COMP7180: Quantitative Methods for DAAI

(Credits from Prof. Andrew Ng)

(Credits from HKBU)

Course Instructors: Dr. Yang Liu and Dr. Bo Han

Teaching Assistant: Mr. Minghao Li

Course Contents

- Continuous and Discrete Random Variables (Week 7)
- Conditional Probability and Independence (Week 8) ← Our Focus
- Maximum Likelihood Estimation (Week 9)
- Mathematical Optimization (Week 10)
- Convex and Non-Convex Optimization (Week 11)
- Course Review (Week 12)

Independence and Conditional Independence

- Two events E and F are independent if $P(E \cap F) = P(E)P(F)$.

- How is independence useful?

Suppose you have *n* coin flips, and you want to calculate the joint distribution $P(E1 \cap E2 \cap ... \cap En)$. If the coin flips are not independent, you need 2^n values in the table.

If the coin flips are independent, then

$P(E1 \cap E2 \cap ... \cap En) = P(E1)P(E2)...P(En)$.

Each P(Ei) table has 2 entries and there are n of them for a total of 2n values

Independence and Conditional Independence

- If events E and F are independent, then

$$P(E|F) = P(E\cap F)/P(F)=P(E)P(F)/P(F)=P(E).$$

So

$$P(E|F) = P(E)$$

We can also obtain that

$$P(F|E) = P(F)$$

Independence and Conditional Independence

- Two random variables X and Y are independent if their probability distribution can be expressed as a product of two factors:

$$P(X, Y) = P(X)P(Y)$$

If X and Y are both discrete random variables, then X and Y are independent, if for any x, y

$$P(X=x, Y=y) = P(X=x)P(Y=y)$$

We can also obtain that $P(X=x|Y=y) = P(X=x)$ and $P(Y=y|X=x)=P(Y=y)$.

Independence and Conditional Independence

If X and Y are continuous random variables, then X and Y are independent, if for any x, y,

$$p_{\mathrm{XY}}(x,y)=p_{\mathrm{X}}(x)p_{\mathrm{Y}}(y)$$

where $p_{\mathrm{XY}}(x,y)$ is the density function of the joint distribution, $p_{\mathrm{X}}(x)$ is the density function with respect to random variable X, and $p_{\mathrm{Y}}(x)$ is the density function with respect to random variable Y.

Independence: An Example

- Toss a coin twice. The second time won't be affected by the first time.

- If we set $X1 = 1$ if the outcome is head at the first time; $X1 = -1$ if the outcome is tail at the first time.

- If we set $X2 = 1$ if the outcome is head at the second time; $X2 = -1$ if the outcome is tail at the second time.

- $X1$ and $X2$ are independent.

Exercise

In this table, please answer

1. Are X and Y independent?

2. Are Y and Z independent?

3. Are X and Z independent?

X	Y	Z	P(X,Y,Z)
0	0	0	0.1
0	0	1	0.2
0	1	0	0.05
0	1	1	0.05
1	0	0	0.3
1	0	1	0.1
1	1	0	0.05
1	1	1	0.15

Credits from Dr. Liu Yang

Exercise

1. $P(X=1) = 0.6$, $P(Y=1)=0.3$
   $P(X=1,Y=1) = 0.2$
   $P(X=1,Y=1)>P(X=1)P(Y=1)$.
   So X and Y are not independent.
2. $P(Z=1) = 0.5$, $P(Y=1,Z=1)=0.2$
   $P(Z=1)P(Y=1)<P(Y=1,Z=1)$.
   So Y and Z are not independent.
3. $P(X=1, Z=1)=0.25$
   $P(Z=1)P(X=1)>P(X=1,Z=1)$
   So X and Z are not independent.

X	Y	Z	P(X,Y,Z)
0	0	0	0.1
0	0	1	0.2
0	1	0	0.05
0	1	1	0.05
1	0	0	0.3
1	0	1	0.1
1	1	0	0.05
1	1	1	0.15

Credits from Dr. Liu Yang

Exercise

In this table, please answer

Are X and Y independent?

X	Y	P(X,Y)
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

- Solution:

P(X=1)=P(Y=1)=P(X=0)=P(Y=0)=0.5

P(X=1,Y=1)=P(X=1,Y=0)=P(X=0,Y=1)=P(X=0,Y=0)=0.25

So it is easy to check that for any i=1 or 2 and j=1 or 2

P(X=i,Y=j)=P(X=i)P(Y=j)=0.25

So X and Y are independent.

Exercise

In this table, please answer

1. Are X and Y independent?

2. Are Y and Z independent?

X	**Y**	**Z**	**P(X,Y,Z)**
0	0	0	0.05
0	0	1	0.2
0	1	0	0.1
0	1	1	0.15
1	0	0	0.1
1	0	1	0.15
1	1	0	0.05
1	1	1	0.2

Exercise

- Solution

1. Let i=1 or 2 and j=1 or 2
$P(X=i)P(Y=j)=P(X=i,Y=j)=0.25$
So X and Y are independent.

2. $P(Y=1)=0.5$, $P(Z=1)=0.7$,
$P(Y=1,Z=1)=0.35=P(Y=1)P(Z=1)$
$P(Y=1,Z=0)=0.15=P(Y=1)P(Z=0)$
$P(Y=0,Z=1)=0.35=P(Y=0)P(Z=1)$
$P(Y=0,Z=0)=0.15=P(Y=0)P(Z=0)$
So Y and Z are independent.

X	**Y**	**Z**	**P(X,Y,Z)**
0	0	0	0.05
0	0	1	0.2
0	1	0	0.1
0	1	1	0.15
1	0	0	0.1
1	0	1	0.15
1	1	0	0.05
1	1	1	0.2

Independence and Conditional Independence

Given three events A, B and C, we say A and B are independent with respect to C, if

$$P(A \cap B | C) = P(A|C)P(B|C)$$

It is easy to check that if A and B are independent with respect to C, then

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

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

Independence and Conditional Independence

1. $P(A \cap B|C) = P(A|C)P(B|C)$
2. $P(A|B,C) = P(A|C)$
3. $P(B|A,C) = P(B|C)$

How to derive the first equation from the second or third one?

Independence and Conditional Independence

How to derive

$P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C)$ from $P(A \mid B, C) = P(A \mid C)$ or $P(B \mid A, C) = P(B \mid C)$?

$$P(A \mid B, C) = P(A \mid C)$$

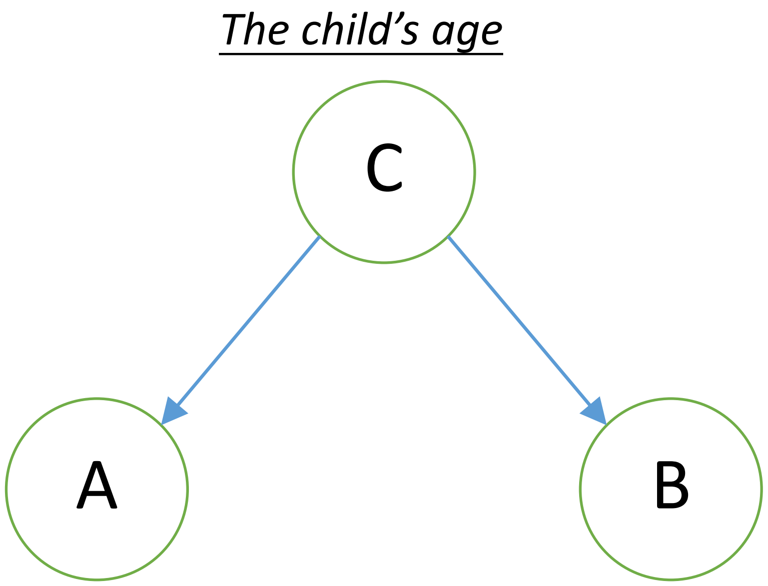
$$P(A \cap B \cap C) / P(B \cap C) = P(A \cap C) / P(C)$$

$$P(A \cap B \cap C) / P(C) = (P(A \cap C) / P(C)) * (P(B \cap C) / P(C))$$

$$P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C)$$

Credits from Dr. Liu Yang

Conditional Independence: An Example



A is the *height of a child* and
B is the *number of words that the child knows*.
It seems when A is high, B is high too.

There is, however, a single piece of information that will make A and B completely independent. What would that be?
The child’s age!

The height and the # of words known by the kid are NOT independent, but they are conditionally independent if you provide the kid’s age.

Credits from Dr. Liu Yang

Conditional Independence: An Example

- A box contains two coins: a regular coin and one fake two-headed coin ($P(H)=1$). I choose a coin at random and toss it twice. Define the following events.

A = First coin toss results in an H.

B = Second coin toss results in an H.

C = Coin 1 (regular) has been selected.

Find $P(A|C), P(B|C), P(A \cap B|C), P(A), P(B)$, and $P(A \cap B)$.

Show that A and B are NOT independent, but they are conditionally independent given C.

Credits from Dr. Liu Yang

Conditional Independence: An Example

A = First coin toss results in an H.

B = Second coin toss results in an H.

C = Coin 1 (regular) has been selected.

We have $P(A|C)=P(B|C)=1/2$. Also, given that Coin 1 is selected, we have $P(A,B|C)=(1/2)*(1/2)=1/4$. So A and B are conditionally independent on C!

To find $P(A)$,$P(B)$, and $P(A\cap B)$, we use the law of total probability:

$P(A) =P(A|C)P(C)+P(A|C^c)P(C^c) =(1/2)*(1/2)+1*(1/2) =3/4$.

Similarly, $P(B)=3/4$. For $P(A\cap B)$, we have

$P(A\cap B)= P(A\cap B|C)P(C)+P(A\cap B|C^c)P(C^c) = P(A|C)P(B|C)P(C) + P(A|C^c)P(B|C^c)P(C^c) = (1/2)*(1/2)*(1/2)+1*1*(1/2)=5/8$.

Credits from Dr. Liu Yang

Conditional Independence: An Example

A = First coin toss results in an H.

B = Second coin toss results in an H.

C = Coin 1 (regular) has been selected.

As we see, $P(A \cap B)=5/8 \neq P(A)*P(B) = 9/16$, which means that A and B are not independent. We can justify this intuitively:

If we know A has occurred (i.e., the first coin toss has resulted in heads), we would guess that it is more likely that we have chosen Coin 2 than Coin 1. This in turn increases the conditional probability that B occurs. This suggests that A and B are not independent. On the other hand, given C (Coin 1 is selected), A and B are independent.

Credits from Dr. Liu Yang

Conditional Independence: An Example

One important lesson:

Conditional independence neither implies (nor is it implied by) independence.

Thus, we can have two events that are conditionally independent but they are not unconditionally independent (such as A and B above). **Also, we can have two events that are independent but not conditionally independent, given an event C**.

Credits from Dr. Liu Yang

Conditional Independence: An Example

Alarm System Example

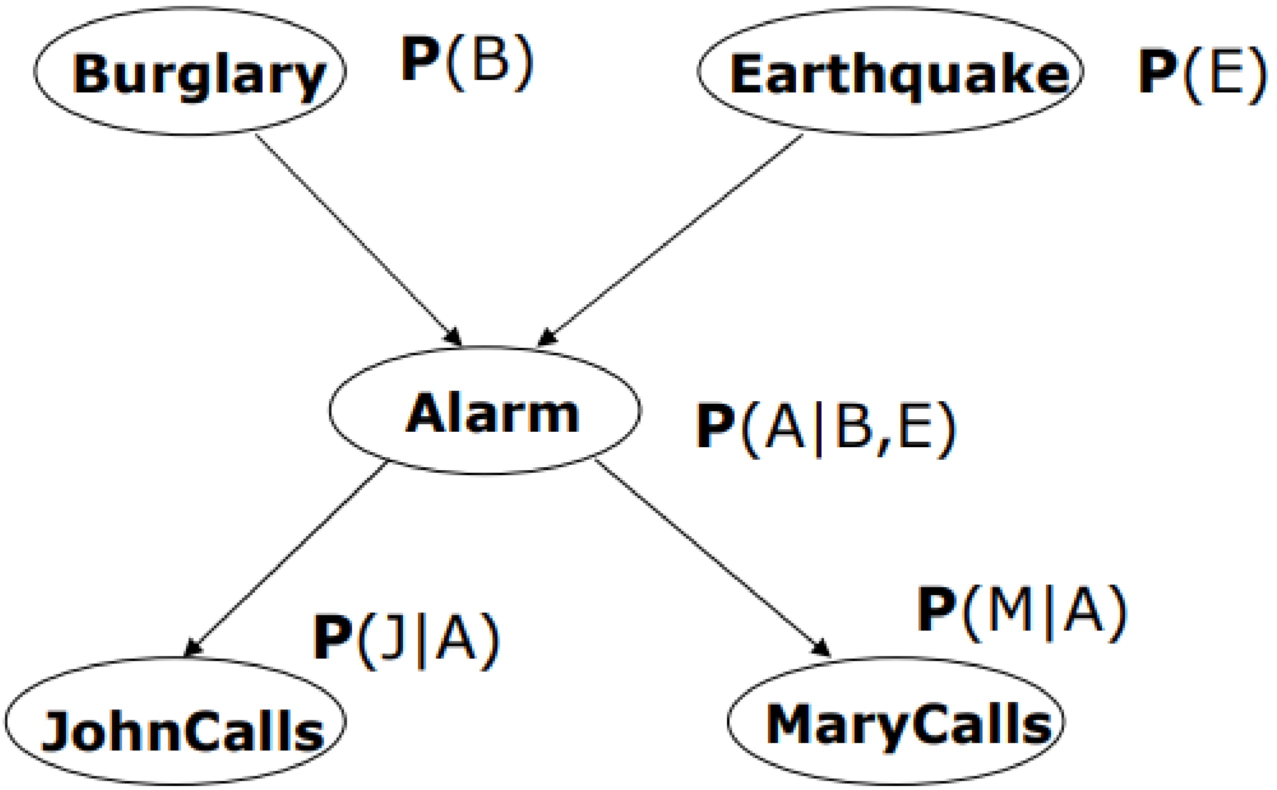
- Assume your house has an **alarm system** against **burglary**. You live in the seismically active area and the alarm system can get occasionally set off by an **earthquake**. You have two neighbors, **Mary** and **John**, who do not know each other. If they hear the alarm they might call you, but this is not guaranteed.

- Now we represent the probability distribution of events:
    - Burglary, Earthquake, Alarm, Mary calls, and John calls

Credits from Credit: CS 2740 Knowledge Representation, M. Hauskrecht

Conditional Independence: An Example

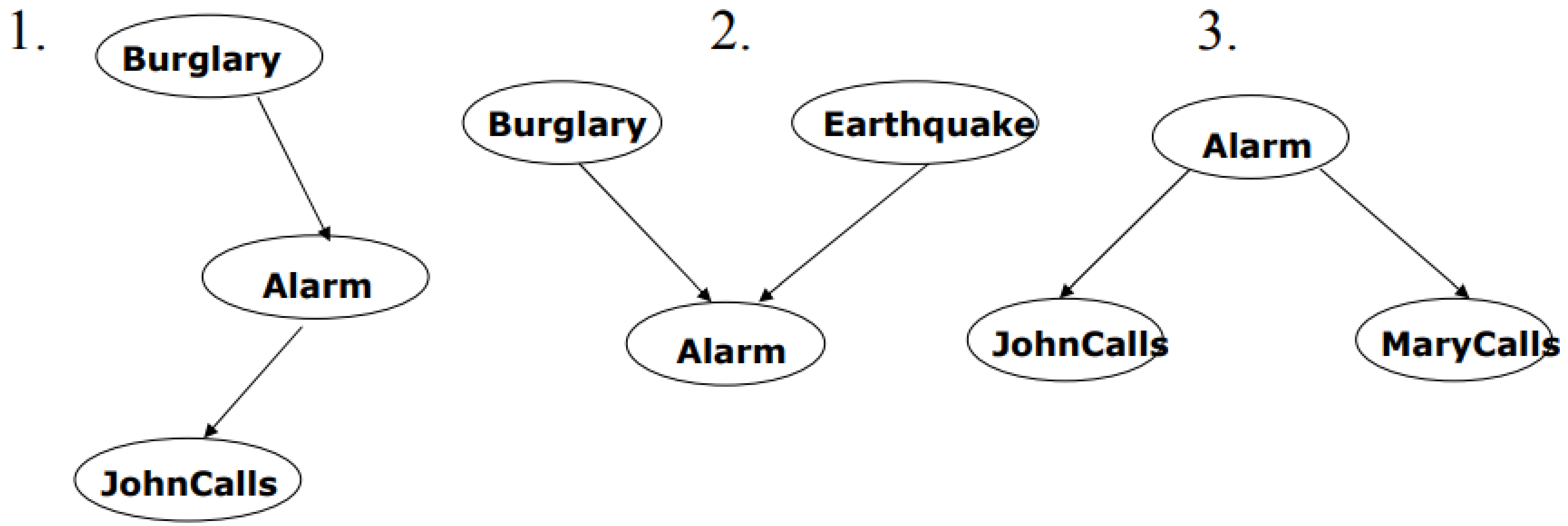
Alarm System Example



Credits from Credit: CS 2740 Knowledge Representation, M. Hauskrecht

Conditional Independence: An Example

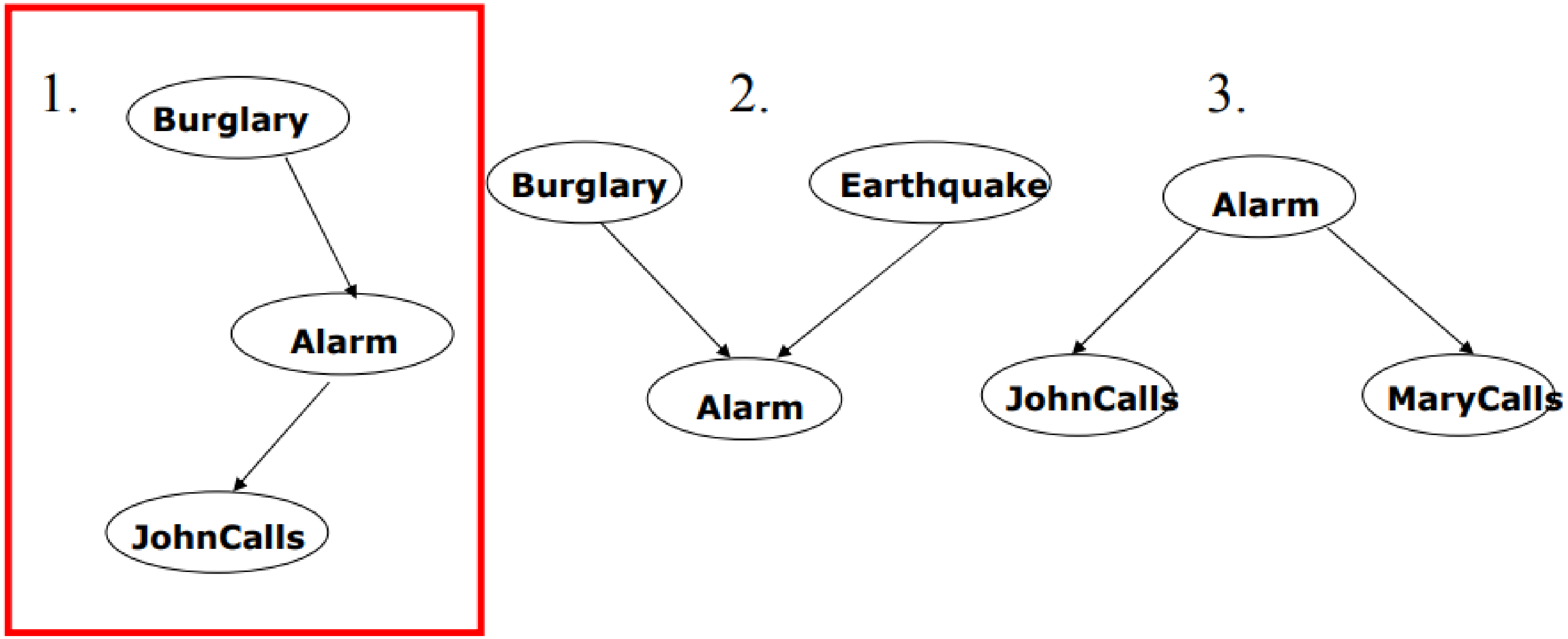
Alarm System Example: 3 basic structures



Credits from Credit: CS 2740 Knowledge Representation, M. Hauskrecht

Conditional Independence: An Example

Alarm System Example: 3 basic structures



1. JohnCalls **is independent** of Burglary given Alarm

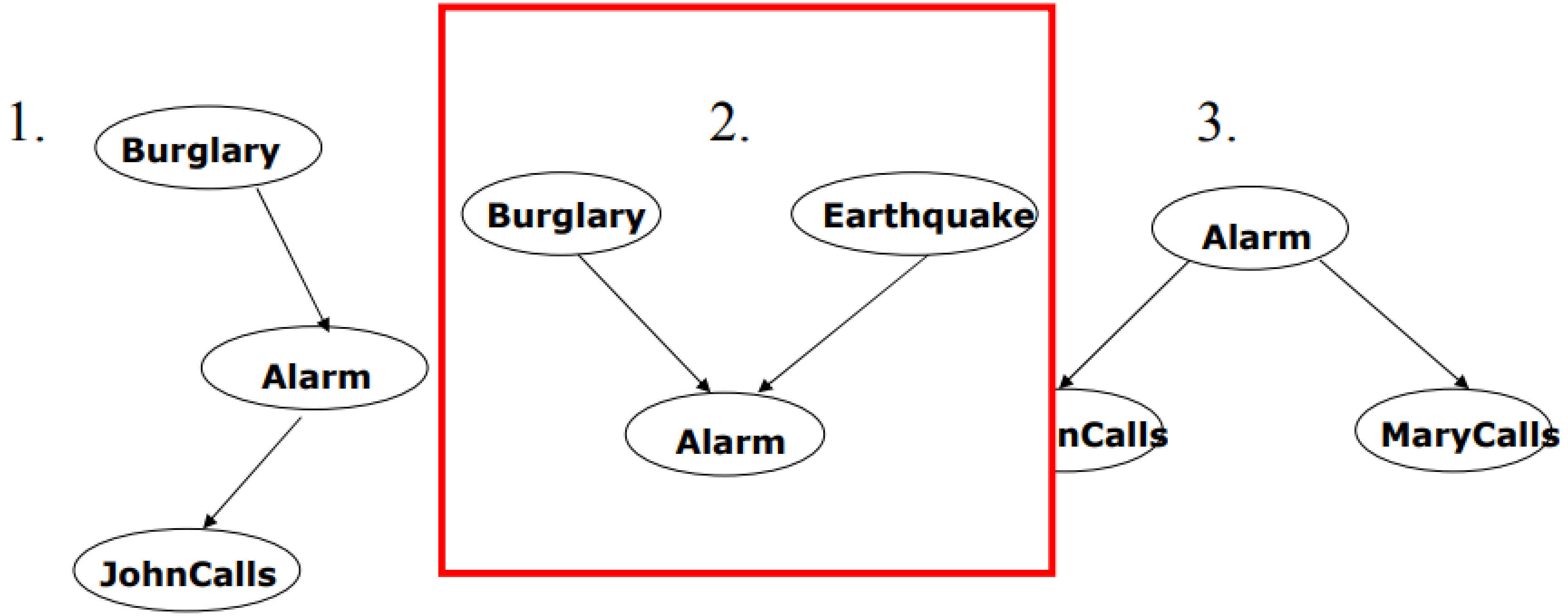
$$P(J \mid A, B) = P(J \mid A)$$

$$P(J, B \mid A) = P(J \mid A) P(B \mid A)$$

Credits from Credit: CS 2740 Knowledge Representation, M. Hauskrecht

Conditional Independence: An Example

Alarm System Example: 3 basic structures



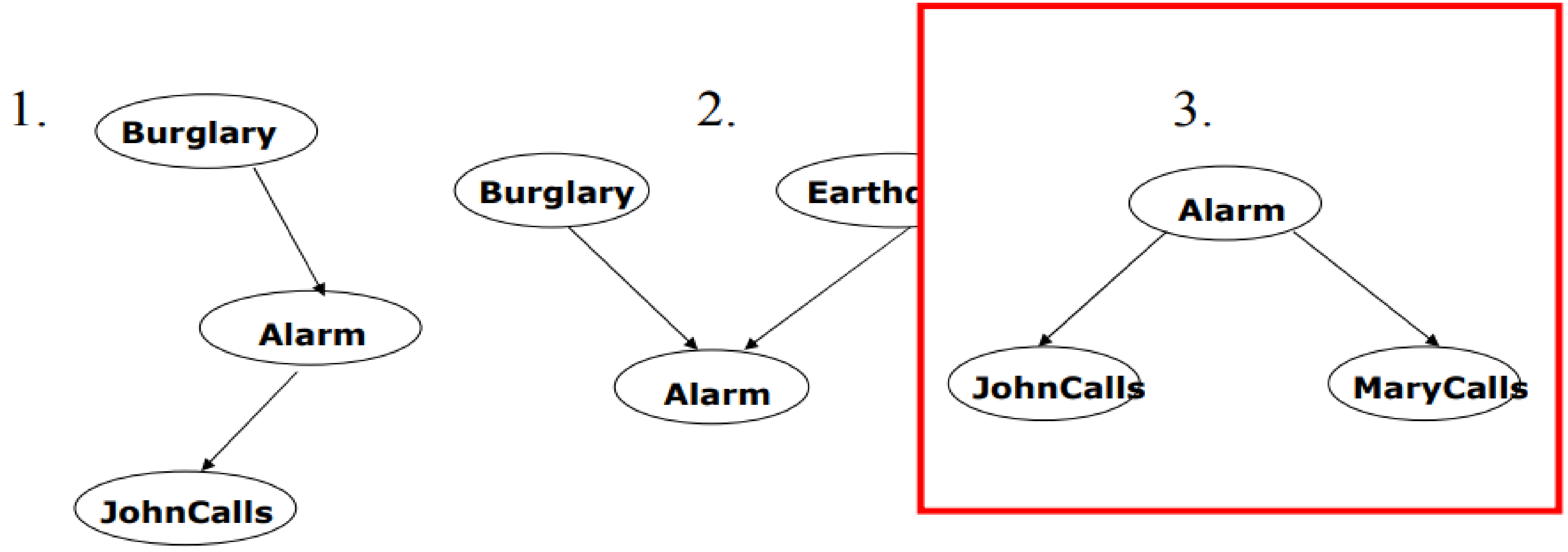
2. Burglary **is independent** of Earthquake (not knowing Alarm)
   Burglary and Earthquake **become dependent** given Alarm !!

$$P(B,E) = P(B)P(E)$$

Credits from Credit: CS 2740 Knowledge Representation, M. Hauskrecht

Conditional Independence: An Example

Alarm System Example: 3 basic structures



3. MaryCalls **is independent** of JohnCalls given Alarm

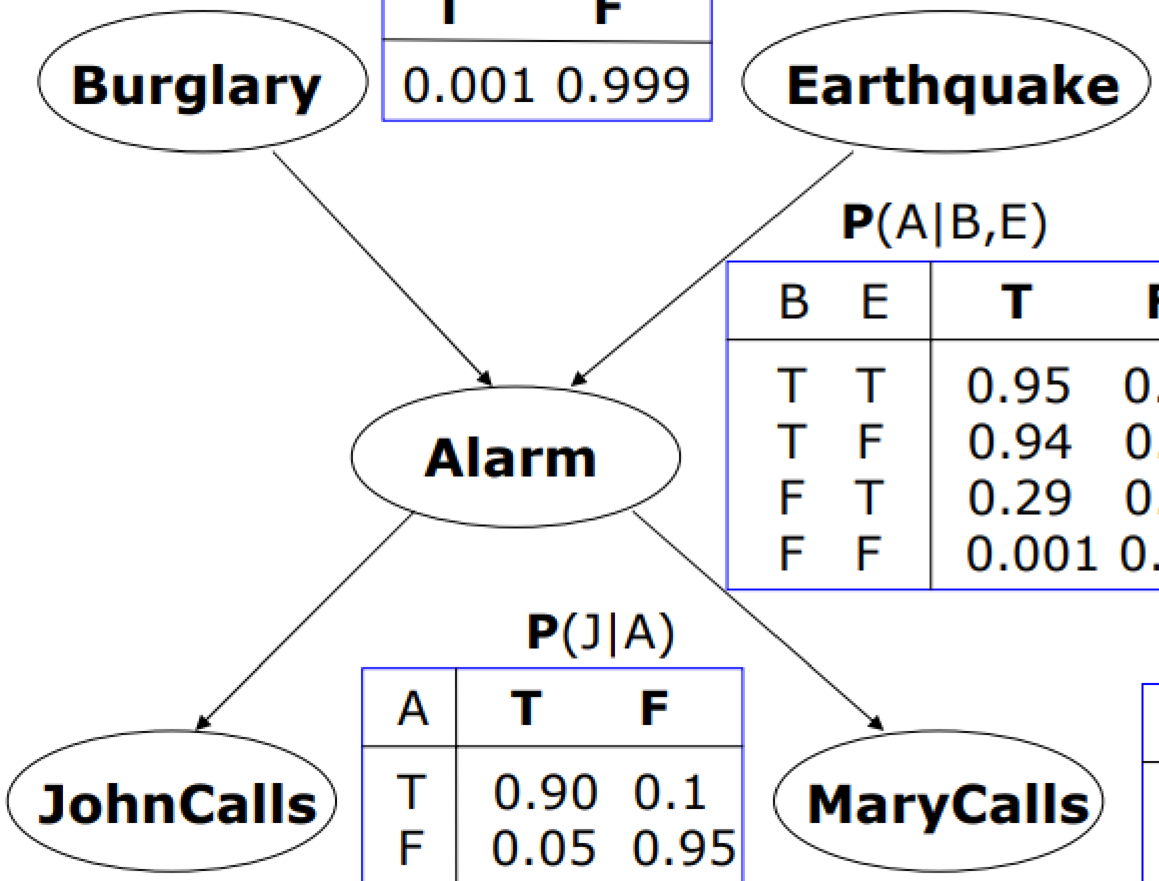
$$P(J \mid A, M) = P(J \mid A)$$

$$P(J, M \mid A) = P(J \mid A) P(M \mid A)$$

Credits from Credit: CS 2740 Knowledge Representation, M. Hauskrecht

Conditional Independence: An Example

Alarm System Example: 3 basic structures



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Credits from Credit: CS 2740 Knowledge Representation, M. Hauskrecht

Conditional Independence to Random Variables

Two random variables X and Y are conditionally independent given a random variable Z if the conditional probability distribution

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

When X, Y and Z are discrete variable variables, if X and Y are conditionally independent given a random variable Z: for any x, y, z

$$P(X=x,Y=y|Z=z) = P(X=x|Z=z)P(Y=y|Z=z)$$

Conditional Independence to Random Variables

When X, Y and Z are continouos variable variables, if X and Y are conditionally independent given a random variable Z: for any x, y, z,

then the probability density functions satisfy that

$$p_{\mathrm{XY|Z}}(\mathrm{x},\mathrm{y}|\mathrm{z}) = p_{\mathrm{X|Z}}(\mathrm{x}|\mathrm{z})\, p_{\mathrm{Y|Z}}(\mathrm{y}|\mathrm{z})$$

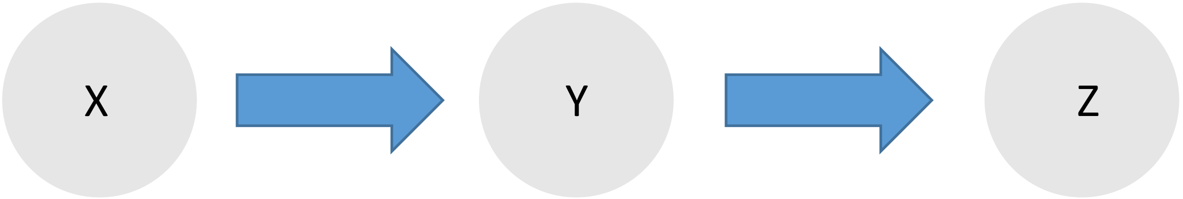
So

$$\frac{p_{\mathrm{XYZ}}(\mathrm{x,y,z})}{p_{\mathrm{Z}}(\mathrm{z})} = \frac{p_{\mathrm{XZ}}(\mathrm{x,z})}{p_{\mathrm{Z}}(\mathrm{z})} \frac{p_{\mathrm{YZ}}(\mathrm{y,z})}{p_{\mathrm{Z}}(\mathrm{z})}$$

$$p_{\mathrm{XYZ}}(\mathrm{x},\mathrm{y},\mathrm{z}) = p_{\mathrm{XZ}}(\mathrm{x},\mathrm{z})\, p_{\mathrm{YZ}}(\mathrm{y},\mathrm{z})$$

Exercises

- Given random variables X, Y, and Z



This graph is associated with $P(X,Y,Z) = P(Z|Y)P(Y|X)P(X)$

To show that X and Z are conditional independent given Y

Credits from Prof. Schmidt-Thieme, Randolf Scholz

Exercises

- Solution: By chain rule, $P(X, Y, Z)=P(X)P(Y|X)P(Z|X,Y)$.

Because $P(X, Y, Z) = P(X)P(Y|X)P(Z|Y)$.

We obtain that $P(Z|X,Y)=P(Z|Y)$. So $P(Z|X,Y)P(X|Y) = P(X|Y)P(Z|Y)$.

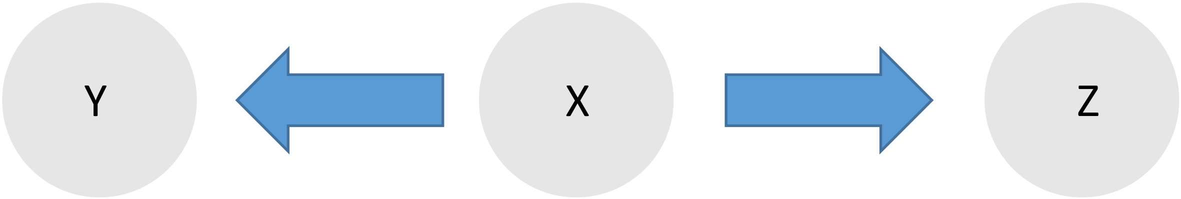
Note that $P(Z|X,Y)P(X|Y)=P(X,Z|Y)$, so $P(X,Z|Y)=P(X|Y)P(Z|Y)$.

Answer: X and Z are independent given Y.

Credits from Prof. Schmidt-Thieme, Randolf Scholz

Exercises

- Given random variables X, Y, and Z



This graph is associated with $P(X,Y, Z) = P(Y|X)P(Z|X)P(X)$

To show that Y and Z are conditional independent given X

Exercises

- Solution: By chain rule, $P(X, Y, Z)=P(X)P(Y|X)P(Z|X,Y)$.

Because $P(X, Y, Z) = P(X)P(Y|X)P(Z|X)$.

We obtain that $P(Z|X,Y)=P(Z|X)$. So $P(Z|X,Y)P(Y|X) = P(Z|X)P(Y|X)$.

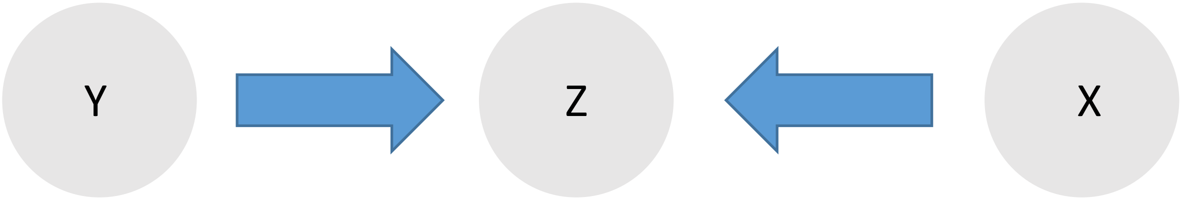
Note that $P(Z|X,Y)P(Y|X)=P(Y,Z|X)$, so $P(Y,Z|X)=P(Z|X)P(Y|X)$.

Answer: Y and Z are independent given X.

Credits from Prof. Schmidt-Thieme, Randolf Scholz

Exercises

- Given random variables X, Y, and Z



This graph is associated with $P(X,Y,Z) = P(Z|X,Y)P(X)P(Y)$

To show that Y and X are independent.

Credits from Prof. Schmidt-Thieme, Randolf Scholz

Exercises

- Solution: By chain rule, $P(X, Y, Z)=P(X)P(Y|X)P(Z|X,Y)$.

Because $P(X, Y, Z) = P(Z|X,Y)P(X)P(Y)$.

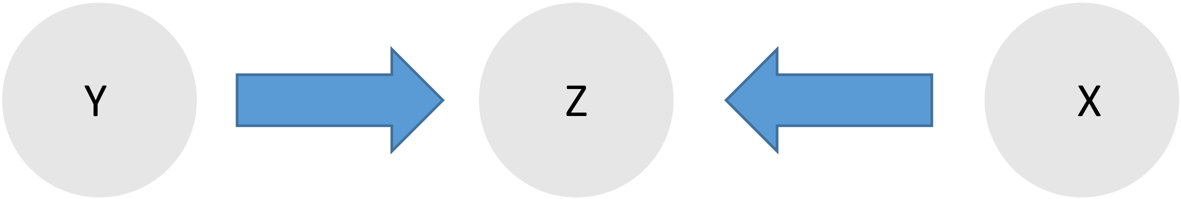
We obtain that $P(Y|X)=P(Y)$. So $P(X,Y)=P(Y|X)P(X)=P(Y)P(X)$.

Answer: X and Y are independent.

Credits from Prof. Schmidt-Thieme, Randolf Scholz

Exercises

- Given random variables X, Y, and Z



This graph is associated with $P(X,Y,Z) = P(Z|X,Y)P(X)P(Y)$

Are Y and X independent given Z?

Exercises

- Solution: There exists cases that X and Y are not independet given Z.

Case. Suppose that Z indicates whether our lawn is wet one morning; X and Y are two explanations for it being wet: either it rained (indicated by X), or the sprinkler turned on (indicated by Y).

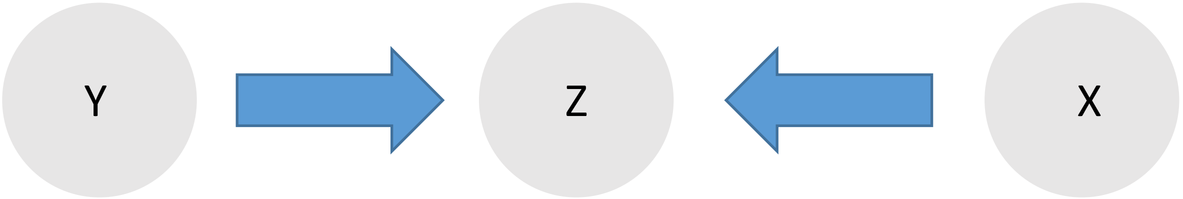
If we know that the grass is wet (Z is true) and the sprinkler didn’t go on (Y is false), then the probability that X is true must be one, because that is the only other possible explanation.

Hence, in this case, X and Y are not independent given Z.

Credits from Prof. Schmidt-Thieme, Randolf Scholz

Exercises

- Given random variables X, Y, and Z



X and Y are independent. But there exists case that X and Y are not independent given Z.

This exercise indicates that
It is possible that X and Y are not independent given some other random variables, even if X and Y are independent.

Credits from Prof. Schmidt-Thieme, Randolf Scholz

Bayes' Rule

- Bayes' theorem is an important tool in statistics and machine learning:

Given events A and B, then $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Proof:

$$P(A|B) = P(A \cap B)/P(B) = (P(B|A)P(A))/P(B).$$

Bayes' Rule

- How to use Baye's Rule?

Example: A bag I contains 4 white and 6 black balls while another bag II contains 4 white and 3 black balls.

One ball is drawn at random from one of the bags, and it is found to be black.

Find the probability that it was drawn from bag I.

Bayes’ Rule

Soultion: Let E1 be the event of choosing bag I, E2 the event of choosing bag II, and A be the event of drawing a black ball.

Then, $P(E1) = P(E2) = 0.5$.

$P(A|E1)$ = P(drawing a black ball from Bag I) = 6/10 = 3/5.

$P(A|E2)$ = P(drawing a black ball from Bag II) = 3/7

$P(A) = P(A|E1)P(E1)+P(A|E2)P(E2) = 18/35$

By using Bayes’ theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(E1|A) = P(A|E1)P(E1)/P(A) = 0.6*0.5/(18/35) = 7/12$$

Bayes' Rule

- How to use Baye's Rule?

Example: A man is known to speak the truth 2 out of 3 times. He throws a dice and reports that the number obtained is a four. Find the probability that the number obtained is four.

Bayes' Rule

Soultion: Let A be the event that the man reports that number four is obtained. Let E1 be the event that four is obtained and E2 be its complementary event.

Then, $P(E1)$ = Probability that four occurs = 1/6.

$P(E2)$ = Probability that four does not occur = 1- P(E1) = 1 – (1/6) = 5/6.

$P(A|E1)$= Probability that man reports four and it is actually a four = 2/3

$P(A|E2)$ = Probability that man reports four and it is not a four = 1/3.

So $P(A) = P(A|E1)P(E1)+P(A|E2)P(E2)$ = 1/9+5/18 = 7/18.

By using Bayes' theorem, probability that number obtained is actually a four, $P(E1|A) = P(A|E1)P(E1)/P(A) = 2/18/(7/18) = 2/7$.

Expectation (or Mean)

The expectation or mean of a random variable X *is denoted by* E[X] *and defined as:*

- For discrete variable,

$$\mathrm{E}[X] = \sum_x \mathrm{P}(\mathrm{X} = \mathrm{x})\mathrm{x}$$

- For continuous variable,

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} p_{\mathrm{X}}(\mathrm{x})\mathrm{x}\, \mathrm{dx}$$

In words, we are taking a weighted sum of the values that x can take on, where the weights are the probabilities of those respective values. The expected value has a physical interpretation as the “center of mass” of the distribution.

Expectation (or Mean): An Example

- Assume that X is a discrete random variable with 4 possible values: x_1, x_2, x_3, x_4, and with the equal probability on these 4 values. What is the expectation of X?

Solution: According to the definition, we have

$$E[X] = \sum_x P(X = x)x$$

$$= \frac{1}{4}x_1 + \frac{1}{4}x_2 + \frac{1}{4}x_3 + \frac{1}{4}x_4$$

$$= \frac{x_1+x_2+x_3+x_4}{4}$$

Credits from Dr. Liu Yang

Expectation (or Mean): An Example

- An exercise:

Let us consider an European roulette game that players can bet on any single number from 1, 2, …, 36. The number 0 is considered as winning for the casino. The probability of the appearance of any number from 0-36 is equal, i.e., 1/37.

Bob places 1 dollar on a specific number. If he wins, he can take the 1 dollar back and gains the extra 35 dollars from the casino; if he loses, he will lose the 1 dollar he bet on that number.
Now, the question is: what is the expectation of the gain of Bob?

Credits from Dr. Liu Yang

Expectation (or Mean): An Example

- An exercise:

Let us consider an European roulette game that players can bet on any single number from 1, 2, …, 36. The number 0 is considered as winning for the casino. The probability of the appearance of any number from 0-36 is equal, i.e., 1/37.

Bob places 1 dollar on a specific number. If he wins, he can take the 1 dollar back and gains the extra 35 dollars from the casino; if he loses, he will lose the 1 dollar he bet on that number. ***<u>Now, the question is: what is the expectation of the gain of Bob?</u>***

Solution:

$E[X] = (35)\times(1/37) + (-1)\times(36/37)$

$= -1/37$

Credits from Dr. Liu Yang

Expectation of Functions

The expectation or mean of f(X) (a function of random variable X) is denoted by E[f(X)] and defined as:

For discrete variable,

$$E[f(X)] = \sum_{x} P(X = x) f(x)$$

For continuous variable,

$$E[f(X)] = \int_{-\infty}^{+\infty} p_X(x) f(x) \, dx$$

Properties of Expectation

For any two random variables X and Y, functions f and g, and any constants a, b ∈ R, the following equations hold:

- $E[a] = a$, $E[f(a)] = f(a)$
- $E[X+Y] = E[X]+E[Y]$ $E[f(X)+g(Y)] = E[f(X)]+E[g(X)]$
- $E[aX] = aE[X]$ $E[af(X)] = aE[f(X)]$
- $E[aX+bY]=aE[X]+bE[Y]$ $E[af(X)+bg(Y)]=aE[g(X)]+bE[g(Y)]$

Properties of Expectation

- For any two random variables X and Y and any constants a, b ∈ R, the following equations hold:

$$E[a] = a \quad (1)$$

Proof: From the definition of expectation, we have

$$E[a] = aP(X=a)=a \times 1 = a.$$

So, we have: the expectation of a constant is the constant itself.

Properties of Expectation

- For any two random variables X and Y and any constants a, b ∈ R, the following equations hold:

$$E[aX] = aE[X] \tag{2}$$

Proof: From the definition of expectation, we have: if X is disctete,

$$E[aX] = \sum_x P(X = x)ax = a\sum_x P(X = x)x = aE[X].$$

If X is continuous, the proof is similar.

Properties of Expectation

- For any two random variables X and Y and any constants a, b ∈ R, the following equations hold:

$$E[X + Y] = E[X] + E[Y] \tag{3}$$

Proof: From the definition of expectation, we have: if X and Y are disctete,

$$E[X + Y] = \sum_x \sum_y (x + y) P(X = x, Y = y)$$

$$= \sum_x \sum_y x P(X = x, Y = y) + \sum_x \sum_y y P(X = x, Y = y)$$

$$= \sum_x x\left(\sum_y P(X = x, Y = y)\right) + \sum_y y\left(\sum_x P(X = x, Y = y)\right)$$

$$= \sum_x x P(X = x) + \sum_y y P(Y = y) = E[X] + E[Y]$$

If X and Y are continuous, the proof is similar.

Credits from Dr. Liu Yang

Properties of Expectation

- For any two random variables X and Y and any constants a, b ∈ R, the following equations hold:

$$E[aX + bY] = E[aX] + E[bY] \qquad (4)$$

Proof: From the definition of expectation, we have: if X and Y are disctete,

$$E[aX + bY] = E[aX] + E[bY]$$

$$= aE[X] + bE[Y]$$

If X and Y are continuous, the proof is similar.

Properties of Expectation

- For any two random variables X and Y, and functios f and g and any constants a, b ∈ R, the following equations hold:

(5)

- $E[f(X)+g(Y)] = E[f(X)]+E[g(X)]$
- $E[af(X)] = aE[f(X)]$
- $E[af(X)+bg(Y)]=aE[g(X)]+bE[g(Y)]$

Proof: Note that
f(X) and g(Y) are both random variables, therefore, using the properites (2), (3) and (4), we can prove the results.

Expectation: Exercise

For any two random variables X and Y, if X and Y are independent, then the following equations hold:

$$E[XY]=E[X]E[Y]$$

Expectation: Exercise

For any two random variables X and Y, if X and Y are independent, then the following equations hold:

$$E[XY]=E[X]E[Y]$$

Solution: According to the definition, we have

$$E[XY] = \sum_x \sum_y (xy)P(X = x, Y = y) .$$

Since X and Y are independent, we know that

$P(X = x, Y = y) = P(X = x)P(Y = y)$. Therefore, we have

$$E[XY] = \sum_x \sum_y (xy)P(X = x)P(Y = y)$$

$$= \left(\sum_x xP(X = x)\right) \left(\sum_y yP(Y = y)\right) = E[X]E[Y].$$

Credits from Dr. Liu Yang

Variance

- Expectation provides measure of the “center” of a distribution, but sometimes we are also interested in what the “spread” is about that center. Therefore, we define the variance Var(X) of a random variable X as follows:

$$\mathrm{Var}(X) = E[(X - E[X])^2]$$

- In words, this is the average squared deviation of the values of X from the mean of X.

Credits from Dr. Liu Yang

Variance of Functions

- Given a function f and random variable X, we define the variance Var(f(X)) as follows:

$$\mathrm{Var}(f(X)) = E[(f(X) - E(f(X)))^2]$$

Properties of Variance

For any random variable X and any a, b ∈ R, the following equations hold:

$$Var(aX+b) = a^2 Var(X)$$

$$Var(X) = E[X^2] - E[X]^2$$

Properties of Variance

For any random variable X and any a, b ∈ R, the following equations hold:

$$Var(X) = E[X^2] – E[X]^2$$

Proof: From the definition of variance, we have

$Var(X) = E[(X – E[X])^2] = E[X^2 – 2XE[X] + E[X]^2]$

$= E[X^2] – E[2XE[X]] + E[E[X]^2]$.

Note that E[X] and $E[X]^2$ are constants, so that

$Var(X) = E[X^2] – 2E[X]^2 + E[X]^2 = E[X^2] – E[X]^2$.

Credits from Dr. Liu Yang

Properties of Variance

For any random variable X and any a, b ∈ R, the following equations hold:

$$Var(aX+b) = a^2 Var(X)$$

Proof: From $Var(X) = E[X^2] - E[X]^2$, we have

$$Var(aX+b) = E[(aX+b)^2] - E[aX+b]^2$$

$$= E[a^2X^2 + 2abX + b^2] - (aE[X]+b)^2$$

$$= (a^2E[X^2] + 2abE[X] + b^2) - (a^2E[X]^2 + 2abE[X] + b^2)$$

$$= a^2E[X^2] - a^2E[X]^2 = a^2 (E[X^2] - E[X]^2) = a^2Var(X)$$

Credits from Dr. Liu Yang

Properties of Variance

If two random variables X and Y are independent, then

$$Var(X + Y) = Var(X) + Var(Y)$$

Proof: Using the linearity of the expectation and the identity $E(XY) = E(X)E(Y)$, which holds by the independence of X and Y, we can write

$$Var(X+Y) = E[(X+Y)^2] - E[X+Y]^2$$

$$= E[X^2+2XY+Y^2] - (E[X]^2+2E[X]E[Y]+E[Y]^2)$$

$$= (E[X^2] - E[X]^2) + (E[2XY] - 2E[X]E[Y]) + (E[Y^2] - E[Y]^2)$$

$$= \quad Var(X) \quad + \quad 0 \quad + \quad Var(Y)$$

Credits from Dr. Liu Yang

Exercises

X has values 0, 1, 2 with P(X=0)=0.3; P(X=1)=0.3; P(X=2)=0.4

Y has values 0, 1, 3 with P(Y=0)=0.3; P(Y=1)=0.2; P(Y=3)=0.5

Please answer:

1. Var(X) and Var(2X+1000000)

2. Var(Y) and Var(3Y+10000000000)

3. If X and Y are independent, what is Var(2X+3Y+10000000)?

Exercises

X has values 0, 1, 2 with P(X=0)=0.3; P(X=1)=0.3; P(X=2)=0.4
Y has values 0, 1, 3 with P(Y=0)=0.3; P(Y=1)=0.2; P(Y=3)=0.5

- Solution:

1. $E[X]=0*0.3+1*0.3+2*0.4=1.1$, $E[X^2]=0*0*0.3+1*1*0.3+2*2*0.3=1.5$

$Var(X) = E[X^2]-E[X]=1.5-1.1=0.4$, $Var(2X+1000000)=2*2*Var(X)=1.6$

2. $E[Y] = 0*0.3+1*0.2+3*0.5=1.7$, $E[Y^2]=0+1*1*0.2+3*3*0.5=4.7$

$Var(Y) = E[Y^2]-E[Y]= 3.0$, $Var(3Y+10000000000)=3*3*Var(Y)=27$

3.Because X and Y are independent: $Var(2X+3Y+10000000)=$
$Var(2X+3Y)=Var(2X)+Var(3Y)=4Var(X)+9Var(Y)=1.6+27=28.6$

Covariance

- The covariance gives some sense of how much two values are linearly related to each other, as well as the scale of these variables

The covariance of two random variables X and Y is denoted by Cov(X, Y) and defined as

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] .$$

Covariance of functions: given functions f and g, then

$$\mathrm{Cov}(f(X), g(Y)) = E[(f(X) - E[f(X)])(g(Y) - E[g(Y)])] .$$

Properties of Covariance

It is clear that

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X);$$

$$\mathrm{Cov}(f(X),f(X)) = \mathrm{Var}(f(X)).$$

Properties of Covariance

If two random variables X and Y are independent, then

$$\mathrm{Cov}(X, Y) = 0 \,.$$

Proof: From the definition of covariance, we have

$\mathrm{Cov}(X, Y) = E[(X – E[X])(Y – E[Y])]$

$= E[XY – E[X]Y – XE[Y] + E[X]E[Y]]$

$= E[XY] – E[E[X]Y] – E[XE[Y]] + E[E[X]E[Y]]$

$= E[XY] – E[X]E[Y] – E[X]E[Y] + E[X]E[Y]= E[XY] – E[X]E[Y] = 0.$

Credits from Dr. Liu Yang

Covariance: Exercise

Please compute:

Cov(X,Y)

X	**Y**	**P(X,Y)**
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

Covariance: Exercise

Please compute:

Cov(X,Y)

X	Y	P(X,Y)
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

Solution:
Because X and Y are independent, so

Cov(X,Y)=0

Covariance matrix

If we have multiple random variables:

$$\mathbf{X} = (X1, X2, \ldots, Xn),$$

then we can define the convariance matrix

Cov(**X**) is a n*x*n matrix and the ij-th element of the matrix is

$$Cov(Xi, Xj)$$

The covariance between two variables is positive when they tend to move in the same direction and negative if they tend to move in opposite directions.

Covariance matrix

Let **X** = (X,Y)

Please compute: Cov(**X**)

X	**Y**	**P(X,Y)**
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

Covariance matrix

Let **X** = (X,Y)

Please compute: Cov(**X**)

X	Y	P(X,Y)
0	0	0.25
0	2	0.25
2	0	0.25
2	2	0.25

Because X and Y are independent ,

Cov(X,Y) = 0.

Cov(X,X) = Var(X) = E[X^2]-E[X]= 2*2*0.5 - 2*0.5 =1

Cov(Y, Y) = Var(Y) = E[Y^2]-E[Y]= 2*2*0.5 - 2*0.5 =1

So

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Common Probability Distribution

Distribution	PDF or PMF	Mean	Variance
$\mathit{Bernoulli}(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$\mathit{Binomial}(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$	np	$np(1-p)$
$\mathit{Geometric}(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\mathit{Poisson}(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, \ldots$	λ	λ
$\mathit{Uniform}(a, b)$	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathit{Gaussian}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$\mathit{Exponential}(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Credicts from Dr. Griffin Young

Bernoulli Distribution

Bernoulli(p)

P(X=1) = p
P(X=0) = 1-p

Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes–no question.

For example, tossing a coin (only one time). If the possibility to appear the head is p, then fliping a coin is a Bernoulli Distribution with parameter p.

It is easy to check that
$E[X] = p$, $Var(X)=p(1-p)$

Credits from Wikipedia

Binomial Distribution

Binomial(p,n)

It is a generalization of the Bernoulli distribution to a distribution over integers. It can be used to describe the probability of observing m occurrences of X=1 in a set of n samples from a Bernoulli distribution where P(X=1)=p. The Binomial distribution is defined as

$$P(Y = m) = P(m; n, p) = \binom{n}{m} p^m (1 - p)^{n-m}$$

The expectation and variance of the binomial distribution are

E[Y] = np and Var(Y) = np(1-p)

Credits from Dr. Liu Yang

Gaussian Distribution (Normal Distribution)

- **Gaussian Distribution (Normal Distribution)**: It is the most widely used model for the distribution of continuous variables. For a single variable X, the Gaussian distribution can be represented as follows:

$$\mathrm{N}\left(\mu, \sigma^2\right).$$

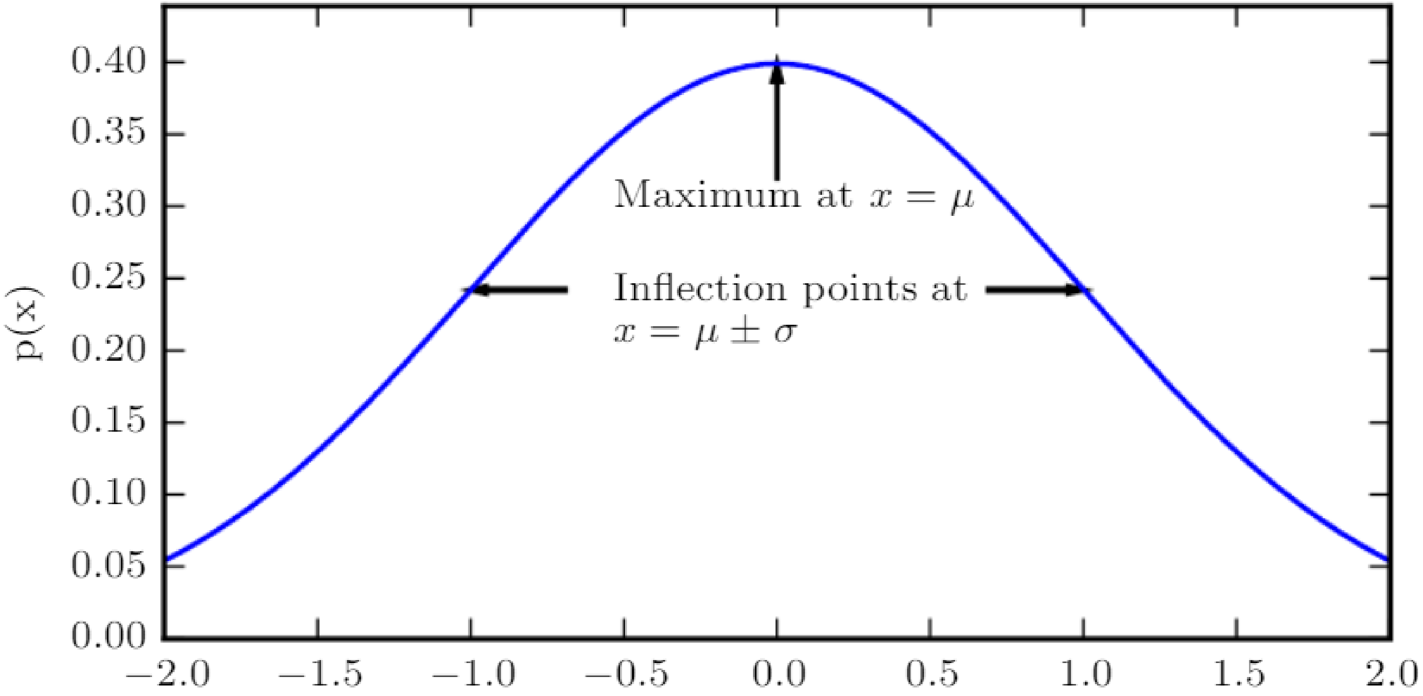
- It is a continuous distribution with the probability density function

$$\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(\mathrm{x}-\mu)^2\right)$$

Gaussian Distribution (Normal Distribution)

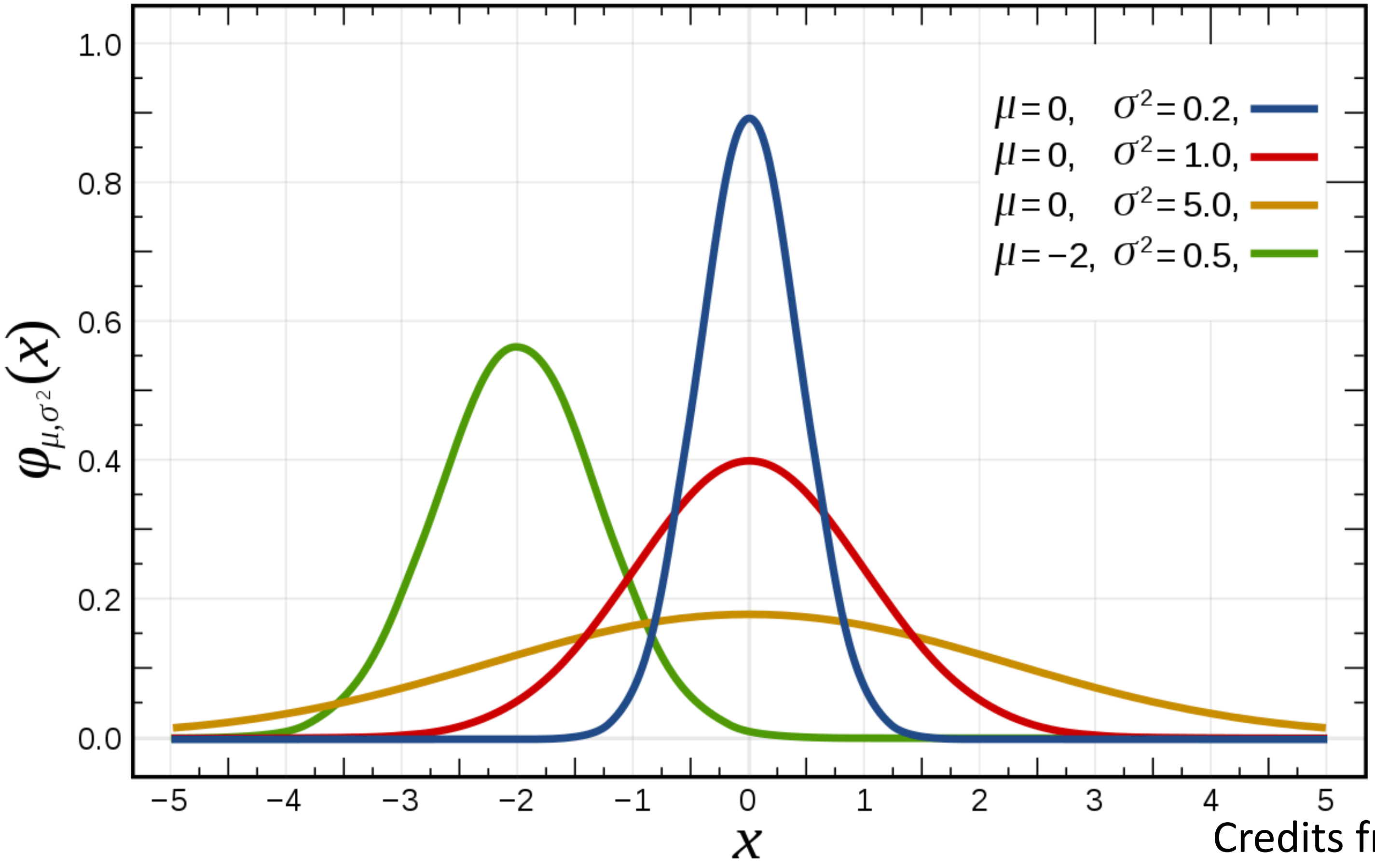
Gaussian distribution depends on two parameters μ, σ.

Then the expectation is μ and the variance is σ^2.



Gaussian Distribution (Normal Distribution)

- Gaussian distributions with different expectations and variances



Credits from Wikipedia

Exponential Distribution

- In the context of deep learning, we often want to have a probability distribution with a sharp point at $\mathrm{x} = 0$. To accomplish this, we can use the exponential distribution: the probability density function is

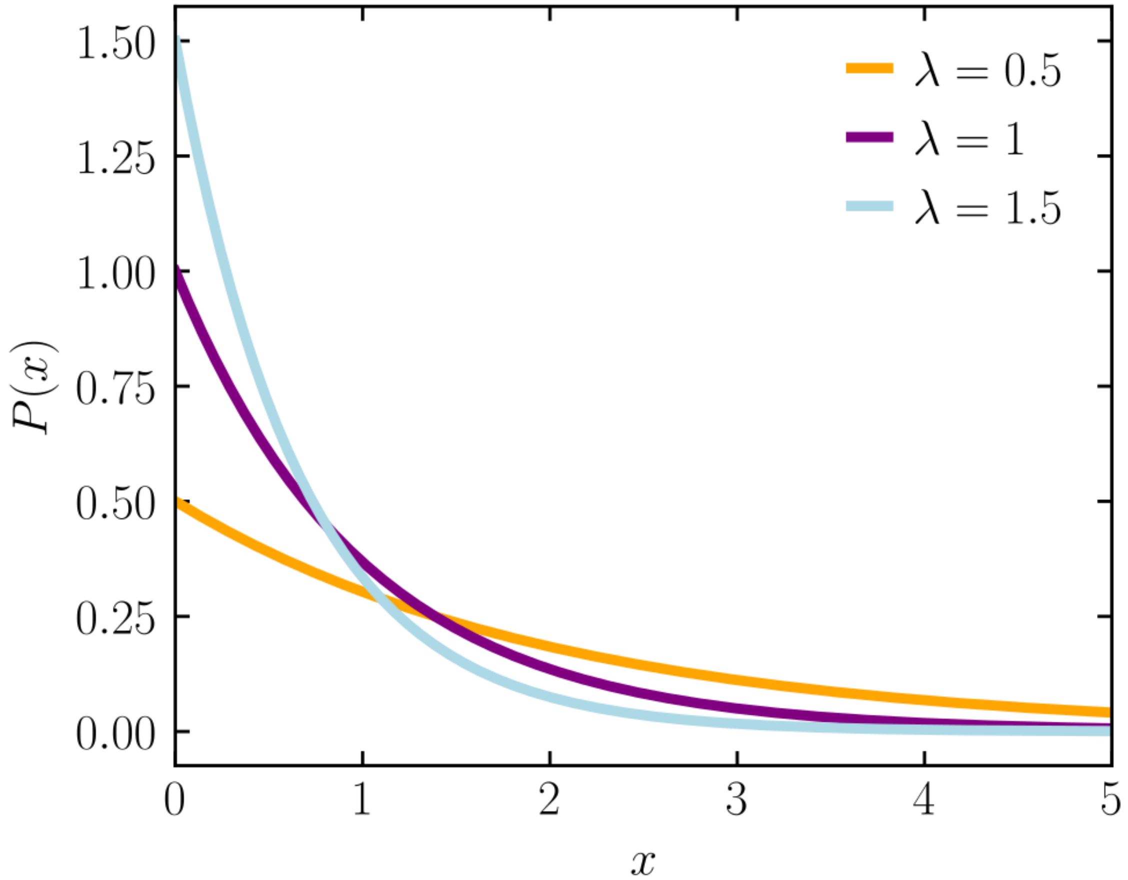
$$p(\mathrm{x}; \lambda) = \lambda \mathbf{1}_{\mathrm{x} \geq 0} \exp(-\lambda \mathrm{x})$$

- The exponential distribution is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate.

Credits from Wikipedia

Exponential Distribution

- Exponential distribution with different parameter λ.



Then the expectation is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$.

Credits from Wikipedia

Laplace Distribution

- A closely related probability distribution that allows us to place a sharp peak of probability mass at an arbitrary point μ is the Laplace distribution: the probability density function is

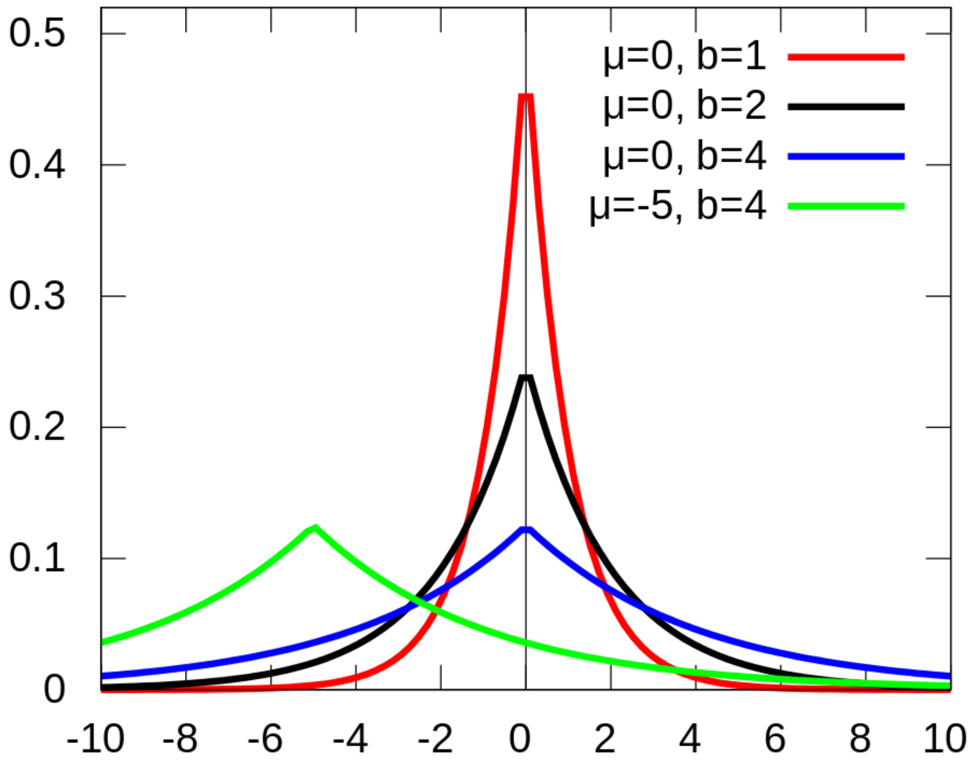
$$\text{Laplace}(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

- Laplace distribution represents the distribution of differences between two independent variables having identical exponential distributions. It is also called double exponential distribution.

Credits from Wikipedia

Laplace Distribution

- Laplace distribution with different parameters b and μ .



Then the expectation is μ and the variance is $2\mathrm{b}^2$.

Credits from Wikipedia

Mixtures of Distributions

How to construct distributions by exisiting distributions?

Mixture of distributions is an important method.

- If we have n discrete distributions P1(x), P2(x),...,Pn(x), then we can construct a mixture distribution by weights w1,...,wn

$$P(x) = \sum_{i=1}^{n} wi * Pi(x) .$$

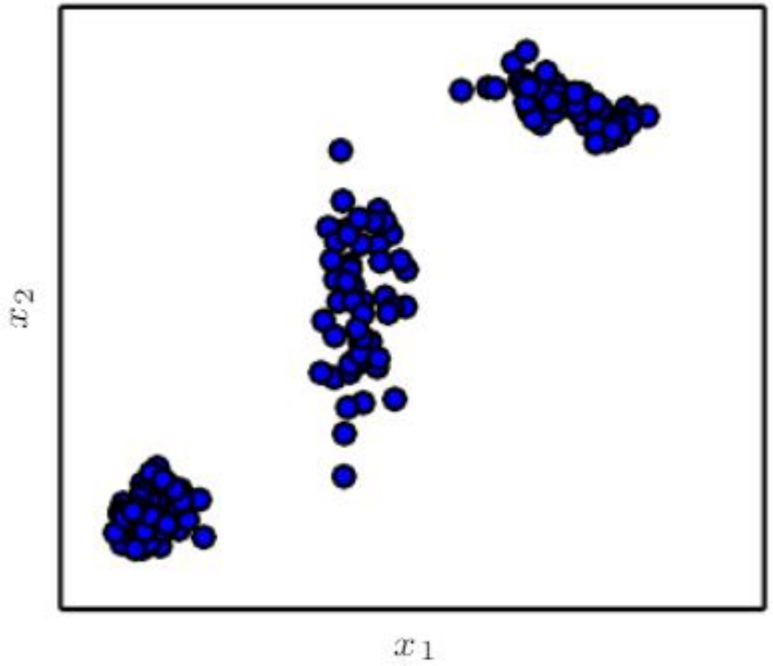
To ensure that P(x) is a distribution, wi should be non-negative and w1+w2+...+wn=1.

Mixtures of Distributions

If we have n continuous distributions with density function p1(x), p2(x),...,pn(x), then we can construct a mixture distribution by weights w1,...,wn: the mixture distribution's density function is

$$p(x) = \sum_{i=1}^{n} wi * pi(x) .$$

- Mixture distributions with three different Gaussian distributions



Change of Variables

How to construct distributions by exisiting distributions?

Contructing a function to deform the random variables is an important method.

- If X is a discrete random variable and g is a function, then

$$P(g(X) = y) = P(X \in g^{-1}(y))$$

Change of Variables

If X is a continuous random variable and g is a differentiable function, then

$$p_{\mathrm{X}}(\mathrm{x}) = p_{\mathrm{Y}}(\mathrm{g}(\mathrm{x}))\left|\frac{\partial \mathrm{g}(\mathrm{x})}{\partial \mathrm{x}}\right|$$

where $p_{\mathrm{X}}(\mathrm{x})$ is the probability density function for X and $p_{\mathrm{Y}}(\mathrm{g}(\mathrm{x}))$ is the probability density function for g(X).

Exercise

For a gaussian distribution with probability density function

$$\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}\mathrm{x}^2)$$

- If the function g(x) = 2x+1, then what is the density function after transforming by function g?

Exercise

Solution: $\frac{\partial g(x)}{\partial x}$ =2. By $p_X(x) = p_Y(g(x)) \left|\frac{\partial g(x)}{\partial x}\right|$, we know that

$$\sqrt{\frac{1}{2\pi}} \exp(-\frac{1}{2}x^2) = p_Y(2x+1) * 2$$

Let y=g(x)=2x+1.

Then x=0.5y-0.5, which implies that $\sqrt{\frac{1}{2\pi}} \exp(-\frac{1}{2}x^2) = p_Y(y) * 2$

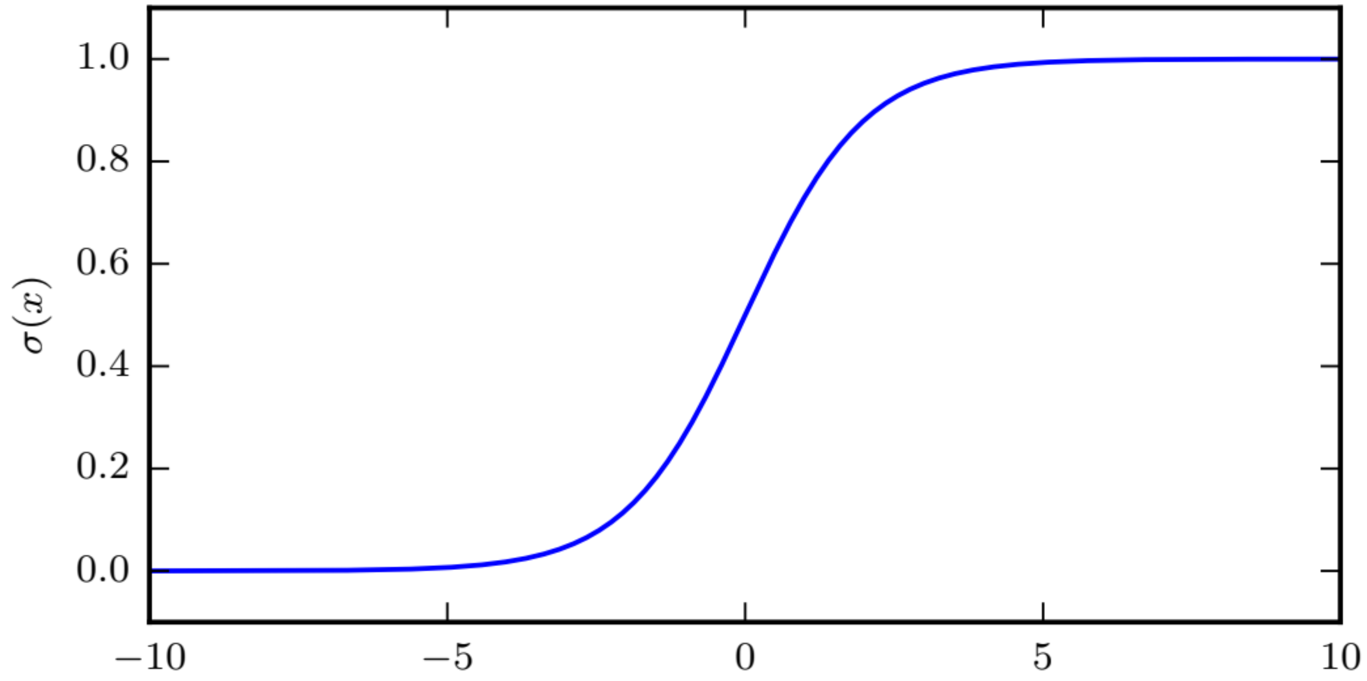
Replacing x by 0.5y-0.5, we have

$$p_Y(y) = \sqrt{\frac{1}{8\pi}} \exp(-\frac{1}{8}(y-1)^2)$$

Logistic Sigmoid

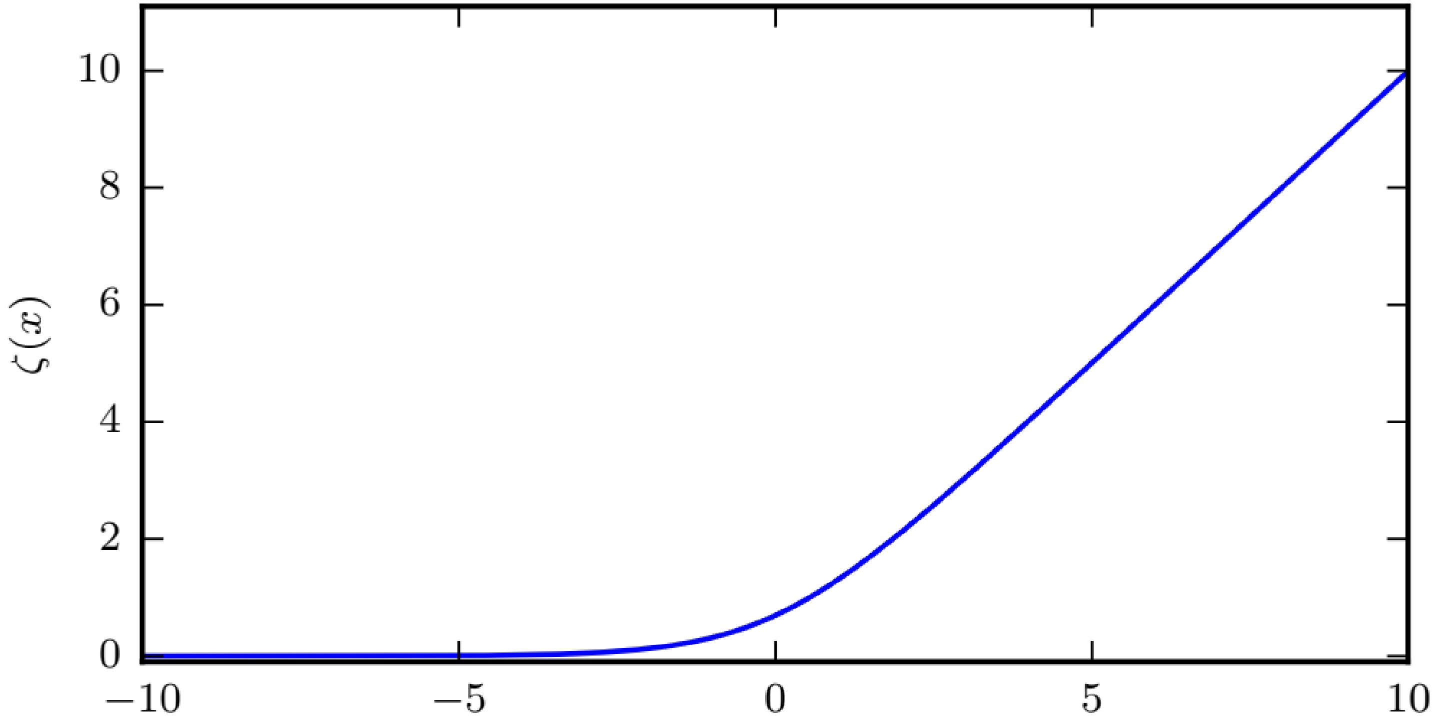
In deep learning, we construct many different functions to deform distributions.

- $\sigma(\mathrm{x}) = \frac{1}{1+\exp(-\mathrm{x})}$



Softplus Function

- $\sigma(\mathrm{x}) = \log(1 + \exp(\mathrm{x}))$



Thank You!