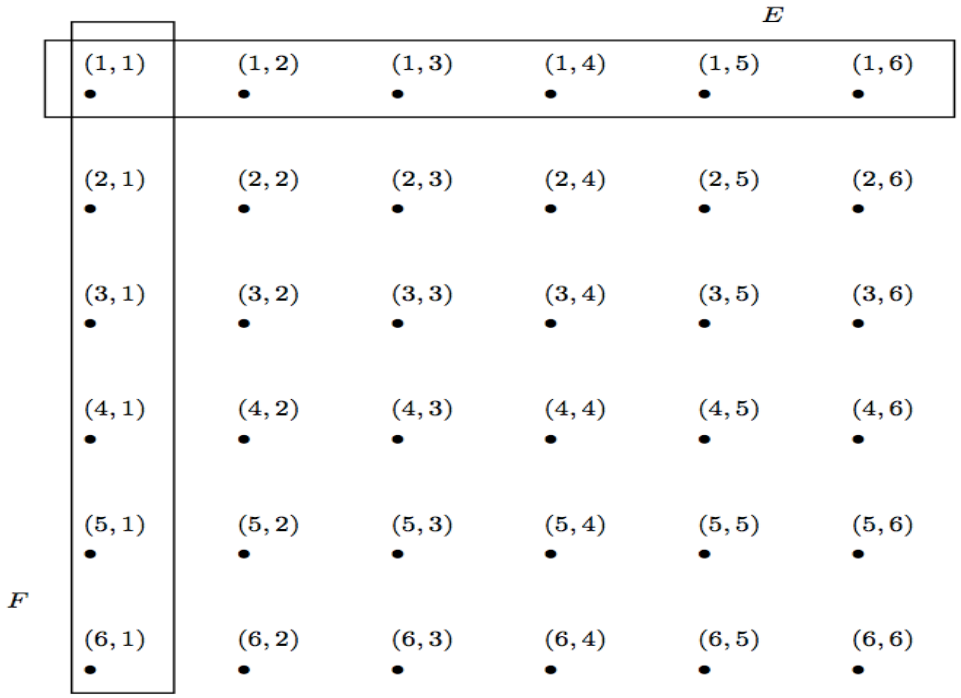
Independent Event

Independent Events: P(E,F) = P(E)P(F)
Example: Probability of getting two tails when flipping a coin two times. (Event E: first time gets tail. Event F: second time gets tail).
How to calculate P(E) and P(F)?
What is P(E,F)?

P(E,F) = ½ * ½ = ¼

Probability of getting a tail and a head:

Conditional probability



Toss of 2 dice

Probability space has 36 elements with equal probability 1/36

E: First comes out 1 (E_1)
F: Second comes out 1 (E_2)

P(E) = ? **= 6/36 = 1/6**
P(F) = ? **= 6/36 = 1/6**
P(F|E) = ? **= 1/6**

The experiments are **independent**, since **P(F) = P(F|E)**.
It does not matter if E occurred or not; the probability of F stays the same.

Conditional probability

Deal of 2 cards from a 52 card deck

Number of points in experiment (probability space):

$\Pi(52,2) = 52 \times 51 = \mathbf{2{,}652}$

E: First card is an ace: $4 \times 51 = \mathbf{204}$

(4 choices for ace, 51 choices for second card)

$P(E) = 204/2{,}652 = 1/13$

F: Second card is an ace: $4 \times 51 = \mathbf{204}$

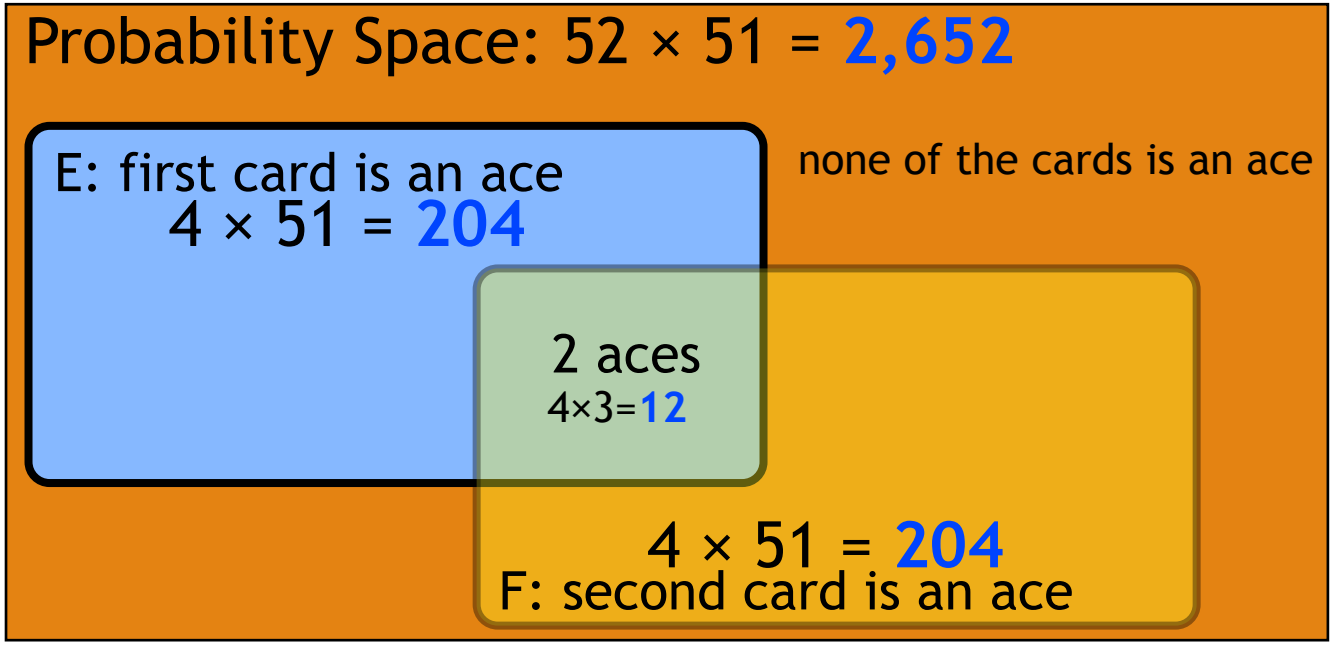
(4 choices for ace, 51 choices for first card)

$P(F) = 204/2{,}652 = 1/13$

$P(F|E) = 12/204 = 1/17$ $(= 3/51)$

since there are $4 \times 3 = \mathbf{12}$ combinations for aces.

Conditional probability



P(E) = 204/2,652

P(F) = 204/2,652

P(F|E) = 12/204

The experiments are **not independent**, since **P(F) ≠ P(F|E)**.
It **does** matter if E occurred or not; the probability of F changes.

Dependent Event

Dependent Events: $P(E|F) = P(E,F)/P(F)$, in which $P(E|F) \neq P(E)$
Examples: There are 5 marbles in a bag. 3 green and 2 red.
P(1st green) = ?

P(1st and 2nd green)= 9/25????

Nope!
They are dependent events
P(1st and 2nd green) = P(1st) * P(2nd green | 1st green)
3/5 * 2/4 = 3/10

You could list all possible outcomes:
1,2; 1,3; 1,4; 1,5; 2,3; 2,4; 2,5; 3,4; 3,5; 4,5
There are 10 overall, and 3 of them is two green marbles.

Practice

There are 300 students in the CS department. Of these students 90 play soccer, 30 play basketball, and 10 play both soccer and basketball. Let A be the event that a randomly selected student plays soccer and B be the event that the student plays basketball.

What is P(A)?

What is P(B)?

What is P(A and B)?

What is P(A|B)?

Break

Probability distribution

The normal distribution is the king of distributions: determined by two parameters - its mean μ (mu) and its standard deviation σ (sigma)

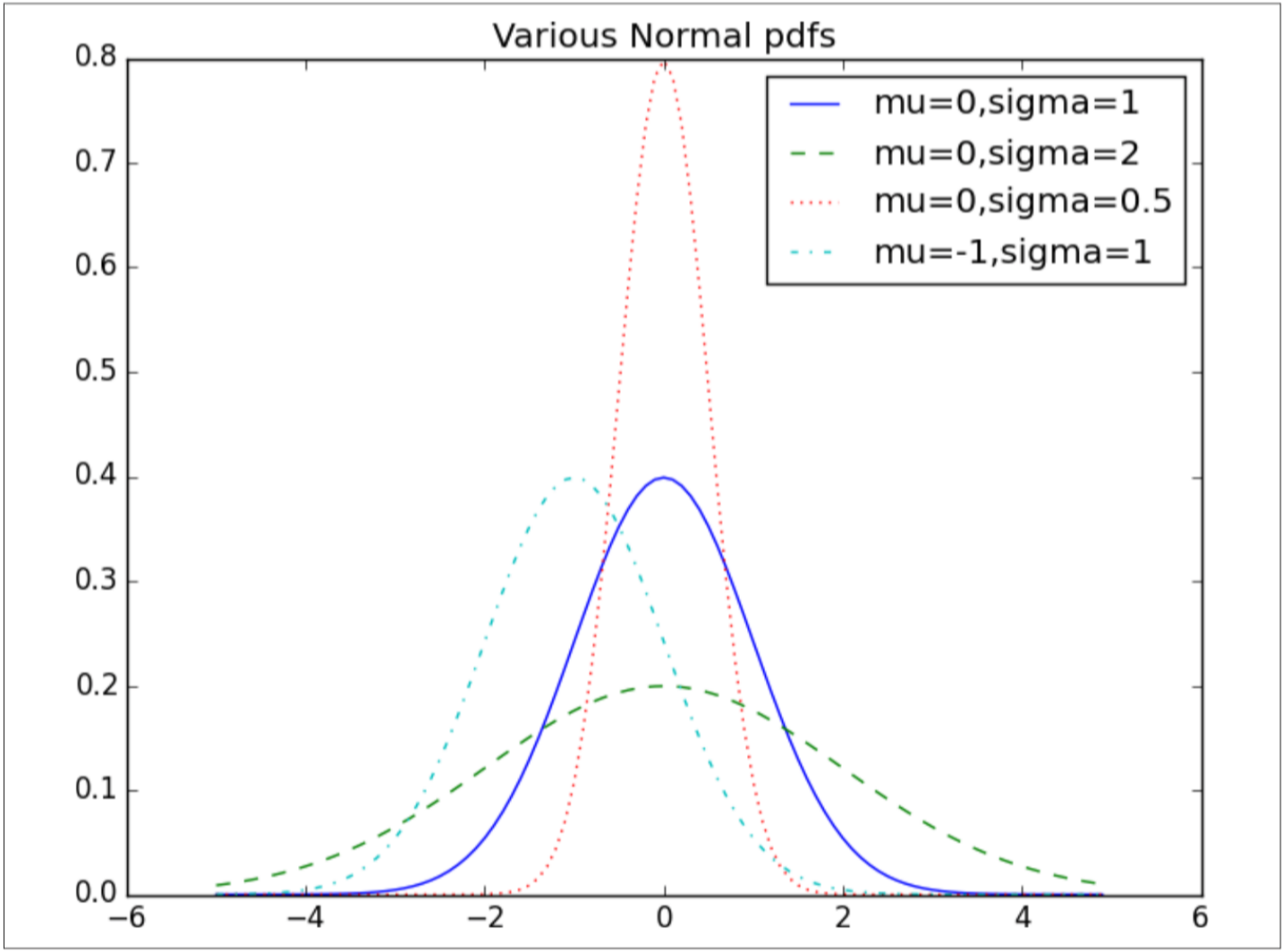
It has the distribution function:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

which we can implement as:

```
def normal_pdf(x, mu=0, sigma=1):
    sqrt_two_pi = math.sqrt(2 * math.pi)
    return (math.exp(-(x-mu) ** 2 / 2 / sigma ** 2) / (sqrt_two_pi * sigma))
```

```
xs = [x / 10.0 for x in range(-50, 50)]
plt.plot(xs,[normal_pdf(x,sigma=1) for x in xs],'-',label='mu=0,sigma=1')
plt.plot(xs,[normal_pdf(x,sigma=2) for x in xs],'--',label='mu=0,sigma=2')
plt.plot(xs,[normal_pdf(x,sigma=0.5) for x in xs],':',label='mu=0,sigma=0.5')
plt.plot(xs,[normal_pdf(x,mu=-1)   for x in xs],'-.',label='mu=-1,sigma=1')
plt.legend()
plt.title("Various Normal pdfs")
plt.show()
```



More Practice

In any 15-minute interval, there is a 20% probability that you will see a shuttle. What is the probability that you see at least one shuttle in the period of an hour?

More Practice

Solution:

Probability of not seeing a shuttle in 15 minutes is

= 1 – P(Seeing a shuttle)
= 1 – 0.2 = 0.8

Probability of not seeing any shuttle in the period of one hour

= (0.8) ^ 4 = 0.4096

Probability of seeing at least one shuttle in the one hour

= 1 – P(Not seeing any shuttle)
= 1 – 0.4096 = 0.5904

More Practice

How can you generate a random number between 1 – 7 with only a die?

More Practice

Solution:

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.
- To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.
- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.
- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

More Practice

A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

More Practice

Solution:

In the case of two children, there are 4 equally likely possibilities

BB, BG, GB and GG;

where B = Boy and G = Girl and the first letter denotes the first child.

From the question, we can exclude the first case of BB. Thus from the remaining 3 possibilities of BG, GB & GG, we have to find the probability of the case with two girls.

Thus, P(Having two girls given one girl) = 1 / 3

More Practice

You could validate it by “generating” a lot of families:

```
def random_kid():
    return random.choice(["boy", "girl"])

both_girls = 0
older_girl = 0
either_girl = 0
random.seed(0)
for _ in range(10000):
    younger = random_kid()
    older = random_kid()
    if older == "girl": older_girl += 1
    if older == "girl" and younger == "girl":  both_girls += 1
    if older == "girl" or younger == "girl":  either_girl += 1

print "P(both | older):", both_girls / older_girl # 0.514 ~ 1/2
print "P(both | either): ", both_girls / either_girl # 0.342 ~ 1/3
```

Homework of Python

Register Twitter and follow my account @cao_renzhi (caora@plu.edu). And get your consumer key and secret:

(1). Go to https://apps.twitter.com/.
(2). If you are not signed in, click Sign in and enter your Twitter username and password.
(3). Click Create New App.
(4). Give it a name (such as “Data Science”) and a description, and put any URL as the website (it doesn’t matter which one).
(5). Agree to the Terms of Service and click Create.
(6). Take note of the consumer key and consumer secret.

Twython

python -m pip install twython

Install twython library. Please don’t run it in the interactive mode of python, run it in command prompt if you could.

Twython

Instantiate the client:

```
import os
CONSUMER_KEY = "TWITTER_CONSUMER_KEY"
CONSUMER_SECRET = “TWITTER_CONSUMER_SECRET"

import webbrowser
from twython import Twython

# Get a temporary client to retrieve an authentication URL
temp_client = Twython(CONSUMER_KEY, CONSUMER_SECRET)
temp_creds = temp_client.get_authentication_tokens()
url = temp_creds['auth_url']

# Now visit that URL to authorize the application and get a PIN
print(f"go visit {url} and get the PIN code and paste it below")
webbrowser.open(url)
PIN_CODE = input("please enter the PIN code: ")
```

Twython

```
# Now we use that PIN_CODE to get the actual tokens
auth_client = Twython(CONSUMER_KEY,
                      CONSUMER_SECRET,
                      temp_creds['oauth_token'],
                      temp_creds['oauth_token_secret'])
final_step = auth_client.get_authorized_tokens(PIN_CODE)
ACCESS_TOKEN = final_step['oauth_token']
ACCESS_TOKEN_SECRET = final_step['oauth_token_secret']

# And get a new Twython instance using them.
twitter = Twython(CONSUMER_KEY,
                  CONSUMER_SECRET,
                  ACCESS_TOKEN,
                  ACCESS_TOKEN_SECRET)
```

Twython

At this point you may want to consider saving the ACCESS_TOKEN and ACCESS_TOKEN_SECRET somewhere safe, so that next time you don’t have to go through this rigmarole.

Google Cloud

Dear Students,

Here is the URL you will need to access in order to request a Google Cloud coupon. You will be asked to provide your school email address and name. An email will be sent to you to confirm these details before a coupon is sent to you.

Student Coupon Retrieval Link

- You will be asked for a name and email address, which needs to match your school domain. A confirmation email will be sent to you with a coupon code.
- You can request a coupon from the URL and redeem it until: 1/1/2024
- Coupon valid through: 9/1/2024
- You can only request ONE code per unique email address.

Please contact me if you have any questions or issues.

Thanks,
Prof. Renzhi Cao

Homework of Python

1. Read book chapter 6
2. Recommend to learn more details from statistics textbook.
3. No homework for today. Continue to work on Project 2.
4. A quiz on next Tuesday, study guide will be posted on Sakai.

Explore twitter robot on next Tuesday!