Analytics Vidhya
Learn everything about analytics

Ace
Data Science
Interviews

Comprehensive

Handbook on Data Science INTERVIEW QUESTIONS

Table of Contents

Puzzles

Question 1

Two trains X and Y (80 km from each other) are running towards each other on the same track with a speed of 40km/hr. A bird starts from the train X and travels towards train Y with constant speed of 100km/hr.

Once it reaches train Y, it turns and starts moving toward train X. It does this till the two trains collide with each other. Find the total distance traveled by the bird?

Answer 1

To simplify the problem, let's consider that only one train is moving (with the added velocity of both) and the other is stationary. So the new velocity for our train will be - (40 + 40) km/hr = 80km/hr

Now we need to find the total time before collision. As we know, the time = distance/speed. So total time the trains will take to collide = 80km/80km/hr = 1 hour.

Total distance travelled by the bird in this duration = 100km/hr * 1hr = 100 km.

Question 2

There are 8 batteries, but only 4 of them work. You have to use them for a flashlight which needs only 2 working batteries. To guarantee that the flashlight is turned on, what is the minimum number of battery pairs you need to test?

Answer 2

The first step involves naming the batteries, for instance, A, B, C, D, E, F, G, and H. In this problem, we can't compare 2 items directly. If a combination of two batteries fail to turn the light on, it means either one or both the batteries aren't working.

The puzzle must be approached in a circular manner.
The batteries are put test consecutively in the order AB, BC, and AC. At most, one of the three batteries between A, B, and C is working, only if none of the pairs work. This also implies that at least three batteries between D, E, F, G, and H must be functional.
DE combination is tried next. If they don't work, at least 2 out of F, G, and H must work. Similarly, try the combinations FG, GH, and FH to positively asset which batteries really work.

Question 3

You pull out 2 balls, one after another, from a bag which has 20 blue and 13 red balls in total. If the balls are of similar colour, then the balls are replaced with a blue ball, however, if the balls are of different colours, then a red ball is used to replace them.

Once the balls are taken out of the bag, they are not placed back in the bag, and thus the number of balls keep reducing. Determine the colour of last ball left in the bag.

Answer 3

If we pull out 2 red balls, we need to replace them with a blue ball. On the other hand, if we draw one red and one blue, it is replaced with a red one. This implies that the red ball would always be in odd numbers, whether we remove 2 together, or remove 1 while adding 1.
This also indicates that the last ball to stay in the bag would be a red one.

Question 4

10 coins are placed before you on a table, while you stay blindfolded. The candidate is permitted to touch the coins, however conditions to the puzzle dictates that he can't really determine which way up they are by feel.

5 coins are placed heads up, while the other 5 are kept tails up, without the interviewee knowing which ones are which. If you're allowed to flip the coins any number of times, how would you build two piles of coins each with the same number of heads up.

Answer 4

This problem can be solved by initially creating two piles of coin, with 5 randomly selected coins in each pile.

Let's assume the first pile looks like H, H, H, H, T and the other pile can be imagined as T, T, T, T, H. The final bit in solving the puzzle involves feeling out the certain type of coin and its number in a pile, now we just have to match the number same type of coin in other pile.

Question 5

There are 10 stacks of 10 coins each, where each coin weighs 10gms. However, one of the stacks is defective, and that stack contains coins which weigh 9gms. Determine the minimum number of weights needed to identify the defective stack.

Answer 5

Trick lies in creating a weighted stack for measurement, which will enable the candidate to identify the defective stack in one measurement.

A coin is taken from the first stack, 2 from the second, 3 from the third, and so on. This will give a total of 55 coins in hand. If none of them are defective, they would weigh 550 gms together.

However, if stack 1 turns defective, the total weight would stand at 549 gms; defect in stack 2 would result in a total weight of 548gms; and so on. Therefore, just one measurement can help the candidate identify the faulty stack.

Question 6

There are 3 switches in a room, where one of them is assigned for a bulb in the next room. You can't see whether the bulb is on or off, until you leave the room. Find the minimum number of times you have to go into the room to identify which switch corresponds to the bulb in the other room.

Answer 6

The person has to initially turn on the first switch and keep it on for 2-3 minutes. Next, turn off the first switch and turn on the second one. Rush to the other room as soon as you turn on the second switch.

If the bulb is glowing, the second switch corresponds to the light bulb; however, if the bulb doesn't glow, but touching it feels warm, the first switch is the one that turns the bulb on. If it's neither lit, nor warm, then the third switch is the desired switch. So, a person must go only once to the other room to find out the accurate switch.

Question 7

There are 3 mislabeled jars, with apple and oranges in the first and second jar respectively. The third jar contains a mixture of apples and oranges. You can pick as many fruits as required to precisely label each jar. Determine the minimum number of fruits to be picked up in the process of labeling the jars.

Answer 7

A noticeable aspect in this puzzles is the fact that there's a circular misplacement, which implies if apple is wrongly labelled as Apple, Apple can't be labelled as Orange, i.e., it has to be labeled as A+O. We are acquainted with the fact that everything is wrongly placed, which means A+O jar contains either Apple or Orange (but not both).

The candidate picks one fruit from A+O, and let's assume he gets an apple. He labels the jar as apple, however, jar labelled Apple can't have A+O. Thus, the third jar left in the process should be labelled A+O.

Basically, picking only one fruit helps in correctly labeling the jars.

Question 8

There are 5 pirates in a ship. Pirates have hierarchy C1, C2, C3, C4 and C5. C1 designation is the highest and C5 is the lowest. These pirates have three characteristics: a. Every pirate is so greedy that he can even take lives to make more money. b. Every pirate desperately wants to stay alive. c. They are all very intelligent.

There are total 100 gold coins on the ship. The person with the highest designation on the deck is expected to make the distribution. If the majority on the deck does not agree to the distribution proposed, the highest designation pirate will be thrown out of the ship (or simply killed). Only the person with the highest designation can be killed at any moment. What is the right distribution of the coins proposed by the captain so that he is not killed and does make maximum amount?

Answer 8

The solution of this problem lies in thinking through what will happen if all the pirates were thrown one by one and then thinking in reverse order.

Let us name pirates as A,B,C,D and E in hierarchy (A being highest).

If only D and E are left at end, D will simply give 0 coins to E and still escape because majority cannot be reached. Hence, even if E gets 1 coin he will give his vote to the distributor.

If C, D and E are there on the deck, C will simply give one coin to E to get his vote. And D simply gets nothing. Hence, even if D gets 1 coin he will give his vote to the distributor.

If B,C,D and E are there on the deck, B will simply give one coin to D to get his vote. C & E simply gets nothing.

If A,B,C,D and E are there on the deck, A simply gives 1 coin each to C and E to get their votes.

Hence, in the final solution A gets 98 coins and only C & E get 1 coin each.

Question 9

There are 100 prisoners all sentenced to death. One night before the execution, the warden gives them a chance to live if they all work on a strategy together. The execution scenario is as follows –

On the day of execution, all the prisoners will be made to stand in a straight line such that one prisoner stands just behind another and so on. All prisoners will be wearing a hat either of Blue colour or Red. The prisoners don't know what colour of hat they are wearing. The prisoner who is standing at the last can see all the prisoners in front of him (and what colour of hat they are wearing). A prisoner can see all the hats in front of him. The prisoner who is standing in the front of the line cannot see anything.

The executioner will ask each prisoner what colour of hat they are wearing one by one, starting from the last in the line. The prisoner can only speak "Red" or "Blue". He cannot say anything else. If he gets it right, he lives otherwise he is shot instantly. All the prisoners standing in front of him can hear the answers and gunshots.

Assuming that the prisoners are intelligent and would stick to the plan, what strategy would the prisoners make over the night to minimize the number of deaths?

Answer 9

The strategy is that the last person will say 'red' if the number of red hats in front of him are odd and 'blue' if the number of red hats in front of him are even. Now, the 99th guy will see the if the red hats in front of him are odd or even. If it is odd then obviously the hat above him is blue, else it is red. From now on, it's pretty intuitive.

Question 10

There is a bus with 100 labeled seats (labeled from 1 to 100). There are 100 persons standing in a queue. Persons are also labelled from 1 to 100.

People board on the bus in sequence from 1 to n. The rule is, if person 'i' boards the bus, he checks if seat 'i' is empty. If it is empty, he sits there, else he randomly picks an empty seat and sit there. Given that 1st person picks seat randomly, find the probability that 100th person sits on his place i.e. 100th seat.

Answer 10

The final answer is the probability that the last person ends in up in his proper seat is exactly 1/2

The reasoning goes as follows:

First, observe that the fate of the last person is determined the moment either the first or the last seat is selected! This is because the last person will either get the first seat or the last seat. Any other seat will necessarily be taken by the time the last guy gets to 'choose'.

Since at each choice step, the first or last is equally probable to be taken, the last person will get either the first or last with equal probability: 1/2.

Question 11

N persons are standing in a circle. They are labelled from 1 to N in clockwise order. Every one of them is holding a gun and can shoot a person on his left. Starting from person 1, they starts shooting in order e.g for N=100, person 1 shoots person 2, then person 3 shoots person 4, then person 5 shoots person 6……..then person 99 shoots person 100, then person 1 shoots person 3, then person 5 shoots person 7……and it continues till all are dead except one. What's the index of that last person?

Answer 11

Write 100 in binary, which is 1100100 and take the complement which is 11011 and it is 27. Subtract the complement from the original number. So 100 – 27 = 73.

Try it out for 50 people. 50 = 110010 in binary.

Complement is 1101 = 13. Therefore, 50 – 13 = 37.

For the number in form 2^n, it will be the first person. Let's take an example:

64 = 1000000

Complement = 111111 = 63.

64-63 = 1.

You can apply this for any 'n'.

Question 12

Four glasses are placed on the corners of a square Lazy Susan (a square plate which can rotate about it's center). Some of the glasses are upright (up) and some upside-down (down).

A blindfolded person is seated next to the Lazy Susan and is required to re-arrange the glasses so that they are all up or all down, either arrangement being acceptable (which will be signalled by say ringing of a bell).

The glasses may be rearranged in turns with subject to the following rules: Any two glasses may be inspected in one turn and after feeling their orientation the person may reverse the orientation of either, neither or both glasses. After each turn the Lazy Susan is rotated through a random angle.

The puzzle is to devise an algorithm which allows the blindfolded person to ensure that all glasses have the same orientation (either up or down) in a finite number of turns. (The algorithm must be deterministic, i.e. non-probabilistic)

Answer 12

This algorithm guarantees that the bell will ring in at most five turns:

1. On the first turn, choose a diagonally opposite pair of glasses and turn both glasses up.
2. On the second turn, choose two adjacent glasses at least one will be up as a result of the previous step. If the other is down, turn it up as well. If the bell does not ring, then there are now three glasses up and one down.
3. On the third turn, choose a diagonally opposite pair of glasses. If one is down, turn it up and the bell will ring. If both are up, turn one down. There are now two glasses down, and they must be adjacent.
4. On the fourth turn, choose two adjacent glasses and reverse both. If both were in the same orientation then the bell will ring. Otherwise there are now two glasses down and they must be diagonally opposite.
5. On the fifth turn, choose a diagonally opposite pair of glasses and reverse both. The bell will ring.

Question 12

There are 10 incredibly smart boys at school: A, B, C, D, E, F, G, H, I and Sam. They run into class laughing at 8:58 am, just two minutes before the playtime ends and are stopped by a stern looking teacher: Mr Rabbit.

Mr Rabbit sees that A, B, C and D have mud on their faces. He, being a teacher who thinks that his viewpoint is always correct and acts only to enforce rules rather than thinking about the world that should be, lashes out at the poor kids.

“Silence!”, he shouts. “Nobody will talk. All of you who have mud on your faces, get out of the class!”. The kids look at each other. Each kid could see whether the other kids had mud on their faces, but could not see his own face. Nobody goes out of the class.

“I said, all of you who have mud on your faces, get out of the class!”

Still nobody leaves. After trying 5 more times, the bell rings at 9 and Mr Rabbit exasperatedly yells: “I can clearly see that at least one of you kids has mud on his face!”.

The kids grin, knowing that their ordeal will be over soon. Sure enough, after a few more times bawling of “All of you who have mud on your faces, get out of the class!”, A, B, C and D walk out of the class.

Explain how A, B, C and D knew that they had mud on their faces. What made the kids grin? Everybody knew that there was at least one kid with mud on his face. Support with a logical statement that a kid did not know before Mr Rabbit’s exasperated yell at 9, but that the kid knew right after it.

Answer 13

After Mr Rabbit’s first shout, they understood that at least one boy has mud on his face. So, if it was exactly one boy, then the boy would know that he had mud on his face and go out after one shouting.

Since nobody went out after one shouting, they understood that at least two boys have mud on their faces. If it were exactly two boys, those boys would know (they would see only one other’s muddy face and they’d understand their face is muddy too) and go out after the next shouting.

Since nobody went out after the second shouting, it means there are atleast three muddy faces And so on, after the fourth shouting, A, B, C and D would go out of the class.

This explanation does leave some questions open. Everybody knew at least three others had mud on their faces, why did they have to wait for Mr. Rabbit’s shout at the first place? Why did they have to go through the all four shoutings after that as well?

In multi-agent reasoning, an important concept arises of common knowledge. Everybody knows that there are at least three muddy faces but they cannot act together on that information without knowing that everybody else knows that too. And that everybody knows that everybody knows that and so on. This is what we’ll be analyzing. It requires some imagination, so be prepared.

A knows that B, C and D have mud on their faces. A does not know if B knows that three people have mud on their faces. A knows that B knows that two people have mud on their faces. But A can't expect people to act on that information because A does not know if B knows that C knows that there are two people with mud on their faces. If you think this is all uselessly complicated, consider this:

A can imagine a world in which he does not have mud on his face. (Call this world A) In A's world, A can imagine B having a world where both A and B do not have mud on their faces. (Call this world AB)

A can imagine a world where B imagines that C imagines that D imagines that nobody has mud on their faces. (Call this world ABCD). So when Mr Rabbit shouted initially, it could have been that nobody was going out because a world ABCD was possible in which nobody should be going out anyway.

So here's a statement that changes after Mr. Rabbit's yell. World ABCD is not possible i.e. A cannot imagine a world where B imagines that C imagines that D imagines that nobody has mud on their faces. So now in world ABC, D knows he has mud on his face. And in world ABD, C knows he has mud on his face and so on.

Question 14

There are 7 prisoners sitting in a circle. The warden has caps of 7 different colours (an infinite supply of each colour). The warden places a cap on each prisoner's head – he can chose to place any cap on any other's head. Each prisoner can see all caps but her/his own. The warden orders everybody to shout out the colour of their respective caps simultaneously. If any one is able to guess her/his colour correctly, he sets them free. Otherwise, he send them in a dungeon to rot and die. Is it possible to devise a scheme to guarantee that nobody dies?

Answer 14

Assign to each of the 7 colours a unique number from 0-6. Henceforth, we will only be doing modular arithmetic(modulo 7).

Assign to each of the 7 prisoners a unique number from 0-6. If the number assigned to prisoner P is N, then P always guesses that the sum of the colours assigned to all

prisoners is M (modulo 7). Thus, he calculates his own colour under this assumption (= (M - sum(colours of the 6 prisoners he can see))%7).

There will always be a prisoner who guesses the correct sum (as the sum lies in 0-6), and this prisoner therefore correctly guesses his own colour.

If there is a solution, then exactly one prisoner is correct (no more). This is because there are 7^7 scenarios.

Each prisoner's response is a function of the colours of the other 6, so if you fix their colours and vary his colour, you can see that he will be correct in exactly one-seventh of the cases (=7^6). The sum (across all scenarios) of the number of prisoners who are correct is 7*(7^6)=7^7.

If each scenario is to have at least one person right, this implies that each scenario cannot have more than one person who is right.

Being right about one's colour is equivalent to being right about the sum of colours of all prisoners (modulo 7). (The colours of the other 6 are known.) So guessing one's colour is the same as guessing the sum. How do we make sure that at least one person guesses the correct sum? By making sure that everybody guesses a different sum.

Question 15

One day, an alien comes to Earth and does one of four things, each with equal probability to:

(i) Kill himself

(ii) Do nothing

(iii) Split himself into two aliens (while killing himself)

(iv) split himself into three aliens (while killing himself)

Once the alien splits, he will take either of the above actions everyday.What is the probability that the alien species eventually dies out entirely?

Answer 15

The answer is √2 − 1.

Suppose that the probability of aliens eventually dying out is x and there are n number of aliens on some certain day.

Then for n aliens, the probability of eventually dying out is xn because we consider every alien as a separate colony. Now, if we compare aliens before and after the first day, we get:

$x = (1/4) * 1 + (1/4) * x + (1/4) * x^2 + (1/4) * x^3$

$x^3 + x^2 - 3x + 1 = 0$

$(x - 1)(x\ 2 + 2x - 1) = 0$

We get, $x = 1, -1 - \sqrt{2}$, or $-1 + \sqrt{2}$

We claim that x cannot be 1, which would mean that all aliens eventually die out. The number of aliens in the colony is, on average, multiplied by 0+1+2+3 4 = 1.5 every visit, which means in general the aliens do not die out. (A more rigorous line of reasoning is included below.) Because x is not negative, the only valid solution is $x = \sqrt{2} - 1$.

To show that x cannot be 1, we show that it is at most √ 2−1.

Let xn be the probability that a colony of one alien will die out after at most n minutes. Then, we get the relation:

$x_n + 1 = 1/4\ (1 + x_n + x^2 n + x^2 n)$

We claim that $x_n \leq \sqrt{2} - 1$ for all n, which we will prove using induction.

It is clear that $x1 = 1/4 \leq \sqrt{2} - 1$. Now, assume $x_k \leq \sqrt{2} - 1$ for some k. We have:

$x_{k+1} \le 1/4\ (1 + x_k + x^2_k + x^3_k)$

$\le 1/4\ (1 + (\sqrt{2} - 1) + (\sqrt{2} - 1)^2 + (\sqrt{2} - 1)^3)$

$= \sqrt{2} - 1$

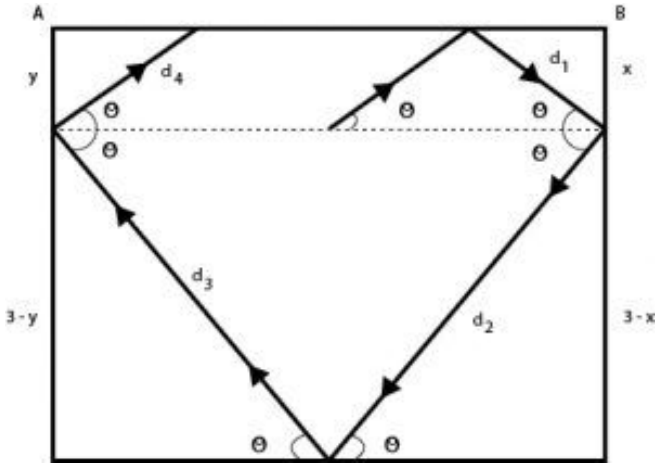
which completes the proof that $x_n \le \sqrt{2} - 1$ for all n. Now, we note that as n becomes large, x_n approaches x. Using formal notation, this is:

$x = \lim_{n \to \infty} x_n \le \sqrt{2} - 1$, so x cannot be 1.

Question 16

A photon starts moving in random direction from the center of square of size 3. Let's say it first colloids to the glass wall AB. What is the expected distance traveled by photon before hitting the wall AB again?

Answer 16



Above is a pictorial representation of the photon. We can calculate it's distance as shown below:

d1 = x cosec (Θ)

d2 = (3 - x) cosec (Θ)

d3 = (3 - y) cosec (Θ)

d4 = y cosec (Θ)

Total distance = d1+ d2 + d3 + d4

= 6 cosec (Θ)

We know, Θ varies between Π/4 and 3Π/4

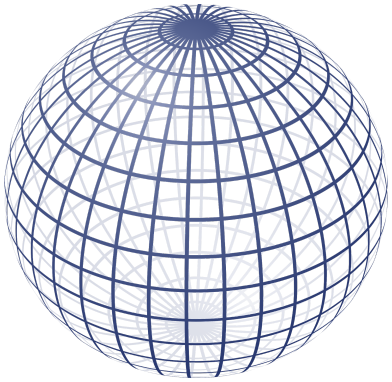
Therefore, E(distance) = 6 E (cosec Θ)

= 6 x (2 /Π) ∫cosec(Θ)dΘ (limits Π/4 to 3Π/4)

= 12/Π ln (√2 + 1/√2 + 1)

Question 17

Consider a unit sphere. 4 points are randomly chosen on it, what is the probability that the centre (of sphere) lies within the tetrahedron (polygon) formed by those 4 points?



Answer 17

Let A, B and C be random points on the sphere with Aa, Bb and Cc being diameters.

The spherical (minor) triangle abc is common to the hemispheres abc, bca and cab (where the notation abc represents the hemisphere cut off by the great circle through a and b and containing the point c, etc), therefore the probability that a further random point, D, lies on this triangle is:

1/2 x 1/2 x 1/2 = 1/8

(For centre to lie in the tetrahedron D should lie in the triangle i.e the opposite hemisphere of ABC)

Question 18

In a country in which people only want boys, every family continues to have children until they have a boy. If they have a girl, they have another child. If they have a boy, they stop. What is the proportion of boys to girls in the country?

Answer 18

Following is the required calculation:

Expected Number of boys for 1 family = 1*(Probability of 1 boy) + 1*(Probability of 1 girl and a boy) + 1*(Probability of 2 girls and a boy) + …

For C couples = 1*(C*1/2) + 1*(C*1/2*1/2) + 1*(C*1/2*1/2*1/2) + …

Expected Number of boys = C/2 + C/4 + C/8 + C/16 + …

Expected Number of boys = C

Expected Number of girls for 1 family = 0*(Probability of 0 girls) + 1*(Probability of 1 girl and a boy) + 2*(Probability of 2 girls and a boy) + …

For C couples = 0*(C*1/2) + 1*(C*1/2*1/2) + 2*(C*1/2*1/2*1/2) + …

Expected Number of girls = 0 + C/4 + 2*C/8 + 3*C/16 + …

Expected Number of girls = C

Therefore, the proportion is C/C = 1:1

Question 19

An bad king has a cellar of 1000 bottles of delightful and very expensive wine. A neighbour queen plots to kill the bad king and sends a servant to poison the wine.

Fortunately (or say unfortunately) the bad king's guards catch the servant after he could poison only one bottle. Alas, the guards don't know which

bottle, but know that the poison is so strong that even if diluted 100,000 times it would still kill the king.

Furthermore, it takes one month to have an effect. The bad king decides he will get some of the prisoners in his vast dungeons to drink the wine. Being a clever bad king, he knows that he needs to murder no more than 10 prisoners – believing he can fob off such a low death rate – and will still be able to drink the rest of the wine (999 bottles) at his wedding party in 5 weeks time.

Explain what is in mind of the king, how will he be able to do so ? (he has only 10 prisoners in his prisons)

Answer 19

The number the bottles are 1 to 1000. Now, write the number in binary format. We can write it as:

bottle 1 = 0000000001 (10 digit binary)

bottle 2 = 0000000010

.

.

.

bottle 500 = 0111110100

bottle 1000 = 1111101000

Now, take 10 prisoners and number them 1 to 10. Let prisoner 1 take a sip from every bottle that has a 1 in its least significant bit. And, this process will continue for every prisoner until the last prisoner is reached. For example:

Prisoner = 10 9 8 7 6 5 4 3 2 1

Bottle 924 = 1 1 1 0 0 1 1 1 0 0

For instance, bottle no. 924 would be sipped by 10,9,8,5,4 and 3. That way if bottle no. 924 was the poisoned one, only those prisoners would die.

After four weeks, line the prisoners up in their bit order and read each living prisoner as a 0 bit and each dead prisoner as a 1 bit. The number that you get is the bottle of wine that was poisoned. We know, 1000 is less than 1024 (2^10). Therefore, if there were 1024 or more bottles of wine it would take more than 10 prisoners.

Question 20

You and your friend are caught by gangsters and made to play a game to determine if you should live or die. The game is simple.

There is a deck of cards and you both have to choose a card. You can look at each other's cards but not at the card you have chosen. You both will survive if both are correct in guessing the card they have chosen. Otherwise both die.

What is the probability of you surviving if you and your friend play the game optimally?

Answer 20

We know, A and B have picked a card at random from a deck. A can see B's card and vice versa. So, A knows (s)he has not picked B's card, but apart from that, (s)he knows that the card is equally probable to be any of the other 51 cards. So, if A guesses B's card, they lose. But if A guesses any other card, there's a 1/51 chance that A is right. This also implies that total probability of success <= 1/51.

A's aim now is to tell any card apart from B's card that gives B the most information about B's own card. So they can plan beforehand as follows:

Consider the sequence of cards Clubs 1-13, Diamonds 1-13, Hearts 1-13, Spades 1-13. A will tell the card after B's card in this sequence. (If A says 4 of Hearts, it means B has 3 of Hearts. If A says Ace of Clubs, it means B has King of Spades)

With A's guess, which is always different from B's card, B gets to know exactly which card (s)he has and can always guess correctly. So the probability of success is 1/51, which is the maximum achievable.

Question 21

You have 12 balls that all weigh the same except one, which is either slightly lighter or slightly heavier. The only tool you have is a balance scale that can only tell you which side is heavier. Using only three weightings, how can you deduce, without a shadow of a doubt, which is the odd one out, and if it is heavier or lighter than the others?

Answer 21

First we weigh {1,2,3,4} on the left and {5,6,7,8} on the right. There are three scenarios which can arise from this:

If they balance, then we know 9, 10, 11 or 12 is fake. Weigh {8, 9} and {10, 11} (Note: 8 is surely not fake). If they balance, we know 12 is the fake one. Just weigh it with any other ball and figure out if it is lighter or heavier.

If {8, 9} is heavier, then either 9 is heavy or 10 is light or 11 is light. Weigh {10} and {11}. If they balance, 9 is fake (heavier). If they don't balance then whichever one is lighter is fake (lighter).

If {8, 9} is lighter, then either 9 is light or 10 is heavy or 11 is heavy. Weigh {10} and {11}. If they balance, 9 is fake (lighter). If they don't balance then whichever one is heavier is fake (heavier).

If {1,2,3,4} is heavier, we know either one of {1,2,3,4} heavier or one of {5,6,7,8} is lighter but it is guarantees that {9,10,11,12} are not fake. This is where it gets really tricky, watch carefully. Weigh {1,2,5} and {3,6,9} (Note: 9 is surely not fake).

If they balance, then either 4 is heavy or 7 is light or 8 is light. Following the last step from the previous case, we weigh {7} and {8}. If they balance, 4 is fake(heavier). If they don't balance then whichever one is lighter is fake (lighter).

If {1,2,5} is heavier, then either 1 is heavy or 2 is heavy or 6 is light. Weigh {1} and {2}. If they balance, 6 is fake (lighter). If they don't balance then whichever one is heavier is fake (heavier).

If {3,6,9} is heavier, then either 3 is heavy or 5 is light. Weigh {5} and {9}. They won't balance. If {5} is lighter, 5 is fake (lighter). If they balance, 3 is fake (heavier).

If {5,6,7,8} is heavier, it is the same situation as if {1,2,3,4} was heavier. Just perform the same steps using 5,6,7 and 8. Unless maybe you are too lazy to try and reprocess the steps, then you continue reading the solution. Weigh {5,6,1} and {7,2,9} (Note: 9 is surely not fake).

If they balance, then either 8 is heavy or 3 is light or 4 is light. Following the last step from the previous case, we weigh {3} and {4}. If they balance, 8 is fake(heavier). If they don't balance then whichever one is lighter is fake (lighter).

If {5,6,1} is heavier, then either 5 is heavy or 6 is heavy or 2 is light. Weigh {5} and {6}. If they balance, 2 is fake (lighter). If they don't balance then whichever one is heavier is fake (heavier).

If {7,2,9} is heavier, then either 7 is heavy or 1 is light. Weigh {1} and {9}. If they balance, 7 is fake (heavier). If they don't balance then 1 is fake (lighter).

Question 22

Robin and Williams are playing a game. An unbiased coin is tossed repeatedly. Robin wins as soon as the sequence of tosses HHT appears. Williams wins as soon as the sequence of tosses HTH appears. The game ends when one of them wins. What are the probabilities of winning for each player?

Answer 22

(Robin) HHT – 2/3 (Williams) HTH – 1/3

Let the probability of Robin winning be p. The probability of Williams winning is (1-p). If the first toss is tails, it is as good as the game has not started, hence the probability of Robin winning is p after the first tail.

p = (1/2)*p + ….

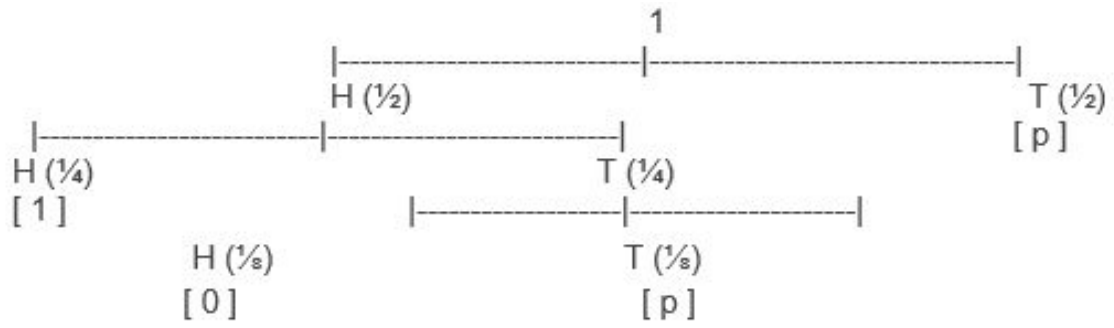
Let the first toss be heads. If the second toss is heads, then Robin definitely wins. Since HH has occurred, and at some point, tails will occur, so HHT will occur. Hence Robin wins with probability 1 for HH.

p = (1/2)*p + (1/2)*((1/2)*1 + ...)

Let the second toss be tails. If the third toss is heads, Robin loses as HTH occurs. If the third toss is tails (HTT) – since two tails have occurred in a row, now it is as good as the game has started from the beginning, so the chances of Robin winning are back to p.

T HH HTH HTT

p = (1/2)*p + 1/2 ((1/2)*1 + 1/2 ((1/2)*0 + (1/2) * p))



p = (1/2)*p + (1/4)*1 + (1/8)*0 + (1/8)*p

Finally, solving this equation gives us p = 2/3.

Question 23

On an island live 13 purple, 15 yellow and 17 maroon chameleons. When two chameleons of different colors meet, they both change into the third color. Is there a sequence of pairwise meetings after which all chameleons have the same color?

Answer 23

Let <p, y, m> denote a population of p purple, y yellow and maroon chameleons. Can population <13, 15, 17> be transformed into <45, 0, 0> or <0, 45, 0> or <0, 0, 45> through a series of pairwise meetings?

We can define function:

X(p, y, m) = (0p + 1y + 2m) mod 3

An interesting property of X is that its value does not change after any pairwise meeting because

X(p, y, m) = X(p-1, y-1, m+2) = X(p-1, y+2, m-1) = X(p+2, y-1, m-1)

Now X(13, 15, 17) equals 1. However,

X(45, 0, 0) = X(0, 45, 0) = X(0, 0, 45) = 0**

This means that there is no sequence of pairwise meetings after which all chameleons will have identical colour.

Guess Estimates

Question 1

Estimate the number of cigarettes consumed monthly in India.

Answer 1

The population of India, i.e., 1.2 billion. Following is an effective way to segment this population:

	Population : 1.2 Bn (100%)								
Segment level I	Age above 22 yrs (60%)				Age between 16 & 22 yrs(10%)				Age <16yrs (30%)
Segment level II	Urban (20%)		Rural (40%)		Urban (3%)		Rural (7%)		
Segment level III	Male (11%)	Female (9%)	Male (25%)	Female (15%)	Male (1.5%)	Female (1.5%)	Male (4%)	Female (3%)	
Avg. cigarettes PM	30	15	5	2	20	10	2	1	0
Population	132000000	108000000	300000000	180000000	18000000	18000000	48000000	36000000	360000000
# cigarettes PM	3960000000	1620000000	1500000000	360000000	360000000	180000000	96000000	36000000	0
Total cigarettes	**8.1 Trillion**								

Following were the key considerations in building the segmentation and the intermediate guesses:

The rural population consumes far lesser cigarettes than urban because of the purchasing power difference.

Male consume more cigarettes than female in both urban and rural populations.

Children below 16 years consume a negligible number of cigarettes.

Male to Female ratio in Urban is closer to 1 than that of Rural.

Male to Female ratio in younger generations is closer to 1 than that of older. This is because of the increase in awareness level.

Bulk of population start smoking after getting into a job and hence the average number cigarettes are higher in older groups.

Total number of cigarettes from the supply side also come to around 10 Trillion, which gives a good sense check on the final number.

Question 2

A startup based out of Delhi is considering selling adverts on WhatsApp. Guesstimate the number of messages sent in NCR everyday so they may estimate the total market volume.

Answer 2

The population of NCR is 30 million.

A message is counted as one message when the user presses the send button or hits 'enter' regardless of the length of the message.

The population of Delhi has been divided into 4 age groups. Further, these age groups have been divided into Heavy, Medium and Light users as per their messaging habits

15-25

- People who form the maximum percentage of the population and who use Whatsapp most heavily.
- School and college students and young professionals. Generally use Whatsapp for regular communication with their classmates/friends and have multiple groups on whatsapp for societies/classes/friends etc.
- Proportion of heavy users of Whatsapp will be considerably more than the Medium and Light users given the age group. Medium users will be comparable to the heavy users in other groups.

25-35:

- working population majorly who use Whatsapp heavily but have lesser time to devote to their app because of their work timings.
- Proportion of Heavy users will be more than that of Medium and Light users, but the difference between the three categories will be lesser.

35-50:

- This demographic generally has a reduced usage of Whatsapp because of family commitments and work pressure.
- Active on Whatsapp groups made of family, friends and colleagues. Moreover, the proportion of Medium users will be more than that of Heavy and Light users.

50-60

- This demographic uses Whatsapp the least, mostly to stay in touch with family members on family groups and some friends.
- Proportion of light users will be much more than Heavy and Medium users.

Table 3.1 — Description of each age group

		% Population share in number of messages		
Age Group	% Share in Total Whatsapp UserPopulation	High	Medium	Low
15-24	50	70	20	10
25-35	25	50	30	20
35-50	15	30	50	20
50-60	10	10	30	60

Table 3.2 — Column 2: Distribution of Age groups of people in total Whatsapp using population. Columns 3 to 5: Distribution of the respective age groups in high, medium and low frequency of messaging basis table 3.1.

4. The number of users of the app is arrived at by dividing the above population into three income groups: Rich, Middle and Poor. Out of these, the poor are assumed to not use Whatsapp. Further exclusions include the extreme age groups and 'people who don't use Whatsapp for other reasons (use other messaging apps, don't use messaging services etc).

The percentage penetration of Whatsapp in each income group is assumed to be 70% for the Rich due to commonly seen widespread usage of the application. In the middle income group, though the absolute numbers will be greater, a lesser percentage of people beyond the age of 35 would be involved.

	%Share in Total Population	People	With Whatsapp	Comments
Rich	20	6,000,000	4,200,000	70% of Total
Middle	50	15,000,000	9,750,000	65% of Total
Poor	30	9,000,000	0	0% of Total
Total (Middle + Rich)	70	21,000,000	**13,950,000**	

Table 4.1 — Distribution of people across Income Groups

5. The number of messages for heavy, medium and light users have been arrived at using educated guesses for each of the age groups keeping in mind their respective habits and purpose of use as specified in Table 3.1.

	Number of Messages		
Age Group	High	Medium	Low
0-15	-	-	-
15-24	700	400	200
25-35	500	200	100
35-50	200	100	50
50-60	100	50	25
60 and above	-	-	-

Table 5.1 — Number of messages as defined by messaging habits across age groups

6. We arrive at the respective total number of messages sent by multiplying the number of people in a particular age group to the number of messages as determined by their messaging habits. The results are collated in Table 6.1.

Age	15-24	25-35	35-50	50 and above
Number of people	6,975,000	3,487,500	2,092,500	1,395,000
Heavy	3,417,750,000	871,875,000	125,550,000	13,950,000
Medium	558,000,000	209,250,000	104,625,000	20,925,000
Light	139,500,000	174,375,000	20,925,000	20,925,000
Total	4,115,250,000	1,255,500,000	251,100,000	55,800,000
Grand Total	5,677,650,000			

Table 6.1 — Calculation of number of messages sent according to age group and messaging habits. The total whatsaspp user population (Table 4.1) has been distributed into different age groups as per table 3.2

Therefore, the total number of messages sent via Whatsapp each day in Delhi-NCR is 5,677,50,000!

Question 3

Estimate the number of tennis balls bought in India per month.

Answer 3

The number of cities in India i.e. ~1700. The catch in this problem is to analyze where all can we use tennis balls. Once we have the number of tennis balls used monthly, we can easily find the number of tennis ball bought in a month using the lifetime of tennis balls. Following is an effective way to segment this population:

Parameters	Possible Tennis ball usage							
Segment Level I	Tennis				Cricket			
Segment Level II	Urban				Urban			
Segment Level III	Metro	Tier-2	Small towns	Rural	Metro	Tier-2	Small towns	Rural
#cities	5	60	1600	5000	5	60	1600	5000
# sectors/cities	100	50	30	10	100	50	30	10
# grounds/sectors	5	3	2	0	50	40	30	10
# daily balls consumed	5	3	2		2	2	2	2
Total daily balls consumed	12500	27000	192000	0	50000	240000	2880000	1000000
Monthly ball consumption	4.4 Million							

Following were the key considerations in building the segmentation and the intermediate guesses:

Rural areas have negligible number of tennis courts.

Metro cities have the highest number of sectors.

For each sectors in metro cities, the number of grounds for both tennis and cricket is higher. This is both because of the bigger area and the higher buying capacity in metros. Number of balls consumed in metro cities per ground is higher because of the higher engagement in metro cities.

Question 4

How many cups of coffee were consumed in the United States in the past week?

Answer 4

Identify the variables to apply to this problem. Number of cups in the past week: This equals number of cups per day × 7 (for 7 days per week).

Assuming people to treat each day equally,
Percent of the population that drinks coffee: this would be an educated guess. Assuming 300 million people in the U.S., we could further assume that 20% are children that (we hope) do not drink coffee.

We could also guess that another 20% of the population does not drink coffee at all (perhaps they prefer tea or other beverages, or just water).

Number of cups per day: here our guess is that of the remaining 60% of people, half drink 2 cups per day, a quarter drink 4 cups per day, and a quarter drink 1 cup per day.

This averages out to 2 × 0.5 + 4 × 0.25 + 1 × 0.25 = 2.25 cups per coffee drinker per day.
Therefore the calculation is:

> 60% × 2.25 × 300,000,000 = 405 million cups each day
> 405 million cups × 7 days per week = 2.84 billion cups per week (you could round it to approximately 2.8 billion cups)

Question 5

What was the revenue for flat screen televisions sold in Australia in past 12 months?

Answer 5

Identify the variables to apply to this problem.
Population of Australia: Approximately 23 million people.

Assume that the average household is 3 people. It is worth noting that families probably have more than 3 people, but this is balanced out by people living alone, such as students and young professionals.

Here is a good example of rounding: you can say 8 million households (which is a little more than 23 million ÷ 3).

Assume households replace their televisions every 4 years.
Assume an average of 1 flat screen television per household. Some households might not have any, but others may have 2 or even 3.

Therefore, (8 million households) × (1 TV per household) ÷ (4 years/purchase) = 2 million televisions purchased in the past year.
Assume an average sale price of $600 isa reasonable average across higher-end TVs, which might cost more than $1,000, and smaller flat screen TVs, which can sell for $200 or even lower.

Therefore, 2 million × $600 = $1.2 billion annual Revenue for television sales in Australia.

Question 6

How many iPhones are currently being used in China?

Answer 6

Identify the variables to apply to this problem.

Population of China: Approximately 1.4 billion people.

There are several different approaches from this point; one approach is to make assumptions around the number of people that can afford iPhones rather than considering the number of households.

Based on very basic knowledge of China, even though the country is experiencing extraordinary economic growth, you might assume that the majority of the population is still very low-income and cannot afford an iPhone. Thus, you might estimate that 20% of the population could afford an iPhone.

Therefore, the total potential market size is 20% × 1.4 billion = 280 million iPhones.
What percent of this total market size is penetrated? There are many competing products that are cheaper, Therefore we estimate that 20% of this segment is currently using an iPhone.

Using these estimates, 20% × 280 million = 56 million iPhones are currently being used in China.

Question 7

What is the revenue of PEUGEOTS (french automobile company) sold in France per year?

Answer 7

Identify the variables to apply to this problem.
Population of France: Approximately 60 million people.
Assume an average household is 3 people. This leads to 20 million households (60 million ÷ 3).

Assume 20% of households have no car, as they are in urban cities such as Paris or Lyon.

Of the remaining households, assume an average of 1.5 cars per household.
Therefore, there are approximately 80% × 1.5 × 20 million households = 24 million cars in France.

Assuming a replacement rate of every 6 years, there will be (24 ÷ 6) = 4 million cars replaced per year.

Of these 4 million, how many are Peugeot brand? You could suggest that the French are quite patriotic, so perhaps 20% of the 4 million cars purchased each year are Peugeot.
Therefore, you estimate that (20 × 4 million) = 800,000 Peugeot cars are purchased in France per year.

Of the 800,000, assume 70% are new cars and 30% are used cars.
Assume that the average price is $30,000 for new cars, and used is $10,000 for used cars (this is assuming similar pricing)

Using these assumptions, (560,000 × $30,000) + (240,000 × $10,000) = $16.8 Billion + $2.4 Billion = $19.2 Billion.

Therefore, total Revenue of Peugeot cars sold in France per year is approximately $20 Billion

Question 8

What is the revenue for board game Monopoly sold in India per year?

Answer 8

Identify the variables to apply to this problem.

Population of India: Approximately 1.2 billion people.

Percent of population that is aged 8-15 in India: Assume 15%.

Total population of 8-15 year old children: 180 million.

Assumed percent living in areas where the board game is available for sale: 50%

Assumed percent of such children playing board games: 20%

Assumed number of board games purchased per child in this age range per year: 2

Average price for Monopoly: Assume Rs 600

Using these estimates, the annual Monopoly sales in India are as follows: 180 million × 50% × 20% × 2 × 10% × Rs 600 = Rs 2.16 billion annual Revenue.

Total Revenue for the children's version of Monopoly in India per year appears to be approximately $36 million. Note that because of the large chain of assumptions made, this estimate could be off significantly; in particular, the estimate is highly sensitive to the percentage breakdown assumptions for the relevant demographic (percent of 8-15 year olds living where the game is available; percent of those individuals who play board games; number of board games purchased by those customers annually etc.)

Question 9

How many square feet of pizza are eaten in the United States each month?

Answer 9

Take your figure of 300 million people in America. How many people eat pizza? Let's say 200 million. Now let's say the average pizza-eating person eats pizza twice a month, and eats two slices at a time. That's four slices a month.

If the average slice of pizza is perhaps six inches at the base and 10 inches long, then the slice is 30 square inches of pizza. So four pizza slices would be 120 square inches. Therefore, there are a billion square feet of pizza eaten every month.

To summarize:

300 million people in America

200 million eat pizza

Average slice of pizza is six inches at the base and 10 inches long = 30 square inches (height x half the base)

Average American eats four slices of pizza a month

Four pieces x 30 square inches = 120 square inches (one square foot is 144 inches), so let's assume one square foot per person

200 million square feet a month.

Question 10

Guess + Estimate the cost of painting the pillars of the metro lines in Delhi as of today. Also, every pillar consists of an advertisement sized 4X4 feet on each side of the pillar. Thus, every team is required to find the estimated net expenditure. (Cost of painting pillar—Revenue received from the advertiser).

Answer 10

Question is pretty straightforward. All you're required to do is estimate the number of pillars, assume values for the cost of painting and revenue earned per pillar. Please note that there are many more parameters that could've been considered. We'll limit it to the below assumptions for simplicity's sake.

The crucial step here is to estimate the number of pillars. We try to visualize how any metro line looks like. It starts from a station and ends at another. There are turns, splits and intermediary stations.

The line itself may be underground, on-ground or above the ground. Let's assign the number of pillars for each of these situations, taking care to choose simple numbers (multiples of 2, 5 or 10):

- Since the underground/ground level stations will not have any pillars, we will assume the number of pillars for them to be zero. The connections to the preceding and subsequent stations, however, will be taken into account.
- The station itself, will have more number of pillars, owing to the extra strength required for the stability of the station premises, therefore, we will assume 8 pillars for each station above ground level.
- The number of pillars increases for every turn/split(like the one after Yamuna Bank) in the tracks. We will assume the increase on the basis of the length of the track.
- For every 100m, there is a pillar. Therefore, there are 10 pillars in each km.

Basis the above assumptions, the length of the various Metro Lines, and subsequently, distance between two stations is assumed to be as follows:

Line	**Length (L)** (in km)	**Number of stations (S)**	**Number of Underground /Ground Level stations (U)**	**Number of stations above ground level (A)**	**Distance between two stations (L/S)** (in km)	**Distance to be considered (A X L/S)**
Blue	52	44	4	40	1.2	48
Red	25	21	4	17	1.2	20.4
Yellow	45	34	25	9	1.3	11.7
Orange	22	7	5	4	3.1	12.4
Violet	23	18	6	12	1.3	15.6
Green	20	17	0	17	1.2	20.4

Data taken from official DMRC website

Though the data above has been taken from the official website, it is not necessary to do so. You may assume whole numbers by general knowledge. For instance, you may already know that the blue line is the longest followed by yellow. Assign say, 50 and 45 to each respectively.

Line	Number of pillars for each station above ground level (8 X A)	Increase in pillars due to turns/splits	Number of pillars for the distance to be considered (10 X A X L/S)	Total number of pillars
Blue	320	20	480	880
Red	136	12	204	352
Yellow	72	0	117	189
Orange	32	0	124	156
Violet	96	8	156	260
Green	136	8	204	348
			GRAND TOTAL	2185

Now, once the number of pillars are calculated, we can estimate the cost of painting the pillars as well as the revenue. The following figures have been taken arbitrarily.
Assume the cost of painting each pillar to be Rs 1200.

- Assume the revenue from each advertisement placed to be 2500.

Number of pillars	Total cost of painting (1500 per pillar)	Revenue earned (2500 per pillar)	Total expenditure
2185	Rs. 32,77,500	Rs. 54,62,500	(Rs. 21,85,000)
		NET EXPENDITURE (FINAL ANSWER)	(Rs. 21,85,000)

The total expenditure, therefore, is Rs. (21,85,000).

Ace
Data Science
Interviews

Analytics Vidhya
Learn everything about analytics

Case study

Question 1

There are multiple cab services these days, and you, as a customer has to make an efficient decision, and select the cab based on your requirement. Some important terms that we will use throughout are defined below:

1. **Base Fare**: Initial amount billed to sit in a cab
2. **Excess km fare**: Billed amount on distance after complimentary ride
3. **Time fare**: Billing on the time spent in the cab
4. **Minimum fare**: This is the minimum amount you will be billed
5. **Tolls and excess fee** : This is the excess charge you need to pay to compensate for the long distances outside main city
6. **Taxes** : Taxes is are over and above the bill
7. **Premium multiplier**: In crowded/congested time, you will be bill something like 1.4 – 2.5 X of the actual bill amount.

Here are the details of the three cab services that you must compare.

1) The first cab service, let's call it A, has the following fare for the three types of cabs - micro, mini, prime.

Standard Rate				
Category	Minimum Bill	Extra km charges	Wait time charges	Ride time charges
Micro**	Rs 40	Rs 6 per Km	N/A	Rs 1 per Min
Prime**	Rs 100 for first 4 Km	Rs 13 per Km	N/A	Rs 1 per Min (Post 5 Min)
Mini**	Rs 80 for first 4 Km	Rs 10 per Km	N/A	Rs 1 per Min (Post 5 Min)

In addition to this, minimum fare of Micro is fixed at Rs. 50.

2) The second service, called B, also has three types of cabs - nano (equivalent to micro), tata indica (equivalent to mini), sedan (equivalent to prime), and here are the prices of the same.

CAR	FARE
Nano	□35 (□5.0/km, □ 1.5/min of trip time)
Tata Indica AC	□49 (□6.0/km after 2.0kms, □ 1.5/min of trip time)
Sedan	□75 (□8.0/km after 2.0kms, □ 1.5/min of trip time)

3) The third company, company C, has the following fare for the cabs offered. They do not have a micro or nano but on the other hand provide an XL.

You need to suggest the most optimised cab, for the following situations-

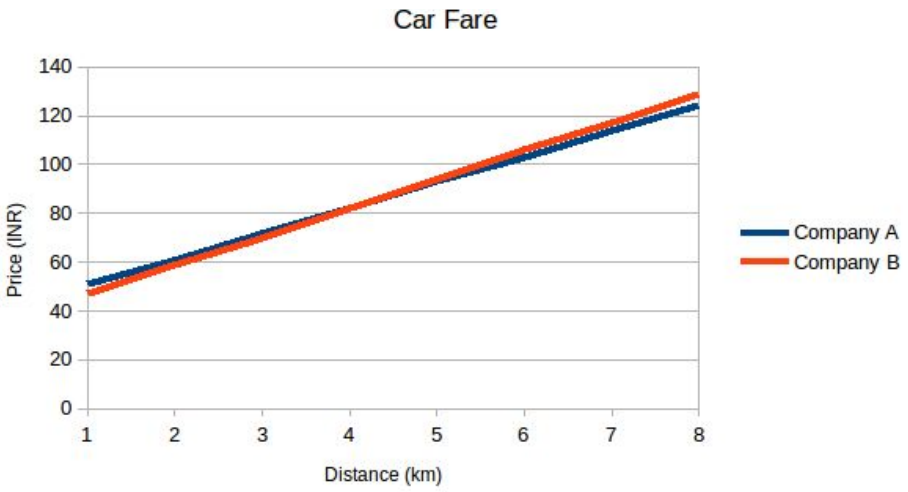
1. Which of the MICRO vehicles will be cheapest if your distance lies between 1 to 8 kms?
2. Which MINI vehicles are the cheapest if your distance is between 1 to 10 kms?
3. If you get a free upgrade from Company A - Micro to Mini, will it be cheaper than Company C, Mini for distance 2-6 kms?
4. Company C is charging a multiplier of 2.1 and Company A is charging a multiplier of 1.4 on their Sedan Vehicles (Company A Prime vs. Company C X). Which one will cost less ?
5. You have already booked UberGo for a multiplier of 1.5 and now you are getting a Company A Mini vehicle without peak charges? The challenge is that if you cancel an Company C, you will incur a cancellation charge penalty. However, if you choose to cancel, you stand a chance to save on peak charges. At what distance will you break even on the cancellation charges on Company C, in case you choose to go ahead with Company A?

Answer 1

We shall take up the questions one by one.

1. Since only company A and B offer the micro cabs, we will choose a cab from one of the two. Since the distance is provided to be 1-8 km range, let's check out the price for this distance for both the company.

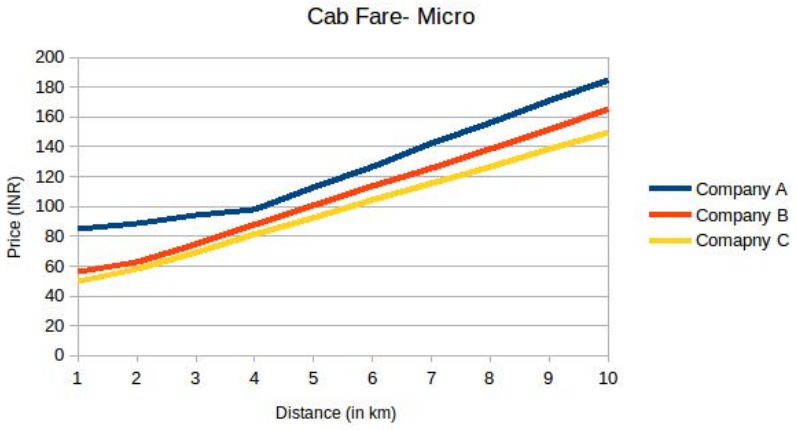
Distance	Company A	Company B
1	51	47
2	61	59
3	72	70
4	82	82
5	93	94
6	103	106
7	114	117
8	124	129



If the distance is less than 4, a cab from company A should be booked while for distance more than 4, cab from company B should be preferred.

2. Since all the companies have an option of a cab type mini, so we must check the cab fare for all the three companies.

Distance	Company A	Company B	Comapny C
1	85	56	50
2	89	63	58
3	94	75	69.5
4	98	88	81
5	113	101	92.5
6	127	114	104
7	142	126	115.5
8	156	139	127
9	171	152	138.5
10	185	165	150

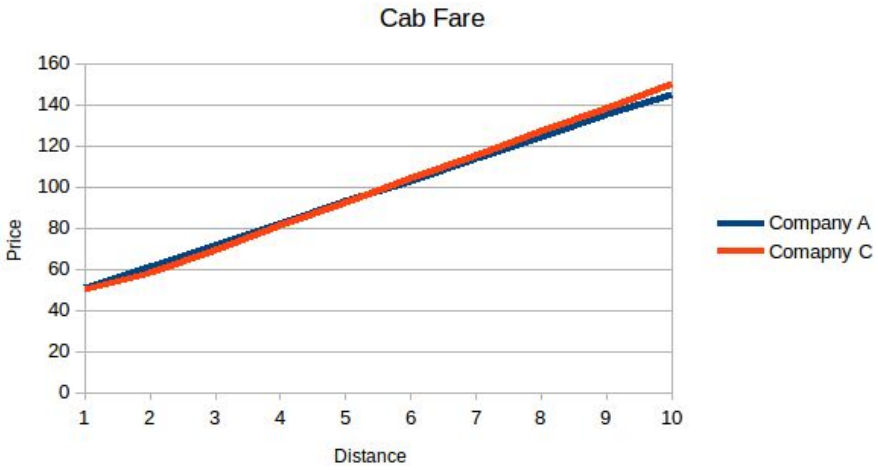


We clearly see that Company C is the cheapest car for the distance range and also seem to be the cheapest option for distances beyond 10 km looking at the

trend.

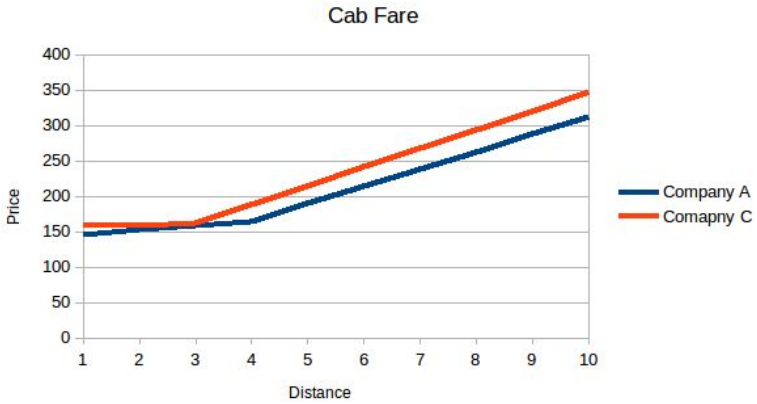
3. We will compare company A micro and company C mini. The trend looks to be very interesting. *Company C mini* starts at a slightly lower rate than *Company A Micro* but Company A takes over after 5 km distance. So, again our answer will be *cannot be determined* in the distance range. However, it is interesting to notice that a *Company A Micro* taxi comes out to be more expensive than *Company C Mini* car for shorter than 5 km distance.

Distance	Company A	Comapny C
1	51	50
2	61	58
3	72	69.5
4	82	81
5	93	92.5
6	103	104
7	114	115.5
8	124	127
9	135	138.5
10	145	150



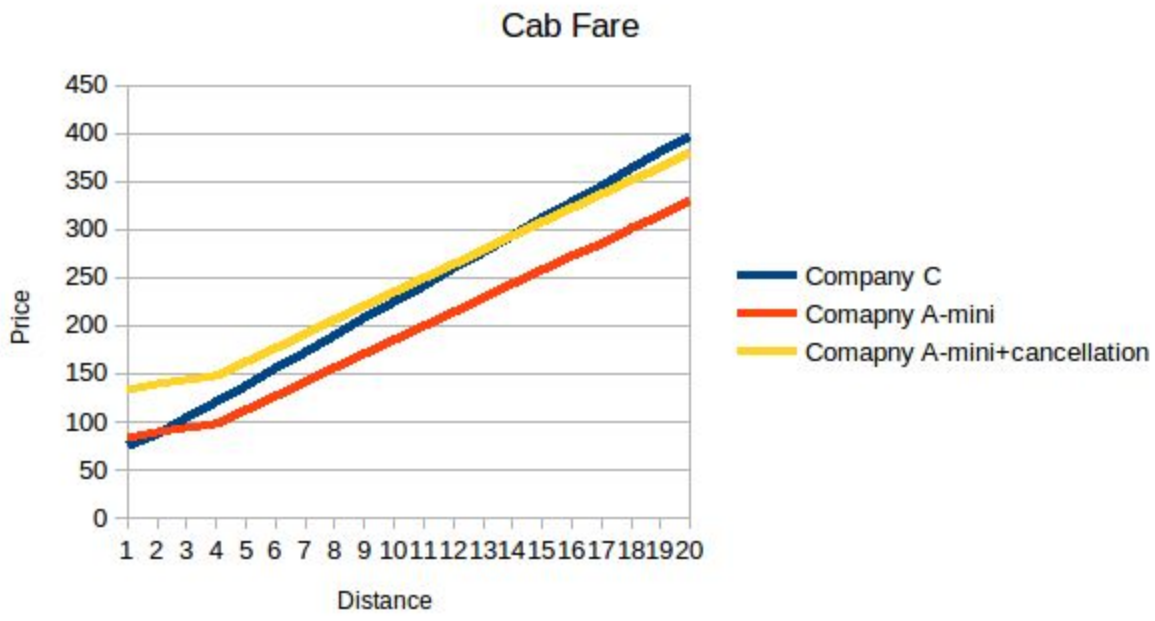
4. Multipliers are often added in peak traffic hours. It becomes very difficult to compare rates in such cases. Here's one of those scenarios. Company A is generally more expensive compared to Company C, but in such extreme multiplier cases, we see that Company A comes out to be cheaper option throughout.

Distance	Company A	Comapny C
1	146	158
2	153	158
3	159	163
4	165	189
5	190	215
6	214	242
7	239	268
8	263	294
9	288	320
10	312	347



5. What you need to keep in mind is the cancellation charges of Company C. Here is the table for Company A vs. Company C :

Distance	Company C	Comapny A-mini	Comapny A-mini+cancellation
1	75	84	134
2	87	89	139
3	104	93	143
4	121	98	148
5	138	112	162
6	156	127	177
7	173	141	191
8	190	156	206
9	207	170	220
10	225	185	235
11	242	199	249
12	259	214	264
13	276	228	278
14	294	243	293
15	311	257	307
16	328	272	322
17	345	286	336
18	364	301	351
19	380	315	365
20	397	330	380

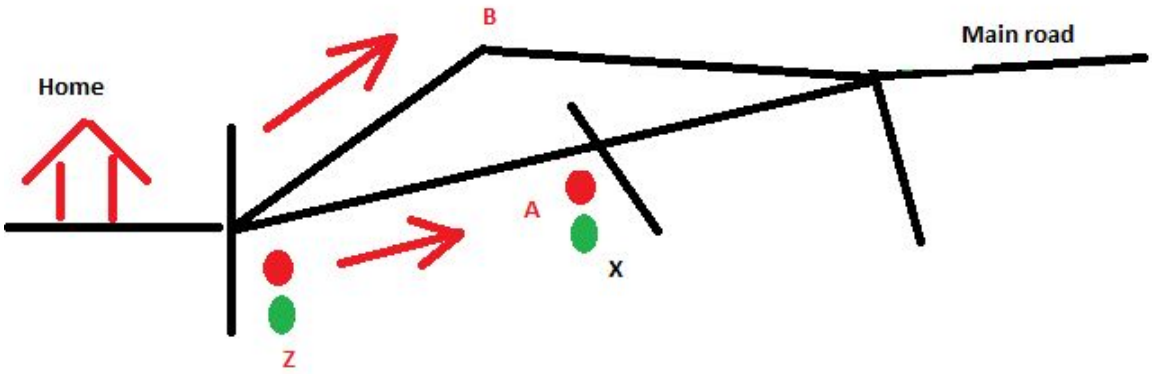


As it can be seen from both table and the graph, the break-even only happens

between 13-14 kms. Hence, you should make a switch only if you want to travel more than 13 kms.

Question 2

Suppose there are two alternate roads which I can take to reach the main road from my home to go to work. The average speed on each of the road comes out around 30 km/hr. Let's call the two roads as road A and road B. Total distance one needs to travel on road A and road B is 1 km and 1.3 km respectively to hit the same point on the main road. Note that, before the two roads split, I see a signal (say Z) which is common to both the roads and hence does not come in this calculation. See figure for clarifications.



Considering the given situation, answer the following questions.

1. What are the possible factors, I should consider to come up with the total time taken on each road?
2. Which road should one take to reach the main road so as to minimize the time taken? And what is the difference in total time taken by the two alternate routes?
3. Recently, one of the junction (say, X) on road A got too crowded and a traffic signal was installed on the same. The traffic signal was configured for 80 seconds red and 20 seconds green. Let's denote the seconds of signal as R1 R2 R3 … G1 G2 G3 . Here, R1 denotes 1 sec after signal switched to red. Does it still makes sense to take

road A, or to switch to road B provided the average speed on the road A is still the same except the halt at signal?

4. If I reach the signal at R1, I will be in the front rows to be released once the signal turns green. Whereas, if I reach the signal at R80, I might have to wait for some time even after signal turns green because the vehicles in the front rows will block me for some seconds before I start. Let's take some realistic guesses for the wait time after signal turns green.

   R1 – R 10 : 0 sec , R11-R20 : 3 sec , R21 – R60 : 10 sec, R61 – R80 : 15 sec, G1-G15 : 5 sec, G15-G20 : 0 sec

   Does it still makes sense to take road A, or to switch to road B provided the average speed on the road A is still the same except the halt at signal?

Answer 2

1. The important factors should be, road length, road quality, traffic on road, signals if any. The area around road (highway/society/parks).

2. The second question simply asks us to determine the time taken in both the cases i.e., taking road A and road B. Here is the calculation

Time taken on road A = 1/30 * 60 min = 2 minutes
Time taken on road B = 1.3/30 * 60 min = 2.6 minutes = 2 min 36 sec
Hence, the clear choice is road A. Road B would have taken 36 sec more than road A.

3. Since road A has a signal now, we will try to find out the average time taken for a trip. So let's say we come to the signal at R1, we'll wait for 80 sec, for R2 - 79 sec… and so on.

E(halt time) = (1+2+3+4+…….80)/(80+20) = (80*81)/(100*2) = 32.4 seconds.
Still we see 32.4 sec < 36 sec. Hence, it still made sense to take road A.

4. We need to calculate the new average time now,

E(halt time) = {(1+2+3+4+…….80) + 3*10 + 10*40 + 15*20 + 5*15}/(80+20) = 40.15 seconds.
This time the game changes and as 40.15 sec > 36 sec, prefer road B over road A.

Python for data science

Question 1

The two most commonly used data structures in python are lists and dictionaries. Can you list down the basic differences between list and dictionary.

Answer 1

Lists store values, in an ordered sequence. Each element is numbered, starting from zero, i.e. the first element in a list is numbered 0, the second 1, the third 2, and so on. We can remove values from the list, and add new values to the end.
Example: list1 = [1,2,3,4,10,100]

Dictionary is an unordered set of ***key: value*** pairs. Unlike lists, instead of using numbers to represent values, we have keys in a dictionary. The values in the dictionary can be removed/ modified and new values can be added.
Example: dict={'English': 90, 'Mathematics': 95, 'Physics':80}

Question 2

Lists and tuples in python are used to store values in an orderly manner (unlike in dictionaries where we do not have an order between the keys). What is the difference between list and tuples?

Answer 2

The only difference between lists and tuples is that lists can be updated and it it a mutable data type, while the same cannot be done with tuples, it is an immutable data type. This means that we can add, update and replace values in a lists but not in tuples.

Question 3

Suppose I have a list list1=['a', 'b', 1, {'Name': 'ABC', 'age': 22}], what will be the result of the following command?

>>list1[1]

>>list1[3]

>>list1[3]['age']

Answer 3

- The list has index starting from 0, so the output of the command list1[1] will ba 'b'.
- Since the fourth element in the list is a dictionary, the result would be {'Name': 'ABC', 'age': 22}
- List1[3]['age'] will give the output 22

Question 4

Suppose I have a list a=[1,2,3,4]. To add a value '5' to the list, I can simply use a.append(5). What if I want to add 10 different values, from 5 to 15, to the list? Will I use append 10 times?

Answer 4

Instead of using append() in this case, we can use extend(). So to add 10 different values, we can use this command: a.extend([5,6,7,8,9,10,11,12,13,14,15])

Question 5

In order to add new elements to a list, we can use one of the two commands - list.append() or list.extend()? For a given list: *a=[1,2,3,4]*, what will be the result of the following commands?

>>a.append([6,7,8])

>>a.extend([6,7,8])

Answer 5

Append is used to add a single value at the end of the list. For the given example, append assumes [6,7,8], to be a list. So the result will be = [1,2,3,4,[6,7,8]]

Extend is used to add multiple elements to the list. In this case, the numbers 6,7 and are considered to be three different elements. The output of second command will be = [1,2,3,4,6,7,8]

Question 6

If we have a list arr = ['a','b','c','d','e','f'], what will be the output of arr[2:-1]?

Answer 6

Python lists support negative indexing. Negative indexing starts from the rights end of the list, because the first element from left is indexed at 0, the first element from right is indexed at -1. Thus arr[2:-1] will return the elements from third position to second last position in the list i.e. ['c','d','e'].

Question 7

What is the output of following?

```
arr = [1,2,3,4,5,6,7,8,9,1,2,3,4,5,6,7,8,9]
s = set(arr)
print(s)
```

Answer 7

This creates a set of all the elements present in the list arr, referenced by s. Since elements cannot be repeated in a set , s will contain only distinct elements from arr i.e. 1,2,3,4,5,6,7,8,9.

Question 8

What will be the output of the following?

- ➢
```
for i in range(2.0):
    print(i)
```
Answer In python range function can only iterate using integer values. Since 2.0 is a floating point number thus the above code will result in an error.

- ➢
```
for i in range(2,10):
    print(i)
```
Answer The above code prints numbers from 2 to 9.

➢ for i in range(2,10,2):
    print(i)

Answer 8

Above code prints all the even numbers from 2 to 8(10 is exclusive), because the third parameter in range is the step-size for iteration.

Question 9

What is the output of the following

>> [1,2,3] + [4,5,6]
>>[1,2,3] + [5]
>>[1,2,3] + 5

Answer 9

- The + operator is overloaded for lists in python, it appends the lists together. The above code gives [1,2,3,4,5,6]
- Similarly for the second command, the result would be a list [1,2,3,5]
- The third command would give an error

Question 10

What is the difference between *break* and *continue* in python?

Answer 10

The “break” statement breaks the execution of the current loop that is being iterated and the code after loop starts executing.

Whereas “continue” forces the control to skip the current iteration of the loop.

Question 11

If a = 2, b = 5 . What is the output of a**b?

Answer 11

The ** operator in python is used for exponentiation. a**b means a^b so 2**5 = 32

Question 12

Code :-

```
def f_1():
    print(a)
def f_2():
    global a
    print(a)

a =  5
f_1()
f_2()
```

What is the output?

Answer 12

To use a global variable in a function , it needs to define it locally using “global variable_name”. Thus f_1 produces and error because there is no variable by the name a in its namespace whereas f_2 prints the value of the global variable a.

Question 13

What is the output of the following command in Python?

>> *True* ***and*** *False*

Answer 13

and is a logical operator in python and returns False if any of its operands is False.

Question 14

Suppose we have a dataframe ‘df’. The command df.head() and df.loc[:5], will give the same results? Why do we need .loc?

Answer 14

In pandas, loc is used to select rows based upon their index numbers.Thus, loc[:5] returns the first 5 rows from the dataframe also the default value for **head** function is 5 so it also returns the first 5 rows from the dataframe.

While using df.head(), we cannot select specific rows. But the loc operator lets us select rows from a dataframe using index numbers.

Question 15

What is the difference between file.tsv and file.csv? How are they used?

Answer 15

- .tsv and .csv are two different file formats referring to tab separated values and comma separated values respectively.

- Tsv are often used in text data as comma can represent a punctuation.
- Tsv and csv files can be loaded into a pandas dataframe using the read_csv function.

Question 16

What is the difference between the following code lines?

```
>> df.drop(['Age'], axis = 1)
>> df.drop(['Age'], axis = 1, inplace=True)
```

Answer 16

The first case, since the value for parameter *inplace* is not specified, the default *inplace=False* is considered. It prints the dataframe without the column specified in the parameter, 'Age' in this case. But this does not make any change to the original dataframe.

When the parameter inplace is set to be true, the column is dropped from the original dataframe.

Question 17

Consider the following function:

```
def greater(x,y):
    return x>y
```

What will be the datatype of the output from the following function, given that the numbers x and y are two integers?

Answer 17

The function compares the two numbers, and would return True or False. Thus the data type would be boolean.

Question 18

What are lambda functions, when to use them and when not to use them.?

Answer 18

Lambda functions are anonymous functions and are not bound to any name/identifier. They are used when we can to perform a “small set of actions” and which does not require defining a separate function. They are lightweight and efficient.
They are not recommended to use when the set of actions is not “small” or are too complex. Defining a separate function is better in this scenario.

Question 19

Which is more efficient, Iteration or recursion?

Answer

Python is not primarily a functional language and therefore there is a little overhead when using recursions, Therefore iterative methods are mostly efficient.
But, there are problem where applying iteration can be tedious (Eg: Tower of Hanoi) and therefore can be efficiently solved using the recursion techniques.

Question 20

What do you mean by scope of a variable?

Answer 20

A scope of a variable refers to the environment to which it is restricted to, A variable of higher scope can used used at lower scope but reverse is not true.

Question 21

What do you mean by list aliasing and list cloning in python?

Answer 22

Aliasing

Python uses the concept of name tagging in list, which implies that when when one list is assigned to another, the two names will represent the same list.

```
x = [ 1, 2, 3, 4, 5, 6, 7]
y = x
x.append(8)
print( x, y)
```

Output:

```
[ 1, 2, 3, 4, 5, 6, 7, 8] [1, 2, 3, 4, 5, 6, 7, 8]
```

Cloning

In list cloning, we have to explicitly mention that we wish to make a copy of a list and store it into another.

```
x = [ 1, 2, 3, 4, 5, 6, 7]
y = x[ : ]
x.append(8)
print( x, y)
```

Output:

```
[ 1, 2, 3, 4, 5, 6, 7, 8] [1, 2, 3, 4, 5, 6, 7]
```

R for Data Science

Question 1

There are different data structures in R - Vector, List, Matrix and Dataframe. How do you decide which to use when?

Answer 1

Vector: Vectors are used when one needs to store a sequence of data elements of the same basic type.
List: Unlike vectors, lists can store elements of different types like – numbers, strings, vectors or another list.
Matrix: A matrix is is used to bind vectors of the same length. Like vectors, matrix has the same condition that all the elements of a matrix must be of the same type (numeric, logical, character, complex).
Dataframe: A data frame is more generic than a matrix, i.e different columns can have different data types (numeric, character, logical, etc). It combines features of matrices and lists like a rectangular list.

Question 2

Suppose I have the following file saved with the name *Dataframe.csv*

Alpha	125.5	0
Beta	235.6	1
Beta	212.03	0
Beta	211.30	0

Alpha	265.46	1

What is the syntax to read this file in R?

Answer 2

Since the given file does not have a header, it is important that we pass the parameter *header=FALSE*. Otherwise, the first row of the file will be read as the header. So the exact syntax would be : csv2('Dataframe.csv',header=FALSE,sep=',')

Question 3

Can we store categorical variables in R? If yes, how?

Answer 3

R has a special data structure called "factor" to store categorical variables. It tells R that a variable is nominal or ordinal by making it a factor.

```
>> gender = c(1,2,1,2,1,2)
>> gender = factor(gender)
>> gender
```

Question 4

For the vector created above, can we plot a frequency table to understand the distribution of data?

Answer 4

table(gender) will return the frequency table for gender as shown below:
(the frequencies for the two levels 1 and 2 are returned)

```
gender
1  2
3  3
```

Question 5

Consider the following list:
list_data <- list("Blue", "Green", c(21,32,11),)

What will be the output for the command:

```
>>list_data[1]
>>list_data[3]
```

Answer 5

The indexing of lists in R starts from 1, so the output of first command will be ‘Blue’ and the output for the second command will be 21, 32, 11

Question 6

Suppose I have the following list of colors-
(“red”, “green”, “blue”, pink”). Now I want to capitalize every letter in the list without having to create a new list or updating every element. How can we do that?

Answer 6

From the stringr package, use the following function str_to_upper(list_name). The result would be

“RED”, “GREEN”, “BLUE”, “PINK”

Question 7

Consider the following list -
list_color=list("red", "blue","pink", "gren", "purple")
I misspelled the last second element. Can I update this element or should I delete and add it again at the end?

Answer 7

Elements can be updated in a list. To do so, we can use the following code:

```
>list_color[5]="green"
```

Question 8

I have two lists
list1=list(1,2,3)
list2=list(4,5,6)

How can I add all the elements of the second list at the end of the first list?

Answer 8

The following command can be used to add the elements of one list at the end of the other

```
>new_list=c(list1,list2)
```

Question 9

How can I print all the odd numbers between 1 to 9?

Answer 9

To print all the odd numbers, we can use either of the two commands -
seq(1, 9, 2) or seq(1, 10, 2).

Question 10

Which of the following is the correct way to assign variables in R?

a. myVar <- 14
b. myVar = 27
c. "helloWorld" -> myVar
d. myVar -> "helloWorld"

Answer 10

(all but d)

Question 11

Suppose we have two vectors a and b, given as:

```
a <- c(2,3,4)
b <- c(1,2)
```

What will be the value of another vector d defined as:

```
d<-a*b
```

Answer 11

Here the elements in the two vectors are multiplied. Since the second vector has only two elements while the first vector has three element, the third element for vector b is taken as 1. So the result will be (2,6,4)

Question 12

If a = 2, b = 5 . What is the output of a**b?

Answer 12

The ** operator in python is used for exponentiation. a**b means a^b so 2**5 = 32

Question 13

Have a look at the following code. What will be the result of f(6):

```
y <- 3
        f <- function(x) {
                    y <- 2
                    y ^ 2 + g(x)
                     }
        g <- function(x) {
                    x * y
                     }
```

Answer 13

The answer will be 22. The function f(x) has the value of y as 2 and x as 6, while gor g(x), the y value is 3 and x value is 6. So the output of g(x) will be 6*3 = 18. The output of f(x) will be 4+18 = 22.

Question 14

Let's say I have the following vector:

C<-c("Delhi is a great city. Delhi is also the capital of India.")
Will there be any difference in the output of the two commands?

>>sub("Delhi","Delhi_NCR",C)

>>gsub("Delhi","Delhi_NCR",C)

Answer 14

sub command only replaces the first occurrence of a pattern and gsub replaces the same throughout. So in the first case, delhi is replaced with Delhi_NCR in the first

statement alone. The output will be : "Delhi_NCR is a great city. Delhi is also the capital of India."

In the second case, on using gsub, the word Delhi is replaced both the times. Hence the result is: "Delhi_NCR is a great city. Delhi_NCR is also the capital of India."

Question 15

Suppose there are 2 data-frames “A” and “B”. A has 34 rows and B has 46 rows. What will be the number of rows in the resultant dataframe after running the following command?

```
merge(A,B,all.x=TRUE)
```

Answer 15

Using all.x forces the merging to take place on the basis of A and hence will contain the same number of rows as of A, which is 34 in this case

Question 16

What should be the output of the following code

```
a <- c(1,1,1,1,2,2,2,2,2)
    b <- c(10,12,15,12,NA,30,42,38,40)
    s <- split(b,a)
    lapply(s,mean)
```

Answer 16

In the first case, the mean is calculated for 10,12,15 and 12 since a has 4 ones. For the second case, when a has 5 twos, one of the numbers is NA, hence the result will be NA. If NA is replaced by a number, the mean will be calculated using the number.
So the result will be (12.25 and NA)

Question 17

How to extract first 3 characters from a word

```
x = "AXZ2016"
```

Answer 17

```
substr(x,1,3)
```

Question 18

What will be the output of the following code:

```
paste(1:3,c("x","y","z"),sep="")
```

Answer 18

The result will be - [1x 2y 3z]

Question 19

Suppose I have the following series of numbers:

```
> x = sample(1:50, 10)
> x
6 31 30 48 40 29  1 10 28 22
```

What will be the result of the following commands:

```
>sort(x)
>rank(x)
>order(x)
```

Answer 19

The sort() function is used to sort a 1 dimension vector or a single variable of data. The rank() function returns the ranking of each value. The order() function returns the indices that can be used to sort the data.

> sort(x)

1 6 10 22 28 29 30 31 40 48 (*It sorts the data on ascending order)*

>rank(x)

2 8 7 10 9 6 1 3 5 4 (2 implies the number in the first position is the second lowest and 8 implies the number in the second position is the eighth lowest.)

>order(x)

7 1 8 10 9 6 3 2 5 4 (7 implies the 7th value of x is the smallest value, so 7 is the first element of order(x) and i refers to the first value of x is the second smallest.)

Statistics and Probability

Question 1

Suppose I have monthly income values of 7 people as: 10000, 20000, 15000, 22000, 18000, 20000 and 25000. Now I add another value to this list which is 100000. Will the mean, median and mode change on adding this new data?

Answer 1

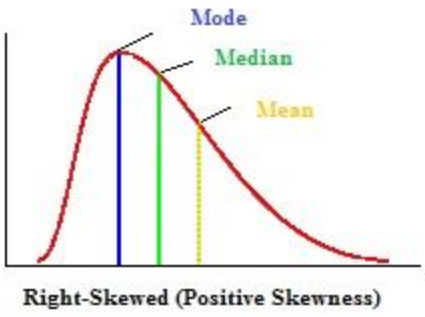
The new value that is added is very large as compared to the values currently present in the series. This value can be treated as an outlier. The mean value of data is drastically affected by outliers. The mode and median value will not be changed in this case.

Question 2

The mean of a distribution is 25, the median is 23, and the mode is 20. Based on the given information can you determine whether the plot for this distribution, will be positively skewed or negatively skewed?

Answer 2

Since the mode value is less that the mean and median, the data will be positively skewed. Here is a diagrammatic representation for the same.



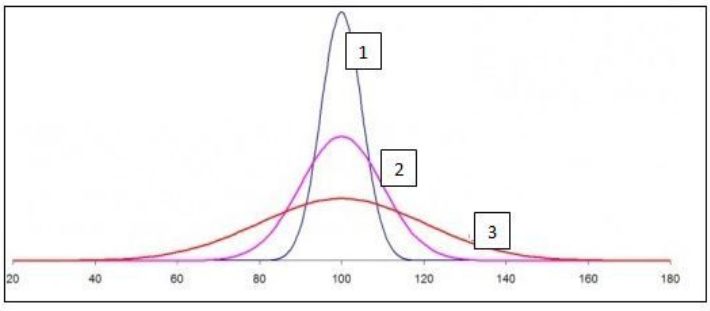
Question 3

For a given normal distribution, to increase the spread and decrease the height of the curve, how should the standard deviation values be changed?

Answer 3

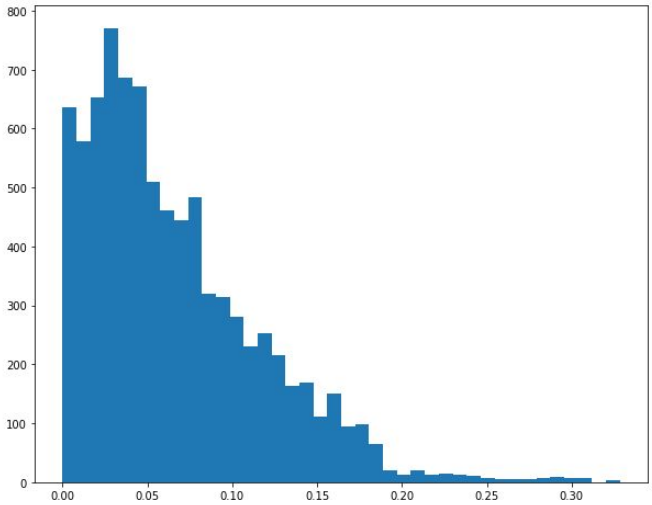
The spread of the data would increase when the value of standard deviation is increase. For the three curves in this image, the standard deviation for curve three is the highest and that of curve 1 is the lowest.

3 > 2> 1



Question 4

The graph shown below represents the distribution of a variable. What is the relationship between mean, median, mode.



Answer 4

From the graph we can interpret that the mode value is close to 0.04. The median value will be approximately between 0.05 and 0.1. Also, since we have a number of very high values the mean value would be closer to 0.15. To define the relationship - mean> median> mode.

Question 5

The interquartile range is the difference between the minimum and maximum value of the numbers in a given series. Is the statement correct?

Answer 5

No, the difference between the minimum and maximum value is the range of the series. The interquartile range is the difference between the the **first quartile** (value below which the lower 25% of the data exist) and the **third quartile** (value below which the lower 75% of the data exist).

Question 6

Suppose I have calculated the variance for a set of numbers. Now I add 20 to all the values in the set, how will this change affect the variance? Will it be increased or decreased by 20?

Answer 6

Variance is the average of squared differences from the mean. Although the mean value of the numbers might change but the overall difference will not be affected. Hence, there will be no change in the value of variance.

Question 7

Consider the same situation given above. How will the standard deviation change after adding a constant value to all the numbers in the set.

Answer 7

There will be no change in the standard deviation as well.

Question 8

How many values, in the below series, fall within one standard deviation of the mean?
180, 313, 101, 255, 202, 198, 109, 183, 181, 113, 171, 165, 318, 145, 131, 145, 226, 113, 268, 108

Answer 8

On calculation, the mean and the standard deviation of the given data comes out to be 181.25 and 64.65 respectively. Now, one standard deviation above an below the mean would give the range 117 to 245. Total 11 numbers fall in this range.

Question 9

If the mean of a certain set of numbers is 14 and variance is 25. Then what would be the standard deviation?

Answer 9

Standard deviation is simply the square root of variance. So it will be 5.

Question 10

How are histograms different from bar charts

Answer 10

Bar charts are used to plot the distribution of categories. It is mainly used for discrete data such as gender, category etc.. While histogram is used for continuous data such as age, salary etc.

Question 11

How can you decide the appropriate bin size for histogram?

Answer 11

A very large bin size might hide some valuable information while a very small bin size will have too many spikes to make a clear inference. Usually the bin size is selected based on the size of the dataset.

For a small amount of data, use wider bins to eliminate noise. While working with a lot of data, it is good to use narrower bins because the histogram will not be that noisy.

Question 12

There are a total of 8 bows with 2 each of green, yellow, orange & red colour. In how many ways can you select 1 bow?

Answer 12

You can select one bow out of four different bows, so you can select one bow in four different ways.

Question 13

The difference between probability density function and probability mass function.

Answer 13

Probability distribution function is used to define the probability distribution. The distribution function is of two types depending on whether we are dealing with continuous or discrete values.

Probability density functions are used for continuous variables while the probability mass functions are for discrete distributions.

Question 14

Two unbiased coins are tossed. What is the probability of getting at most one head?

Answer 14

The probability is ¾. Since the total possible outcomes when two coins are tossed will be 4 (HH, HT, TH, TT). We are to find at most 1 head, so the favorable outcomes are (TT, HT, TH). Therefore, Probability = $^{3}/_{4}$.

Question 15

For a continuous random variable X with mean 100 and standard deviation 10 , what is the probability of getting X = 75.00?

Answer 15

Here since the probabilities are continuous, the probabilities form a mass function. The probability of a certain event is calculated by finding the area under the curve for the given conditions. Here since we're trying to calculate the probability of getting X exactly 75 – the area under the curve would be 0.

Question 16

A fair six-sided die is rolled twice. What is the probability of getting 2 on the first roll and not getting 4 on the second roll?

Answer 16

The two events mentioned are independent. The first roll of the die is independent of the second roll. Therefore the probabilities can be directly multiplied.

P(getting first 2) = 1/6

P(no second 4) = 5/6
Therefore P(getting first 2 and no second 4) = 1/6* 5/6 = 5/36

Question 17

Can you list down the properties of normal distribution?

Answer 17

The normal distribution has the shape of a bell curve
area under curve is 1.
The mean, median and mode of the distribution coincide.

Question 18

Normal distribution is symmetric about the origin (0,0)?

Answer 18

No, the normal distribution is symmetric about the mean. But for a standard normal curve the values are scaled between -1 to1. In this case, the value of mean is 0 and standard deviation is 1. So the standard normal curve is symmetric about 0.

Question 19

Normal Distribution is applied for discrete or continuous variables?

Answer 19

Non linear distribution is used for continuous variables only. Frequency distribution is used for discrete variables.

Question 20

If we are trying to find out the probabilities for different number of heads/tails in N number of coin tosses, what probability distribution will good for such a problem?

Answer 20

Since the given problem has only two outcomes, i.e. heads and tails, the best approach would be to use Binomial distribution.

Question 21

In a binomial distribution, the value of mean and variance is equal?

Answer 21

No, the value of mean is calculated as n*p (where p is the probability of success of the event) while the variance is n*p*q (where q is probability of failure).

Question 22

We are given with a feature from a dataset that has a mean of 150 and standard deviation of 15, find out the Z value for samples having value greater than 180. What is the significance of Z value?

Answer 22

180 is 2 standard deviation away from the mean. So we can directly say that z value is 2. Z value defines how far is the observed value above or below the mean value.

Question 23

What happens to the confidence interval when we introduce some outliers to the data?

Answer 23

The confidence interval depends on the standard deviation of the data. On introducing outliers in the data, the standard deviation increase, and so does the confidence interval.

Question 24

What does confidence interval = 95% mean?

Answer 24

On repetitive sampling, 95% of the times, the interval estimates will contain the population mean.

Question 25

How can you lower the margin of error?

Answer 25

The following is the equation for calculating the margin of error.

$$Z_{a/2}\,\sigma/\sqrt{n}$$

By looking at the expression of margin of error we can say that margin of error will decrease with increase in sample size.

Question 26

What does the p-value for a statistical data signify?

Answer 26

The p value for a statistical test is used to draw conclusions, basically deciding to accept of reject the null hypothesis.The range of p-value is always between 0 and 1. Generally the threshold for p value is set to be 0.05. When the value is below 0.05, the null hypothesis is rejected. If the value is equal to or greater than the threshold, 0.05 in this case, the null hypothesis is not rejected.

Question 27

Can you explain the difference between one sample t-test and paired t-test?

Answer 27

Taking a sample from a population, the one sample t-test is used to identify if the sample is the representative of the overall population. For example t test can be used to determine if the population of delhi is a true representative of the population of India.

Paired t-test is used when we compare the same population before and after an intervention. For example, suppose an evening tution class is introduced for students of class 10th. Paired t-test can be used to identify if the evening tuitions made any difference in the marks of students before and after the change.

Question 28

When do we use a chi square test?

Answer 28

chi squared test is used to find out if two categorical variables are dependent on each other and what is the degree of their correlation.

Question 29

What does the correlation value = -0.95 suggest?

Answer 29

Strong negative relationship between the variables. One variable strongly decreases with increase in other.

Question 30

Suppose we need to determine the average age of all the working professionals in India. As you can imagine, gathering this amount of data is a tedious task. Is there an alternative to collecting the data for every individual across the country?

Answer 30

The concept of central limit theorem can be applied in this case. Collect the data of multiple small samples, suppose working professional from delhi, mumbai, Bangalore and so on, and take the means of these samples. The mean of sample means should represent the whole population (India in this case)

Question 31

Consider the following two situations -

Situation 1: The doctor identifies a male patient as pregnant.

Situation 2: The doctor identifies a female pregnant lady as not pregnant.

Given that the null hypothesis is that the patient is not pregnant, classify the above two situations as Type I and Type II error.

Answer 31

For the first situation, the null hypothesis is incorrectly rejected, which is a type II error. In the second situation, the null hypothesis is falsely accepted hence it is type I error.

Question 32

What is the difference between the one sample and the two sample t-test?

Answer 32

In one sample t-test, we are trying to infer whether a sample belongs to a population distribution.
In two sample t-test, we are trying to infer whether the two independent samples are from same distribution or not.

Question 33

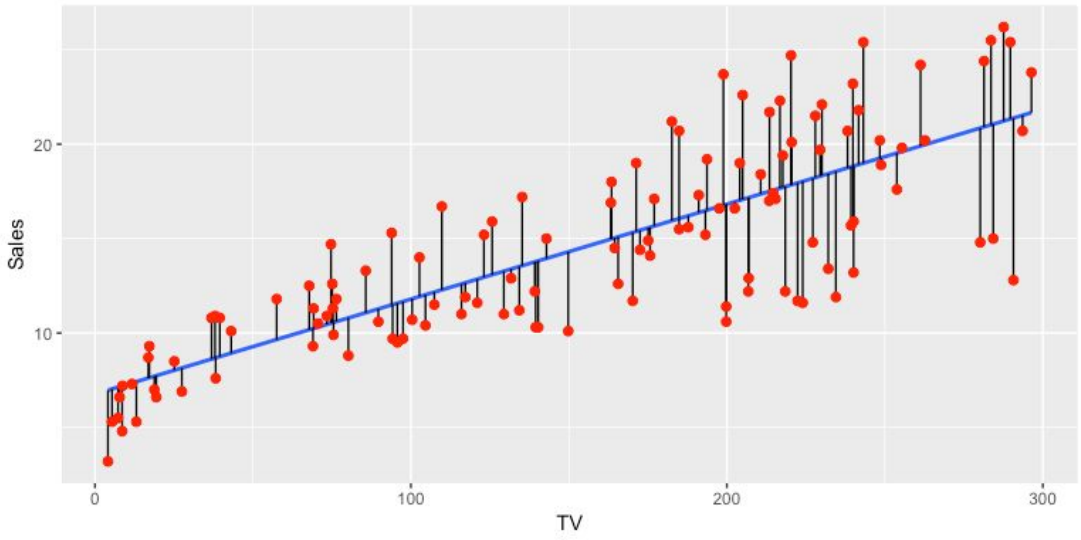
When do we use paired t-test?

Answer 33

Paired t-test is used when we want to check whether the sample before and after the intervention are similar or significantly different.
Or in other words, whether the intervention has any significant effect or not.

Machine learning

Linear Regression



Question 1

Linear regression model is used with continuous target variable. Which dataset will the model perform better on? Are there any underlying assumptions for linear regression model?

Answer 1

There are certain assumptions for the linear regression model that we must check in our dataset before we build a model. Although the model would work on any data

a. Linear regression assumes in order to work well that the relationship between the dependent and independent variables is linear.
b. The residuals must be normally distributed.
c. The second assumption can is that all the variables must follow the pattern of multivariate normal distribution.
d. The independent variables must not be highly correlated. In other words, there must be little or no multicollinearity in the data.
e. Linear regression analysis also assumes that there is no auto-correlation in the data.
f. Lastly the data must follow the pattern of homoscedasticity, which means the standard deviation of residuals should be constant.

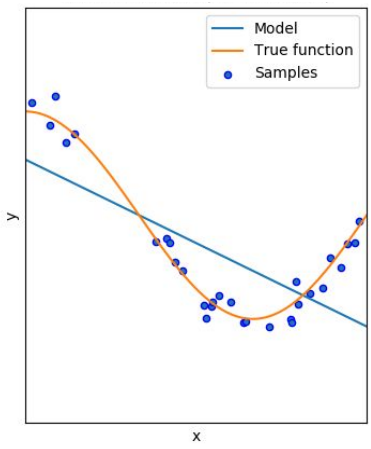
Know more: Assumptions in regression

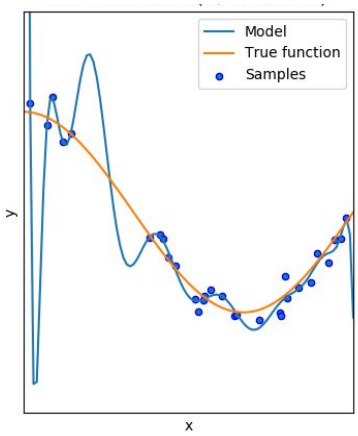
Question 2

Can linear regression learn the non-linear relationships in the dataset?

Answer 2

The linear regression model works well for data that has a linear relationship between the independent and target variables. For a data that has non linear relationship, the model would not give satisfactory results. Suppose we have a non linear data and we use a linear regression model, it might end up underfitting (as in the first image). Instead we can use polynomial regression (image two).





Source

Question 3

Polynomial regression has certain drawbacks, for instance, a high degree model tends to overfit. What are the options to overcome the problem?

Answer 3

Although polynomial regression performs better that linear regression in some cases but it tends to overfit. To avoid overfitting we can use regularization techniques such as ridge regression or lasso regression. In case of ridge regression, the magnitude of the

coefficients decreases, where the values reaches close to zero but not absolute zero. While using lasso, the magnitude of coefficients for certain independent variables is reduced to zero.
We can also use spline regression. In this technique, instead of building one model on the complete dataset, it is divided into multiple datasets and a model is build on each dataset.

(if you are new to spline regression, read this article)

Question 4

The linear regression model fits a line in order to model the relationship between independent and dependent variables. So how does the model decide the right line?

Answer 4

- As in the figure above, there could be multiple lines fit on the data points. The right line is decided such that the error value is minimum (as compared to other lines).
- Selecting the best fit line is equivalent to selecting the right set of coefficients for the line. This is an optimization problem.
- To select the best fit line, we use gradient descent optimization technique.
- Initially the coefficients for independent variables are randomly assigned and the error is calculated. These weights are then continuously updated to reduce the error.

(if you are not familiar with the concept of gradient descent algorithm, you must read this article: Introduction to Gradient Descent Algorithm)

Question 5

Suppose you have a dataset with over 200 features and you find that the independent features are highly correlated. How would you verify this and deal with the situation?

Answer 5

When the independent features are highly correlated, one of the assumptions of linear regression model is violated. We need to check for multicollinearity in the dataset. This can be done using the VIF method. Lower value of VIF value suggests no multicollinearity whereas a high value (>= 10) implies serious multicollinearity.

To deal with the highly correlated variables, we can use ridge or lasso regression. If we want the highly correlated features to be completely removed from the dataset, we can use the lasso regularization technique.

Question 6

Regularization techniques can be used for feature selection. Suppose the dataset has 40 features and we want to use only 20 features for building a linear regression model. How would we decide which regularization to you use- ridge or lasso? Is there any way you can use a combination of both? How can you do that?

Answer 6

Ridge and lasso are both regularization techniques. Lasso is used for feature selection since it reduces the coefficients of less important variables to zero. In case of ridge, the coefficients of certain variables become close to zero (but never equal to zero).

Another type of regularization called the elastic-Net Regression, which is a hybrid of ridge and lasso regression.

Question 7

Suppose we have N independent variables (X1,X2… Xn) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting

the best fit line using least square error on this data. You found that correlation coefficient for one of it's variable (Say X1) with Y is -0.95. So does that mean the variable X1 and Y is very weak?

Answer 7

Since the correlation is decided only by the magnitude only, and not the sign, we can conclude that the relation between the X1 and Y is strong.

Question 8

If you are given the two variables V1 and V2 such that

1. If V1 increases then V2 also increases
2. If V1 decreases then V2 behavior is unknown

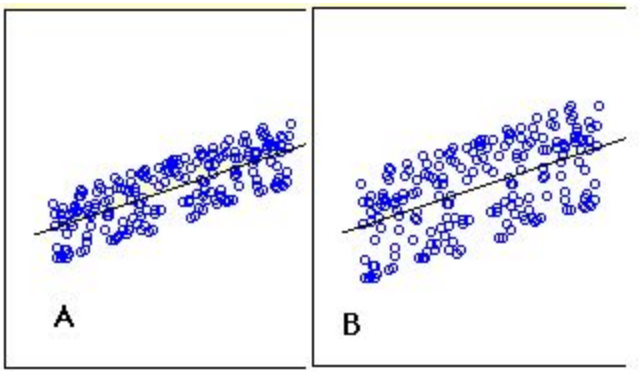
What is the value of pearson's correlation between V1 and V2?

Answer 8

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V1 as x and V2 as |x|. The correlation coefficient would not be close to 1 in such a case.

Question 9

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.



Answer 9

X axis is the independent variable and Y is the dependent variable. Considering the scale is same in both the case, the sum of residuals is higher in which case?
(Sum of residuals will be zero in both the cases)

Question 10

Suppose you have been given the following scenario for training and validation error for Linear Regression. (image below)

Scenario	Learning Rate	Number of iterations	Training Error	Validation Error
1	0.1	1000	100	110
2	0.2	600	90	105
3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

Which of the following scenario would give you the right hyper parameter?

Answer 10

The training and validation error is least in the second case and hence the right hyper parameters would be from parameter 2.

Question 11

Why do we only use ordinary least square during error calculation and not the absolute error.

Answer 11

Squared error penalises the larger error more drastically and therefore is helps optimisation to converge faster, also it useful in problems where error penalty is not linear.

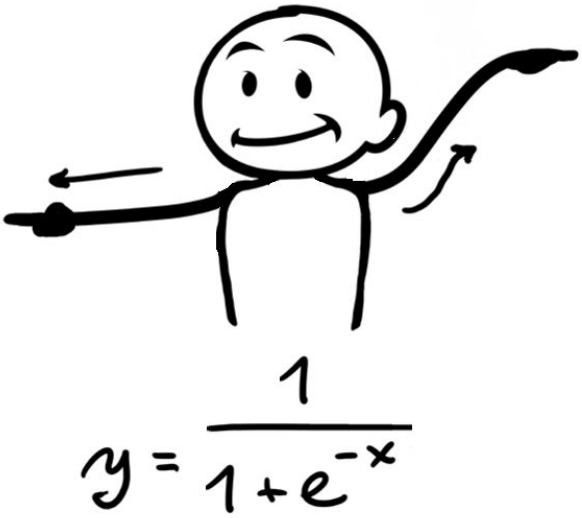
Question 12

What should we do if we want to apply linear regression on the non linear dataset?

Answer 12

Linear regression works well when the dataset has linear relationships, hence to use it on a non linear dataset, we need to use feature transformation to make non-linear data linear.

Logistic regression



Question 1

Suppose we have a classification problem where the target variables have 4 classes - A, B, C, D. How can we use logistic regression for this problem statement?

Answer 1

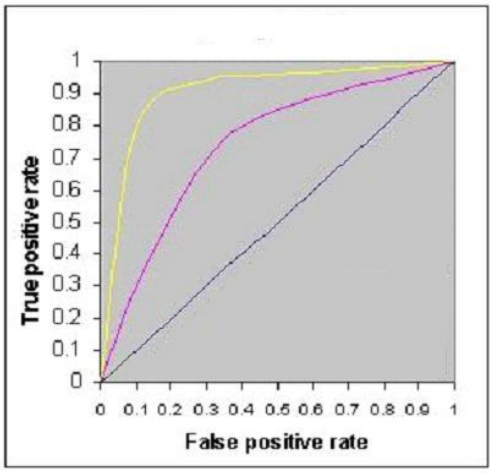
The logistic regression can be used for multiclass classification by implementing the “one-versus-rest” technique where we take up each class and try to classify whether the data points belong to that particular class or not.

Suppose we have four classes A, B, C and D. We can start with classifying the data points as A vs B/C/D (not A), and then for class B vs not B, so on.

Model->	Mode l1	Model 2	Model 3	Model 4
	(A vs B ,C ,D) Calculating probability of outcome A	(B vs A, C, D) Calculating probability of outcome B	(C vs A, B, D) Calculating probability of outcome C	(D vs A, B, C) Calculating probability of outcome D

Question 2

The below figure shows AUC-ROC curves for three logistic regression models. Different colors show curves for different hyper parameters values. Which of the following AUC-ROC will give best result?



Answer 2

ROC is a probability curve and AUC represents degree or measure of separability. Higher the AUC (area under curve), better the model.

From the graph show, we see that the yellow curve has the highest AUC and hence would give the best results.

Question 3

Can you suggest me right evaluation metric basis below problem statement?

a. Need to predict the credit card fraud transaction and as you know fraud transaction is a rare event.
b. Estimate the price of products and here price distribution of products are skewed.
c. Trying to build a recommendation engine for e-commerce company, here our task is to predict the next N products customer will buy.

Answer 3

In the first case, we can expect that the dataset provided would be imbalanced, as it is mentioned that fraud transaction is a rare event. We can use F1 score (confusion matrix) in this case.

B: Since the price distribution of products id skewed, RMSE would not be the right metric to evaluate the model. Instead, RMLSE can be used.

C: The third case is a recommendation problem. A recommender system produces an ordered list of recommendations for users. MAP@k (mean precision recall) is used as evaluation metric for recommendation problems.

Question 4

Suppose I applied a logistic regression model on data and got training accuracy X and testing accuracy Y. Now I add 2 new features in data. Select option(s) which are correct in such case.

Note: Consider remaining parameters are same.

1. Training accuracy always decreases.
2. Training accuracy always increases or remain same.
3. Testing accuracy always decreases
4. Testing accuracy always increases or remain same

Answer 4

Training accuracy will increase but testing accuracy would increases only when the new variables are found to be significant.

Question 5

Two logistic regression models are giving the same accuracy and F1 score, how will you decide the better of the two?

Answer 5

When two logistic regression models are giving the same results in performance, they can be judged on the basis of how well they are separating the classes, this is primarily done by using the log-loss function.

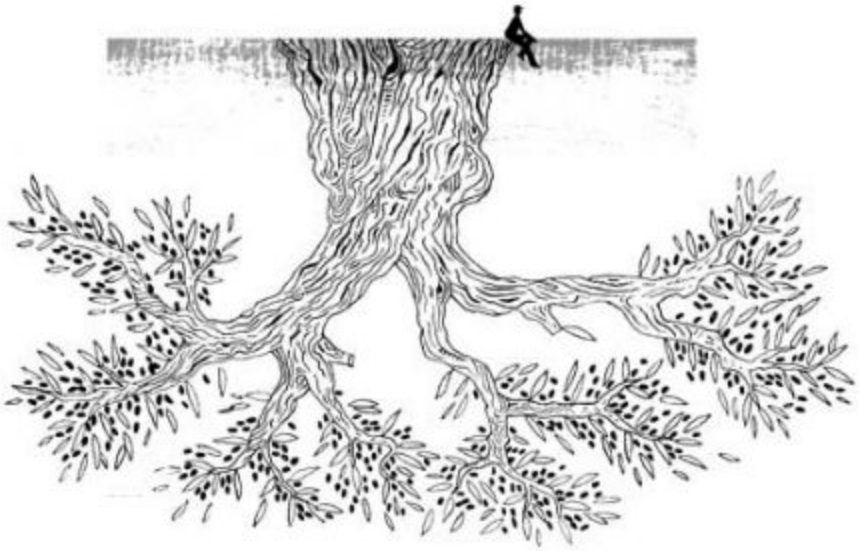
Question 6

Why is logistic regression is called “regression” , where in practicality it is a classification algorithm.

Answer 6

In logistic regression algorithm, we are predicting the probability whether a point belongs to to a certain class or not.
We know that probability lies in the range 0 to 1, therefore it is a regression process. We have to explicitly define the threshold in to be deterministic whether a point is in certain class or not.

Decision Tree



Question 1

For a decision tree algorithm, how will the splitting criteria differ in case of a regression model and a classification model?

Answer 1

The splitting is done using the independent variables and does not change if the target variable is continuous or categorical. The complete data is split into homogeneous groups, such that similar point are in one group. In order to select the best split point, various measures such as gini index, information gain and entropy are used.

Based on the input variables, a split is made on each possible point (for each variable and value). The gini index (or entropy/ information gain) is calculated for each split. Comparing these values, the best split point is selected.

Question 2

Suppose we have 2000 data points and 20 features. What is the maximum number of leaf nodes possible with a decision tree? Also, can you calculate the number of levels to which the tree will grow (depth of tree).

Answer 2

The maximum number of leaf nodes will be equal to the number of datapoints in the dataset (such that each leaf node has only one sample/datapoint). In this case the maximum number of leaf nodes will be 20,000.

Question 3

If we let the tree grow to the maximum length, there are chances of overfitting. How will you make sure that your decision tree model does not overfit?

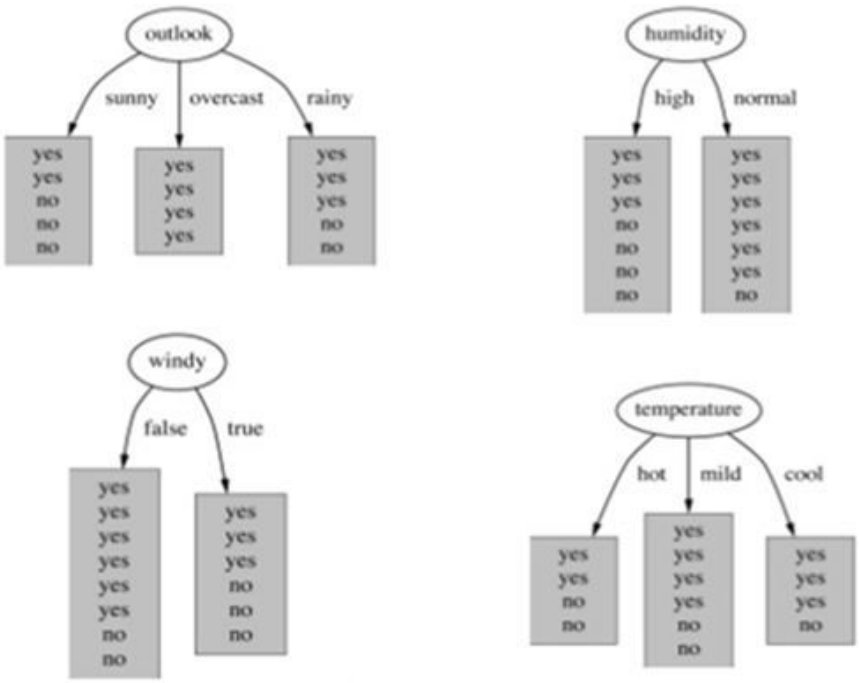
Answer 3

If we let the tree grow to the maximum length, there are high chances that the model overfits on the training data. So we should reduce the size of the tree, which is called tree pruning. To do so, we can set certain conditions or constraints on the various parameters of decision tree model - such as max depth, min leaf sample size etc.

Consider the example of max depth, if we have a total of 500 data points, then the tree grows till each leaf node has only one sample. Instead, if we set a constraint on the parameter max_deapt = 3, the tree would only grow for three levels.

Question 4

Suppose you are building decision tree model, which split a node on the attribute, that has highest information gain? In the below image, select the attribute which has the highest information gain.



Answer 4

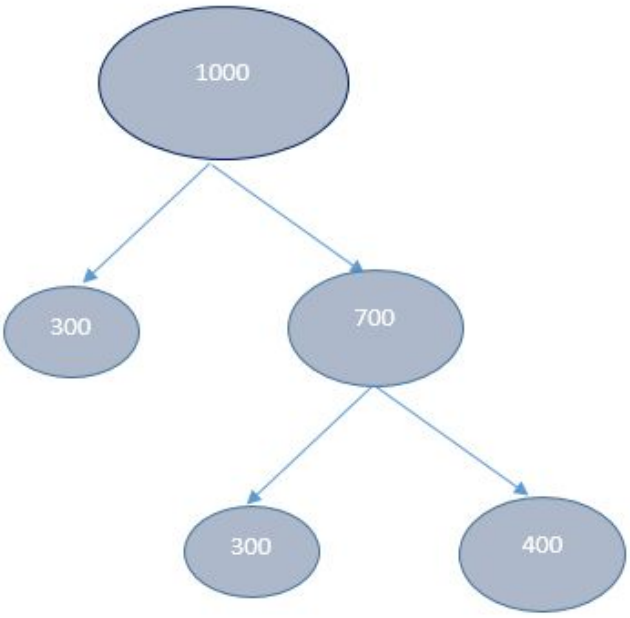
The split point in a decision tree is selected such that similar points are in the same group. While splitting on the variable *outlook,* we see that the most homogeneous groups are created. Hence the best attribute to split would be 'outlook'

Question 5

Given 1000 observations, Minimum observation required to split a node equals to 200 and minimum leaf size equals to 300 then what could be the maximum depth of a decision tree?

Answer 5

The tree split has two constraints, the minimum number of samples in the node must be greater than 200. Also, the split will happen only when the leaf nodes created after the split has at least 300 samples each. The answer will be max depth =2. The tree with 1000 samples would look like this:

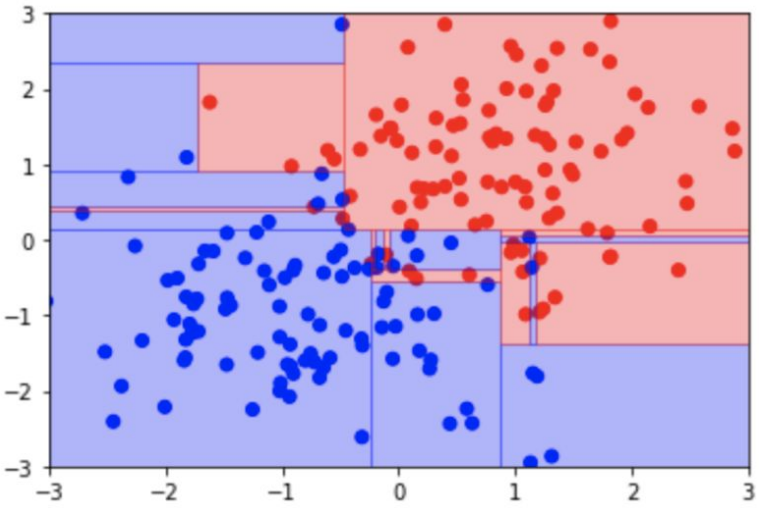


Question 6

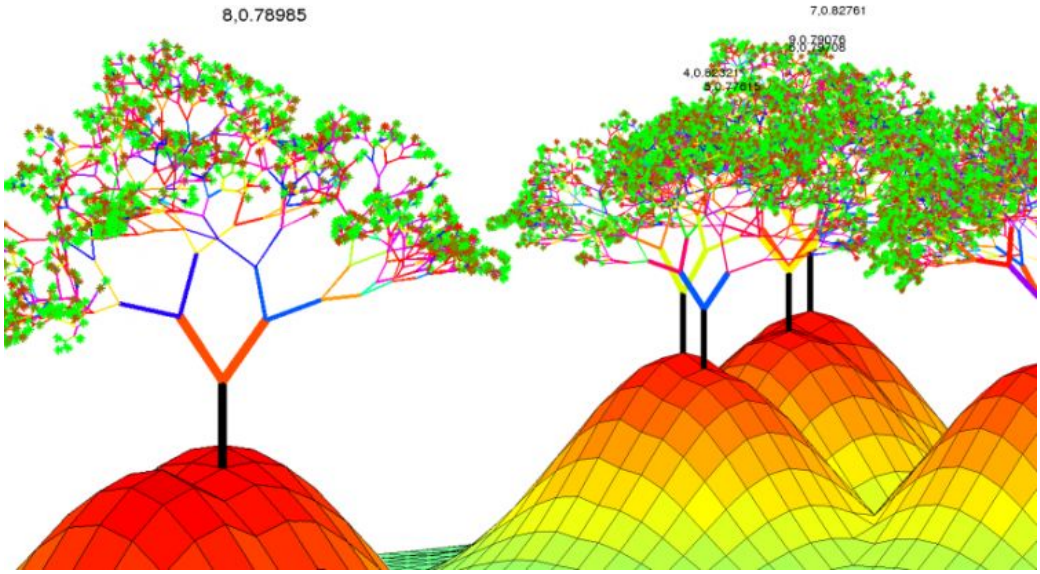
What is the nature of the decision boundary of a decision tree, explain?

Answer 6

At every level in Decision Tree, the splits occur within a single variable or the feature of the input data, therefore when the decision boundaries are plotted, they appear to be in combination of vertical and horizontal lines.



Random Forest



Question 1

A random forest model is an ensemble of multiple decision trees. Consider the following two scenarios -

Case 1: Build a random forest with 10 estimator (i.e. 10 decision trees).
Case 2: Creating 10 decision trees and averaging the predictions from these trees to calculate the final predictions.

Will both the models give the same results? (considering the dataset and other parameters are same in both the cases)

Answer 1

When training a decision tree classifier, splits are done using all of the data points and all of the features whereas in case of random forests every estimator(decision tree) is based only on a random subset of the dataset(thus the name random forest).

Hence, in the case of multiple decision trees trained on the same dataset , all the trees will be same along with their predictions for a given data point whereas because of change in distribution of data for different decision trees in case of random forests, the predictions can be different for the same data point from different estimators.Thus Case 1 and Case 2 will have different results.

Question 2

Increasing the value of n_estimator improves the accuracy of the model? How can one decide the optimal value of estimator?

Answer 2

Increasing the number of estimators for a random forest will always improve the generalization capacity of the model (as long as the number of estimators is not comparable to the number of datapoints), thus we can use as high a number as can be handled by the hardware of the system.

For finding the optimal number of estimators for a random forest , we can try grid search i.e. building the model with different number of estimators and then checking the generalization capacity. Also, it is general practice to initialize the number of estimators with the square root of number of data points for a random forest model.

Question 3

Two most important ensemble learning methods are bagging and boosting. Can you explain the difference between bagging and boosting?

Answer 3

Bagging (or Bootstrap Aggregating): In this technique, multiple subsets of the complete data are created and a model is trained on each of these subsets. The final prediction is made by combining the values predicted from each of these models trained on different subsets. Various algorithms which use bagging technique are Bagging estimator, Random Forest, Extra trees.

Boosting: It is a sequential process where each subsequent model attempts to correct the errors from the previous model. This is done by giving higher weights to the observations which were incorrectly predicted. Final model (strong learner) is the weighted mean of all the models (weak learners). AdaBoost GBM, XGBoost, etc are some of the algorithms which use boosting technique.

Question 4

Random Forest have multiple decision trees and at the end we combine results of all the decision trees using different methods like average, voting

etc. And, to bring randomness/ diversity, we do take bootstrapped samples and random subset of features. Selection of random subset of features can be done in two ways:

a. Select random subset of features for each tree and based on these features individual tree will grow.
b. Every tree has access to all features and at the split point, we are selecting random subset of features and split is happening based in this random subset.

Which of the above mentioned methods can be used in case of a random forest model?

Answer 4

Both the methods can be used. In the first case, the random selection of features is done for each tree. So at each split, a feature from this random subset is selected.

In the second scenario, we are using all the features for each tree. A random selection of features is done for each split.

Question 5

While doing cross validation, we realised that Random Forest model is overfitting. Which of the following steps might help you to overcome overfitting problem?

c. Increase the value of minimum sample size of the leaf nodes
d. Increase Maximum Depth of the trees
e. Decrease the value of minimum number of samples required to split an internal node

Answer 5

In order to avoid overfitting for a random forest model, we will need to prune the tree. To do so, we will have to tune all the parameters mentioned above.

A. Leaf is the end node of a decision tree. A smaller leaf makes the model more prone to capturing noise in train data.

B. Altering the max depth of trees can be used for tree pruning. This controls over-fitting since higher depth allows model to learn relations very specific to a particular sample. Hence the max depth must be decreased.

C. The value for min_samples_split should be increased to control overfitting. When the min sample split is 2, the tree is allowed to grow till each leaf node has only one sample, which increases the chances of overfitting.

Question 6

Neural Networks are universal approximators. Having said that, can a neural network exactly replicate the decision boundary of a random forest. Explain.

Answer 6

Random forest is an ensemble of multiple decision trees, each using a random subset of data points and features. Hence a neural network would not be able to exactly replicate the boundary of a random forest model.

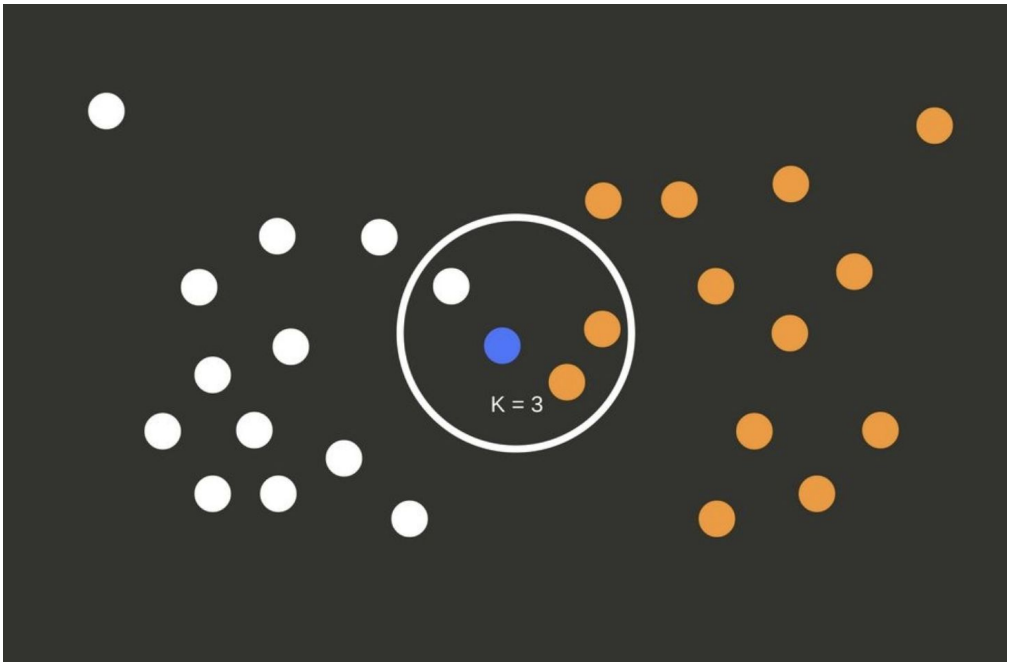
Question 7

Is there any relation between the Random forest and Bagging Algorithms?

Answer 7

Random forest is a special case of Bagging algorithm when the base estimator is Decision tree.

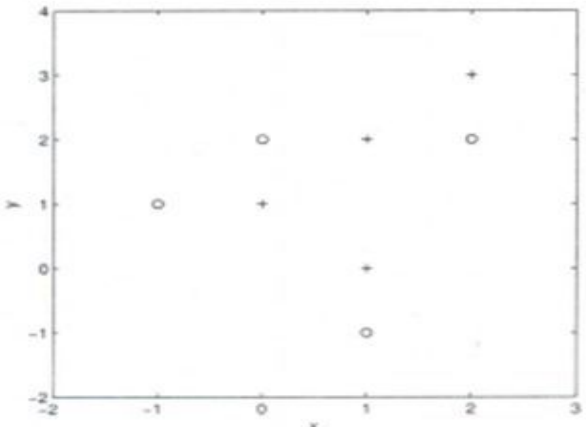
Knn (k- nearest neighbour)



Question 1

You have given the following data where x and y are the 2 input variables and Class is the dependent variable.

x	y	Class
−1	1	−
0	1	+
0	2	−
1	−1	−
1	0	+
1	2	+
2	2	−
2	3	+



Here, the scatter plot shows the above tabular data in 2D space. Suppose, you want to predict the class of new data point x=1 and y=1 using euclidean Dictionary distance in 3-NN. In which class this data point belong to?

1. + Class
2. – Class
3. Can't say

Answer 1

For KNN classifier , we will have to find the distance of all the data points from the test point, the closest 3 points have distance of 1,1,1 and classes +,+,+. Thus the predicted class for the test point is +.

Question 2

The dataset we are given for a problem consists of multiple features , one of them is the customer age ranging from 18-60 ,another feature is the salary/annual income of the individual.We are trying to use KNN algorithm on the data but the results are not satisfactory.Where is the problem?

Answer 2

The problems lies in the fact that the features are from different scales i.e. we need to apply normalization on our dataset. This will ensure that there is equal contribution from all features of the dataset. Here the age feature is constrained to 18-60 but the salary feature can have very large values. Normalization is a very important data preprocessing step for machine learning algorithms as it converts all features to the same scale while maintaining the distribution of the data.

Question 3

We are given a dataset containing 2000 data points, now we want to use knn algorithm for classification , what values of K will you try(1/200)?

Answer 3

When we consider a very large value of K , the knn algorithm under fits. Consider the case when k is infinity , the model is the simplest i.e. it always predicts the same class - the majority class thus leading to low variance and high bias whereas when the value of k is very small the model is very complex i.e. it has high variance and low bias.

For a very low value of k, suppose k = 1, the model overfits, since it looks at only one nearest neighbour. Thus one must select optimum value of k to build a good model.

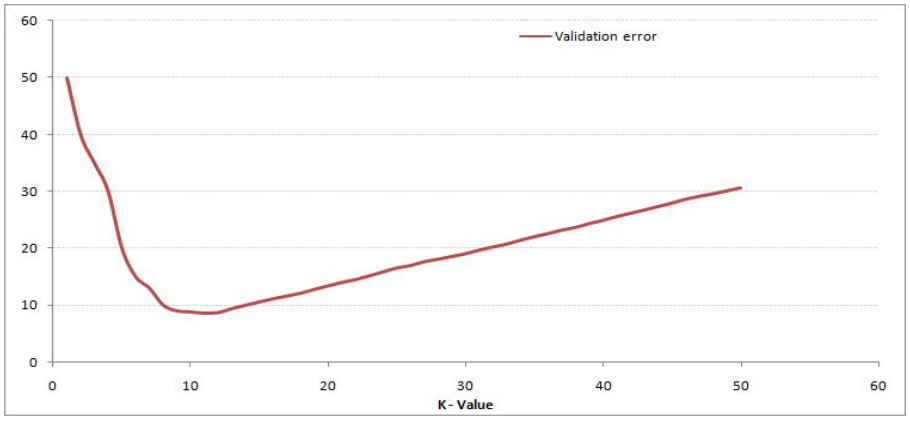
(the elbow curve method can be used).

Question 4

What does the K in KNN mean? How do you find the optimal value for K?

Answer 4

K is a hyperparameter for the KNN algorithm which decides the number of training data points that are used for predicting the label for the data point during the testing phase. For finding the optimal value of k , we can try all possible values of K from 3 upto a specific value and monitor the loss for the validation data. Another possible solution is the elbow technique where we plot the total sum of squared error for different values of k and find the point where the total sum of squares starts increasing.



Here k=8 seems like the optimal value for our dataset.

Question 5

Why is KNN called a lazy learning algorithm?

Answer 5

KNN is called a lazy learning algorithm because it apparently has no process of learning, all the magic happens during the prediction time based on the neighbors of the pont to be predicted.

Question 6

Why KNN is not used in industry?

Answer 6

KNN apparently has no learning process and therefore it requires all the data to infer the class of a new point, because of which, all the data has to be store in the primary memory which is not feasible when the data is very large.

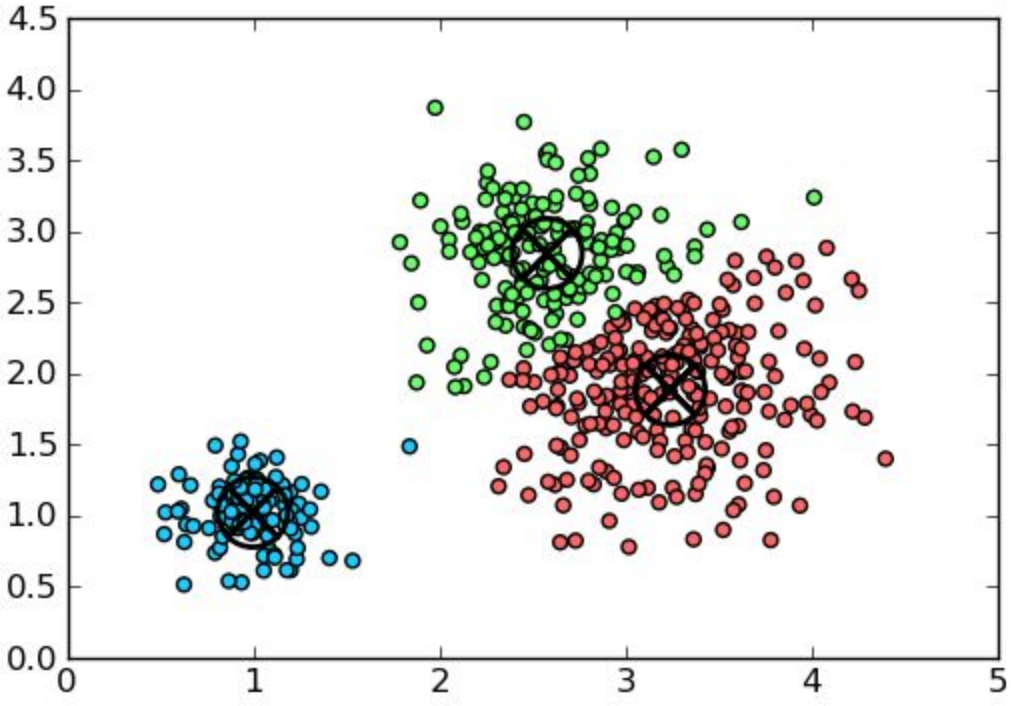
Question 7

List the differences between KNN and K-means Algorithm?

Answer 7

KNN is a Supervised algorithm whereas K-means is the unsupervised learning algorithm.KNN classifies points into predefined classes, whereas K-means segregates data into clusters which are not predefined.

K-means



Question 1

While doing K means algorithm, consider the following two cases-
Case1: The cluster centers are updated after every datapoint assignment
Case2: The cluster centers are updated after assigning every data point to some cluster.

We have applied k means on our dataset to calculate the cluster centers, now one feature of the dataset was scaled by some factor M, do the cluster centers change? What if all the features were scaled?

Answer 1

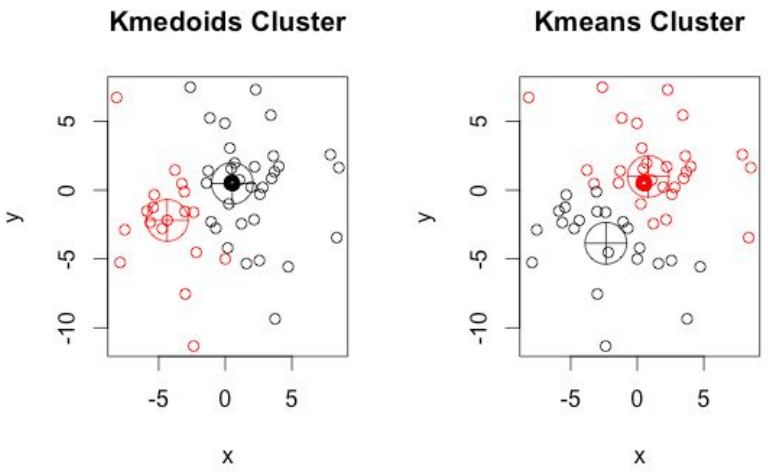
In case only one feature was scaled , the cluster centers will need to be updated as it will increase/decrease the value of single feature that will lead to the change of cluster centers thus there might be some points whose clusters will get changed.But in case all the features were scaled , the cluster centers will also get scaled by the same factor but in this case the clusters will not change.

Question 2

Is there any other statistic than the cluster mean that can be used to calculate the cluster centers? The K medoids algorithm

Answer 2

In case of the K medoids algorithm, the medoid is used as the cluster center instead of the mean. Also, the absolute difference between the cluster center and data point is used instead of euclidean distance.Thus, the k medoids algorithm is more robust to noise and outliers present in the data.



Question 3

Does the k means algorithm have any assumptions or constraints for the dataset? If yes, what are the assumptions for this clustering algorithm?

Answer 3

K means clustering algorithm have two assumptions

1. The clusters are spherical in shape
   K means algorithm tries to find out spherical clusters of data points around its cluster centers, if this assumptions is not held we can convert the data into polar coordinates for the algorithm to work properly
2. The cluster are approximately same size
   K means tries to minimize the within cluster sum of squares

Question 4

Is k-means algorithm prone to outliers?

Answer 4

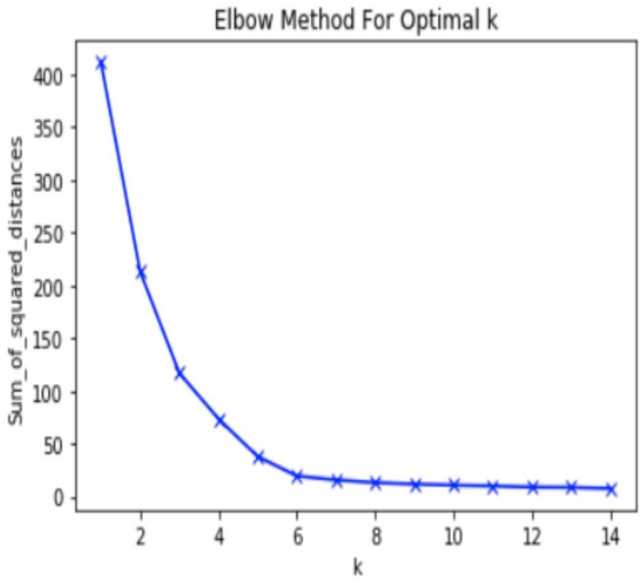
K means is a clustering algorithm, and the presence of outliers in the dataset (if not scaled), can shift the centroid value. Thus these centroid values may not be the as representative as they otherwise would be.

Question 5

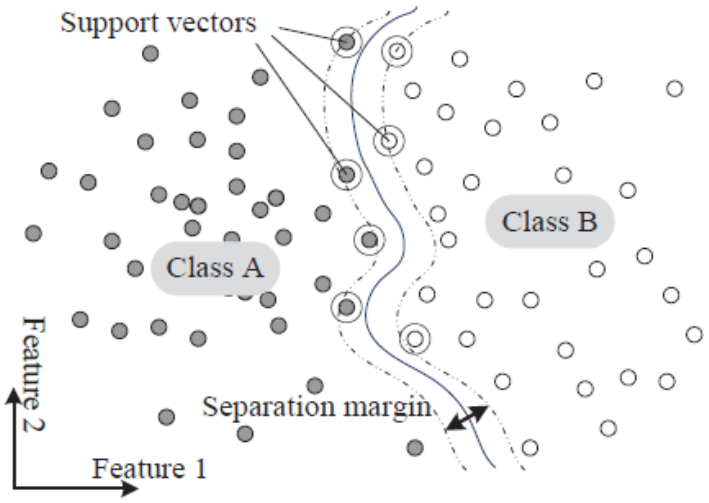
How can we decide the value of K in K-means?

Answer 5

We use the Elbow method in order to find the ideal number for k.

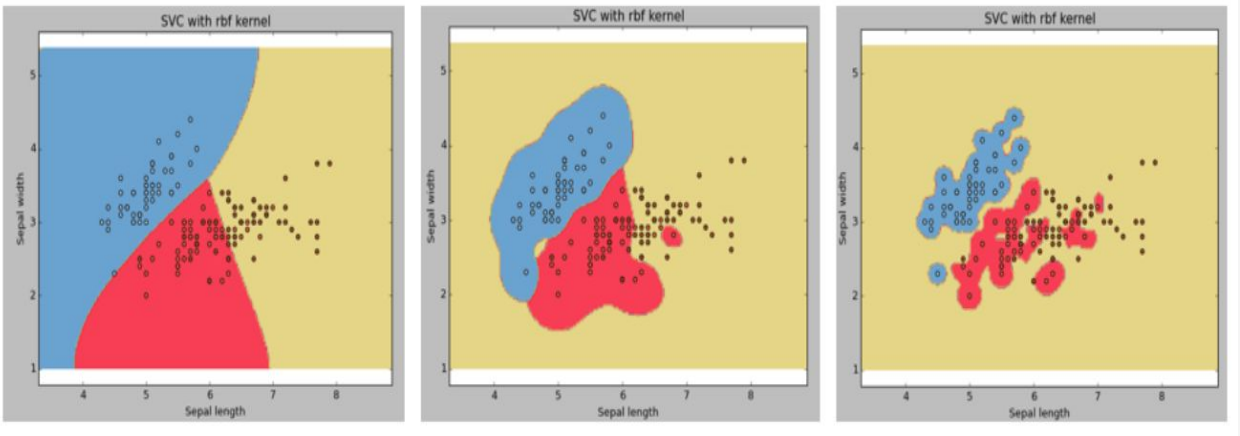


Support Vector Machine



Question 1

Have built a support vector machine with three different value of kernel co-efficient also known as gamma. Can you please arrange below three images in ascending order of gamma values?

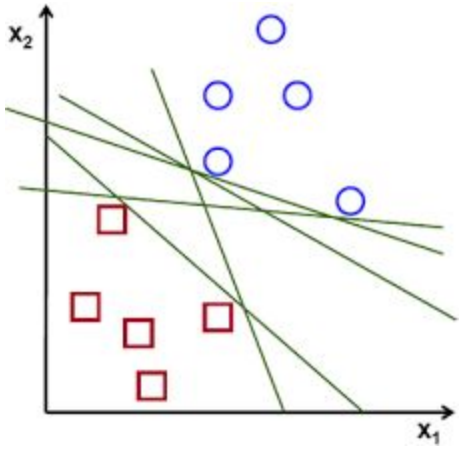


Answer 1

Penalty parameter C of the error term. It also controls the trade-off between smooth decision boundary and classifying the training points correctly. For large values of C, the optimization will choose a smaller-margin hyperplane. Read more here.

Question 2

Suppose we were given a dataset that consists of all numerical features, we applied linear regression but the results were not satisfactory. Further we also used SVM but still the performance did not improve , what seems to be the problem? Find out the best decision boundary from the ones given below

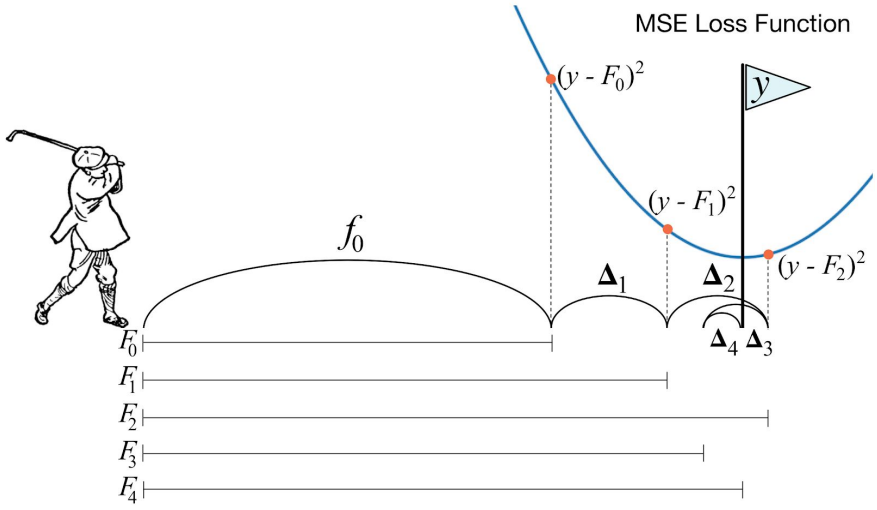


Answer 2

If both linear regression and SVM are not giving satisfactory results on a dataset, this implies that the classes are not linearly separable. In such a situation we can use the kernel trick of SVMs to project the dataset into higher number of dimensions such that the classes become linearly separable.

The best decision boundary is the one that maximizes the margin on both sides.(SVM theory)

GBM/ xgb



Question 1

For XGBoost, how are L1, L2 regularizations different from Gamma?

Answer 1

The value of gamma defines the minimum loss reduction required for making a further split on a leaf node. The range of gamma values is from [0,∞). This is different from L1 and L2 regularization which are used to penalize weights for features in the dataset.

Question 2

Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?

Answer 2

The major difference between random forest and gradient boosted trees lies in the fact that the individual estimators in random forests are independent of each other whereas in case of gradient boosted trees the individual estimators are not independent. The estimators in random forests are built parallely whereas in GBTs the estimators are built incrementally. Also the data points for which the current estimator was not able to perform well are given more preference in the next estimators so that the whole model is able to generalize well over the complete dataset.

Question 3

Are the trees used in gradient boosting independent of each other?How is the dependence established between the weak learners?

Answer 3

The trees in gradient boosting algorithm are built in a sequential order, each individual tree trying to correct the errors of the previous trees. This is done by giving higher weightage to the points that were wrongly predicted by the previous tree. So the individual trees (or weak learners) in GBM are dependent on each other.

Question 4

Which of the following are mandatory data pre-processing step for XGB?

1. Impute Missing Values
2. Remove Outliers
3. Convert data to numeric array / sparse matrix
4. Input variable must have normal distribution
5. Select the sample of records for each tree/ estimators

Answer 4

Converting data to numeric array/sparse matrix is a mandatory pre-processing step for XGBOOST as xgboost cannot handle categorical features.

Question 5

I keep on increasing the number of estimators (say to 10000 trees) for the same train set in case of random forests and GBM. I check its performance on a validation set. Which of these do you think will overfit more & why?

Answer 5

Gradient boosted machines will have a higher chance of overfitting than random forests. In gradient boosting algorithm, each tree is dependent on the previous tree, trying to correct the error. So as the number of trees increase, the model is likely to overfit. On the other hand, for a random forest model, each individual tree in independent and uses a different subset of data points and features.

Naive bayes

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

Question 1

The features of our dataset show strong correlation , is Naive Bayes a good choice for such a dataset? How will you modify the dataset such that Naive Bayes shows good results?

Answer 1

Strong correlation between features implies dependence. Thus Naive Bayes is not a good choice for such a dataset because Naive Bayes assumes that the features of the given dataset are independent of each other. In such a case we can apply PCA to identify the independent dimensions from the dataset and then use Naive Bayes.

Question 2

What does the word ‘naive’ in naive Bayes algorithms suggest?

Answer 2

Naive Bayes assumes that all of the features in a data set are equally important and independent. As we know, these assumption are rarely true in real world scenario.

Question 3

In what scenario is Naive Bayes suitable to use and when it it not suitable to use.?

Answer 3

Naive Bayes is suitable for the datasets where the independent variables are categorical and also are truly independent from each other, it also works well with text data.

Question 4

Can I use Naive Bayes for continuous variable?

Answer 4

It is not recommended to use Naive Bayes for the continuous variables, But other variants of Naive Bayes like: Gaussian Naive Bayes can be used instead.

Question 5

What is the working principle behind the naive bayes.?

Answer 5

Naive Bayes in based on the Backward Probability Reverse probability. It follows a probabilistic approach to model the data.

Data Exploration and Feature Engineering

Question 1

The categorical variables can further be classified as cardinal and ordinal. Can you explain the difference between them and how can we deal with these variables?

Answer 1

The variables which have a certain number of categories fall under the term ‘Cardinal variables’. These variables do not have an order between them. For example, the variable ‘Gender’ can have two categories - Male and Female, and there is no order to these categories. Most machine learning models cannot deal with categorical variables. So we use one hot encoding technique to convert these categorical variables into binary.

While Ordinal variables have categories which contain an order between them, such as ‘class of a student’ which can be I, II,....IX, X, and so on. We know that IV and V are higher than I and II. So we can say that these categories have an order. Although one hot encoding technique can be applied on these variables, it is preferred to use label encoding (since the order in the variables is important)

Question 2

Can you explain the difference between multiclass and multilabel problem using a simple example.

Answer 2

In a multiclass problem, each row can belong to only one class (hence have one label), whereas in a multilabel problem, each row may belong to more than one classes (have multiple labels).

For example, if we are classifying a movie based on the language it can be either english, hindi, tamil, or marathi etc. A single movie cannot be in both tamil and marathi at the same time. This is a multiclass classification problem.

On the other hand if we are trying to classify a movie based on the genre such as comedy, action, romance etc. A movie can belong to two different genres at the same time. This classification problem would be of a multilabel classification.

Question 3

While exploring the distribution of the target variable, you find out that the dataset in imbalanced, how will you deal with it?

Answer 3

There are several ways to handle class imbalance

1. Use appropriate evaluation metrics
2. **Resample the dataset**: If the dataset is of considerable size we , can randomly select training data points from the majority class such that the number of data points in majority and minority class become similar.
   In case we do not have enough data , we can generate data points for the minority class using repetition,bootstrapping or SMOTE.
3. **Cluster the majority class**: Instead of random sampling we can cluster the data points of the majority class into clusters equal to the number of data points of the minority class.
4. **Use boosting**: Boosting models can perform better for dataset with class imbalance because of their inherent ability to learn better estimators for the data points that are difficult to learn.

[How to handle Imbalanced Classification Problems in machine learning?]

Question 4

What is dimensionality reduction?

Answer 4

Dimensionality reduction is a technique used to reduce the dimensions of the input data which can cause algorithms to run more efficiently and

Question 5

How do we fill a missing value in dataset in an “intelligent” way?

Answer 5

Straight-forward and quick way is to impute all the values using median/mode.
The intelligent way would be to impute the values based on the other columns having a high correlation with the column to be imputed.

Question 6

During feature selection - we remove features thus giving less information to the model for training. Will this affect performance of the model?

Answer 6

Feature selection is selecting the most relevant attributes for the predictive model. Only the redundant features are removed from the model so that it does not adversely affect the accuracy of the model. Having fewer but relevant features reduces the model complexity and training time.

Question 7

There are broadly two common methods to convert categorical variable to number, Label Encoding and One Hot Encoding. How do you decide which method to use when?

Answer 7

Most machine learning models cannot deal with categorical variables and hence we need to perform one hot encoding or label encoding before feeding the data to the model.
The variables which have a certain number of categories fall under the term ‘Categorical variables’. These variables do not have an order between them. For example, the variable ‘Gender’ can have two categories - Male and Female, and there is no order to these categories. In this case, we can use the one hot encoding technique to convert the categorical columns into two columns with binary values.

While some variables have categories which contain an order between them, such as ‘class of a student’ which can be I, II,....IX, X, and so on. We know that IV and V are higher than I and II. So we can say that these categories have an order. Although one hot encoding technique can be applied on these variables, it is preferred to use label encoding (since the order in the variables is important)

Question 8

Is rotation necessary in PCA? If yes, Why? What will happen if you don’t rotate the components?

Answer 8

Yes, rotation (orthogonal) is necessary because it maximizes the difference between variance captured by the component. This makes the components easier to interpret. Not to forget, that’s the motive of doing PCA where, we aim to select fewer components (than features) which can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn’t change, it only changes the actual coordinates of the points.

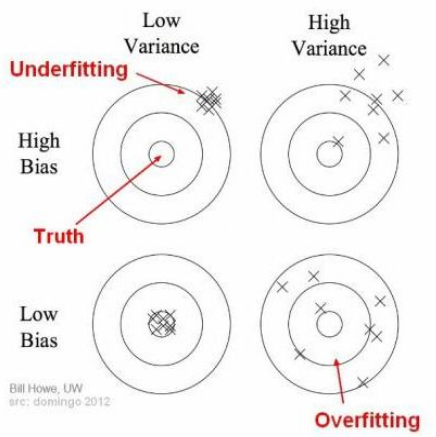
If we don’t rotate the components, the effect of PCA will diminish and we’ll have to select more number of components to explain variance in the data set.

Question 9

Have built two models “Model 1” and “Model 2”. "Model 1" has high bias and low variance where as “Model 2” has low bias and high variance. Which of these two models have higher chance of overfitting or under-fitting?

Answer 9

A model with high bias and low variance tends to overfit on certain feature or data points. The model will not be generalized and performs poorly, thus has a higher chance of underfitting. On the other hand, a model with high variance and low bias will have a higher chance of overfitting. A simple diagram (shown below) can be used to verify the above statements.



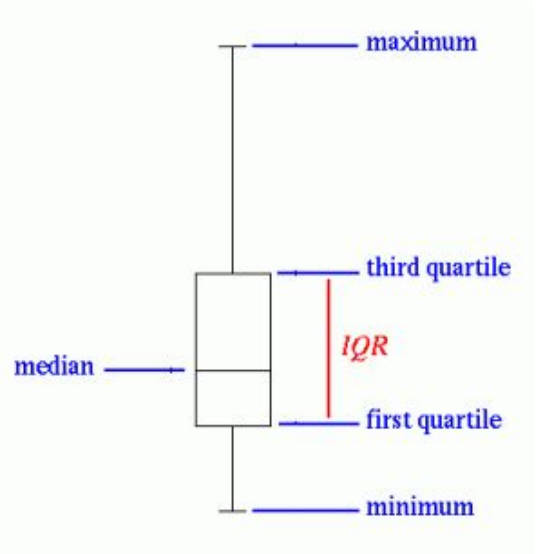
Question 10

During the data exploration process, one needs to deal with the outliers present in the data. How will you find out if your dataset has outliers? Which plot would be useful in this case?

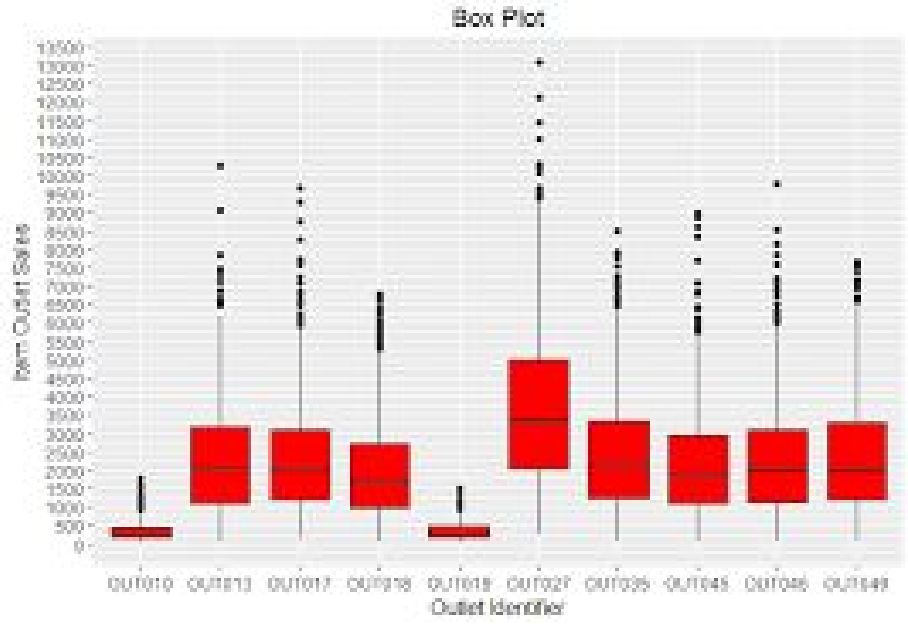
Answer 10

We can use a histogram or a box plot to understand the data distribution. Generally a box plot is used to identify the outliers in the data. A box plot is a graphical method of displaying variation in a set of data.

- Minimum – The minimum value in the data set
- First quartile – The value below which the lower 25% of the data exist
- Median – the value below which the lower 50% of the data exist
- Third quartile – the value below which the lower 75% of the data exist
- Maximum – the maximum value in the data set



Here is an example of box plot. The x-axis contains different product ID and y axis contains the outlet sales values.



Question 11

Your task is to predict delay for a given train at a particular junction. Discuss the features/information that you would collect and use to build an ML model for doing that

Answer 11

Past data for each day which has features

- The time train starts
- Number of junctions the train covers
- Time it reaches the previous junction
- Past data showing number of times the train is late

Question 12

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. The aim is to predict the sales of each product at a particular store. The following variables have missing values. Suggest possibles ways to deal with these missing values in the dataset.

```
Item_Fat_Content                0
Item_Identifier                 0
Item_MRP                        0
Item_Outlet_Sales            5681
Item_Type                       0
Item_Visibility                 0
Item_Weight                  2439
Outlet_Establishment_Year       0
Outlet_Identifier               0
Outlet_Location_Type            0
Outlet_Size                  4016
Outlet_Type                     0
source                          0
dtype: int64
```

Note- ignore the missing values in item_outlet_sales (that is for the test data)

Answer 12

- For the variable *Item_weight*, following are the possible ways to fill the missing values:
    - Using mean of the complete column
    - Using median of the column (if column has outliers, median is preferred)
    - Groupby on item_type and calculate mean.

- For the variable *Outlet_size*, can be imputed using
    - Outlet location type
    - Outlet establishment year

Model selection, parameter tuning and cross validation

Question 1

What is the difference between parameters and hyperparameters?

Answer 1

Hyperparameters for a model are those which we can manually tune while parameters are those which the model learns during the training time.

For example the coefficients in linear regression is a parameter while the value of c or random state is a hyperparameter.

Question 2

There are multiple machine learning algorithms, such as linear regression, logistic regression, random forest, xgb etc. How will you select which model to implement?

Answer 2

The model to be used is often decided based on the dataset we need to work on.

- For a regression problem, we can choose linear regression, random forest regressor etc.
- Similarly for classification problem, the choice of models would be slightly different.
- For imbalanced dataset, boosting techniques work really well.
- In case the dataset has a very high number of categorical variables, we can use CatBoost.

- ➢ For very large datasets, light GBM has proved to work better than other algorithms.

Question 3

How will you evaluate the model performance before deployment? (Different types of validation techniques)

Answer 3

It is necessary to evaluate a model before making predictions on the final set to make sure that the model is not overfitting on the training data. We can split the complete data into training and validation sets and observe its performance on the validation set.

There are a number of ways for creating a validation set-

1. We can simply split the dataset into train and validation set, keeping the first 70 percent to train while the remaining 30 percent to test.
2. The same can also be done by splitting the dataset randomly into train and validation without maintaining the order (using train_test split).
3. Another commonly used validation technique is kfold cross validation, where the complete dataset is split into k folds (using k-1 to train and 1 to test the model).
4. We also have Stratified k-fold cross validation, where each of these folds have the same distribution.

For more detailed explanation, read the following blog: Improve Your Model Performance using Cross Validation (in Python and R)

Optimization, regularization, evaluation metric

Question 1

When does regularization becomes necessary in Machine Learning?

Answer 1

Regularization becomes necessary when the model begins to overfit / underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

Article: How to avoid Over-fitting using Regularization?

Question 2

After training a model on a particular set of data, the model gives and accuracy of 98% on the training data. I use this model to make predictions on the test data and the accuracy drops to 54%. What went wrong?

Answer 2

The model is overfitting on the training data. Overfitting is when the model is performing extremely well on the training data but does not show a similar performance on the test (or validation data).

One way to verify whether the model is overfitting is to check the model performance on a validation set before using it on the test set.

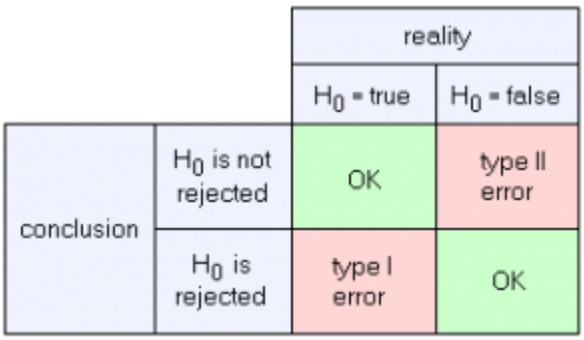
Question 3

What do you understand by Type I vs Type II error ?

Answer 3

Type I error is committed when the null hypothesis is true and we reject it, also known as a 'False Positive'. Type II error is committed when the null hypothesis is false and we accept it, also known as 'False Negative'.

In the context of confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).



Question 4

On adding new features to the dataset, there is a some change in the r-squared and adjusted r-squared values. Consider the following cases

1) r-squared and adjusted r-squared both increase
2) r-squared does not change but adjusted r-squared decrease
3) r-squared and adjusted r squared don't change

Which of the following cases is possible?

Answer 4

However, the problem with R-squared is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables. This is where "Adjusted R squared" comes to help. Adjusted R-squared

penalizes you for adding variables which do not improve your existing model.

In other words, each time you add a feature, R squared either increase or stays constant, but adjusted R squared may increase, stay constant or decrease. So all the above cases are possible.

Question 5

You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?

Answer 5

Low bias occurs when the model's predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results. In such situations, we can use bagging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling.

Question 6

When is it acceptable to have too many false positives or too many false negatives?

Answer 6

This depends on the given problem. Let's say the problem statement is to predict whether a particular patient has cancer or not. False negative would mean that a patient actually having cancer is not treated with the right medication. Thus a higher number of False negatives could be dangerous.

Instead, if we had a problem statement where we need to predict whether a mail is spam or not, if an important mail is actually classified as spam, we might miss out on some important information. Thus a false negative might not be very harmful but a high number of false positives could result in missing out on some important mails.

Time Series

Question 1

How can you differentiate or classify a problem statement as time series problem? For instance, consider the following two problem statements; which one would you apply a time series model on?

Problem 1- Based on - the income per month, date of issuing loan, and the loan type, we have to predict the loan amount for the person.

Loan ID	Date	Income per month	Loan type	Loan amount
ID207	15/07/18	25000	Car Loan	1000000
ID190	15/07/18	50000	Home Loan	2500000
ID007	22/07/18	70000	Personal Loan	1500000
ID433	29/07/18	45000	Education Loan	4500000
ID204	29/07/18	20000	Education Loan	5000000
ID611	08/08/18	80000	Business Loan	9000000
ID947	17/08/18	60000	Personal Loan	3700000
ID200	21/08/18	20000	Car Loan	500000
ID222	29/08/18	30000	Personal Loan	4300000

Problem 2- The target is to predict the hourly temperature using the for a particular region given the past data for the same.

Time	cloud cover	dew point	humidity	wind	Temperature
5:00 am	97%	51 °F	74%	8 mph SSE	59 °F
6:00 am	89%	51 °F	75%	8 mph SSE	59 °F
7:00 am	79%	51 °F	76%	7 mph SSE	58 °F
8:00 am	74%	51 °F	77%	7 mph S	58 °F
9:00 am	74%	51 °F	74%	7 mph S	60 °F
10:00 am	74%	52 °F	70%	8 mph S	62 °F
11:00 am	76%	52 °F	65%	8 mph SSW	64 °F
12:00 pm	80%	52 °F	60%	8 mph SSW	66 °F
1:00 pm	78%	52 °F	58%	10 mph SW	67 °F
2:00 pm	71%	52 °F	54%	10 mph SW	69 °F
3:00 pm	75%	52 °F	52%	11 mph SW	71 °F
4:00 pm	78%	52 °F	52%	11 mph SW	71 °F
5:00 pm	78%	52 °F	52%	12 mph SW	71 °F
6:00 pm	78%	52 °F	54%	11 mph SW	69 °F
7:00 pm	87%	53 °F	60%	12 mph SW	68 °F
8:00 pm	100%	54 °F	66%	11 mph SSW	65 °F
9:00 pm	100%	55 °F	72%	13 mph SSW	64 °F

Answer 1

The data points in a time series dataset are equally spaced and the target variable for each row is dependent on the past dependent and independent variables.

From the given two problems, the second problem represents a time series since the next hour's temperature depends on the previous value. While in the first problem, the loan amount for an individual person does not depend on the loan taken by the previous person.

Question 2

What are the different algorithms used for solving a time series forecasting problem?

Answer 2

Starting from the basic approaches used for forecasting time series, we can use :

- Naive method: In this forecasting technique, the value of the new data point is predicted to be equal to the previous data point.
- Simple averaging: The next value is taken as the average of all the previous values.
- Moving average: This is an improvement over the previous technique. Instead of taking the average of all the previous points, the average of 'n' previous points is taken to be the predicted value.
- Exponential smoothing: In this technique, larger weights are assigned to more recent observations than to observations from the distant past.
- Holt's linear trend method: This method takes into account the trend of the dataset. By trend, we mean the increasing or decreasing nature of the series.
- Holt's Winter seasonal method: This algorithm takes into account both the trend and the seasonality of the series.

[You can read about these methods in detail in the following article: 7 methods to perform Time Series forecasting (with Python codes)]

Some more complex forecasting techniques include:

- ARIMA/ SARIMA: ARIMA stands for Auto-Regressive Integrated Moving Average. It has three components - AR (autoregressive term), I (differencing term) and MA (moving average term).
- Prophet: Facebook's Prophet is an open source library used for time series forecasting. It tries to capture the seasonality in the past data.

Question 3

Consider the dataset for stock price given below.

	Year	Month	Week	Day	Dayofweek	Dayofyear	close_price
0	2013	10	41	8	1	281	155.80
1	2013	10	41	9	2	282	155.55
2	2013	10	41	10	3	283	160.15
3	2013	10	41	11	4	284	160.05
4	2013	10	42	14	0	287	159.45
5	2013	10	42	15	1	288	158.05
6	2013	10	42	17	3	290	162.00
7	2013	10	42	18	4	291	164.20
8	2013	10	43	21	0	294	159.60
9	2013	10	43	22	1	295	161.85

Using the “Date” column, we have extracted features like - week, month, year etc. Since the target variable is continuous, we can use a simple linear regression model to make predictions for next month. Then why do we need time series specific algorithms?

Answer 3

Regression algorithms can be used for forecasting the closing price in this case but they do not take into account the trend and seasonal pattern in the series, which are important components of the time series data.

Although a linear regression model can be used but a time series forecasting model would undoubtedly give better results.

Question 4

Suppose we have the air quality dataset for a particular region. We have the date and time columns and multiple features such as CO(GT), NO2(GT) etc.

Date	Time	CO(GT)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
10/03/2004	18.00.00	2	1360	150	11	1046	166	1056	113	1692	1268	13
10/03/2004	19.00.00	2	112	9	4	103	1174	92	1559	972	13	3
10/03/2004	20.00.00	2	1402	88	9	939	131	1140	114	1555	1074	11
10/03/2004	21.00.00	2	1376	80	9	948	172	1092	122	1584	1203	11
10/03/2004	22.00.00	1	1272	51	6	836	131	1205	116	1490	1110	11
10/03/2004	23.00.00	1	1197	38	4	750	89	1337	96	1393	949	11
11/03/2004	00.00.00	1	1185	31	3	690	62	1462	77	1333	733	11
11/03/2004	01.00.00	1	31	3	3	62	1453	76	1333	730	10	7
11/03/2004	02.00.00	0	1094	24	2	609	45	1579	60	1276	620	10
11/03/2004	03.00.00	0	1010	19	1	561	-200	1705	-200	1235	501	10
11/03/2004	04.00.00	-200	14	1	3	21	1818	34	1197	445	10	1
11/03/2004	05.00.00	0	1066	8	1	512	16	1918	28	1182	422	11
11/03/2004	06.00.00	0	1052	16	1	553	34	1738	48	1221	472	10
11/03/2004	07.00.00	1	1144	29	3	667	98	1490	82	1339	730	10
11/03/2004	08.00.00	2	64	8	0	174	1136	112	1517	1102	10	8

If we want to predict the NO2 content for the next month based on the above information using an ARIMA model, what are the changes that we need to make in this dataset?

Answer 4

ARIMA models work on the following assumptions –

- The data provided as input must be a univariate series.
- The data series is stationary, which means that the mean and variance should not vary with time.

We will first have to convert the above data into a univariate series. Since we want to predict the NO2 content, we will use that as our target variable. The Date and time columns must be combined to create a timestamp that'll be used as the index.

The next step is to make sure that the series is stationary before we apply an ARIMA model.

Question 5

Why do we need to check time series for stationarity? What would happen if we build a model without checking for this assumption?

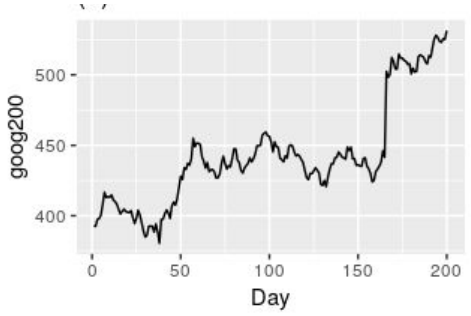
Answer 5

The idea behind making the series stationary is to have a constant mean and variance over time so that the forecast is not biased. For instance, if a series has an increasing trend, the model would be biased focussing on the increasing values alone. This will in turn affect the final predictions of the model.

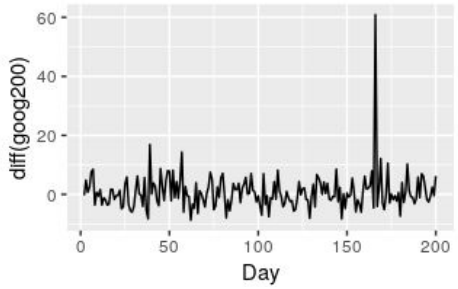
Question 6

Which of the following is are stationary series?

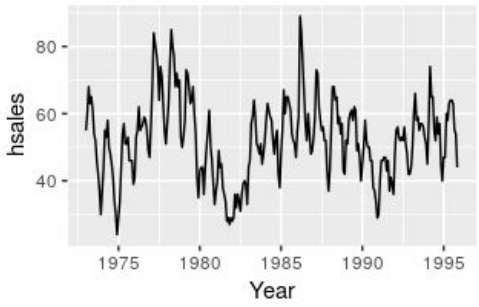
A-



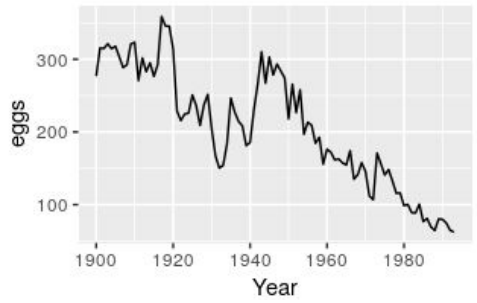
B-



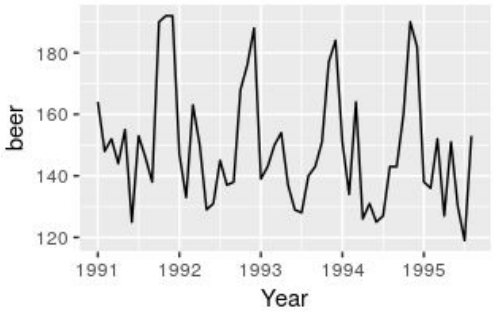
C-



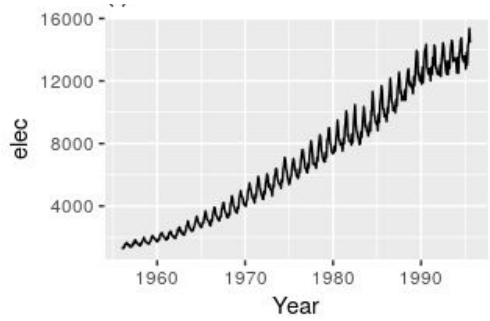
D-



E-



F-



Answer 6

To visually check whether the series is stationary, we need to verify two things -

- The series does not have an increasing or decreasing trend
- The series does not have a seasonal pattern.

The plots A and F have an increasing trend while D has a decreasing trend, so we can rule out A, D and F. Considering the seasonal pattern, we see that C, E and F have have a seasonality and are not stationary. So only B is a stationary series.

Other than visually looking at the plots, we can also use some statistical tests like to check the stationarity of the series.

Question 7

We cannot check the mean and variance at each point in the series to identify if the series is stationary or not. Is there any other way to identify a stationary series?

Answer 7

We can visually look at the plot to identify if the series has a trend or a seasonality. We also have some statistical tests to check is the series is stationary. The most commonly used statistical test is ADF (augmented dickey fuller) test.

The null hypothesis for ADF test is - time series has a unit root (implies non-stationarity). Based on the p-value, one can accept or reject the null hypothesis. Some other statistical tests to check stationarity are Kpss test, PP test etc.

(To study about the various statistical tests in detail, you can refer the following article : A Gentle Introduction to Handling a Non-Stationary Time Series in Python)

Question 8

What are methods to make a series stationary and how can we decide which ones to use when?

Answer 8

The condition for a stationary series is that the mean and variance should not change with time. In other words, we can say that there should be no trend or seasonality in the series. Following are the common methods to make series stationary -

Differencing: To stabilize the mean of the series, we can take lag differences. Each value in the series is subtracted from the previous value.

Seasonal Differencing: When there is a seasonal pattern in the series, differencing with the nth lag would be more appropriate and has proved to give better results.

Transformation: We use some common transformations like log, square root and cube root to deal with high variance in the series.

Often a combination of both is also used to make the series stationary, for instance differencing the series first and taking a log of the series.

Question 9

Once the input is univariate and series is stationary, we satisfy the conditions for ARIMA model. What are the hyperparameters for ARIMA and how can we decide the values for these parameters?

Answer 9

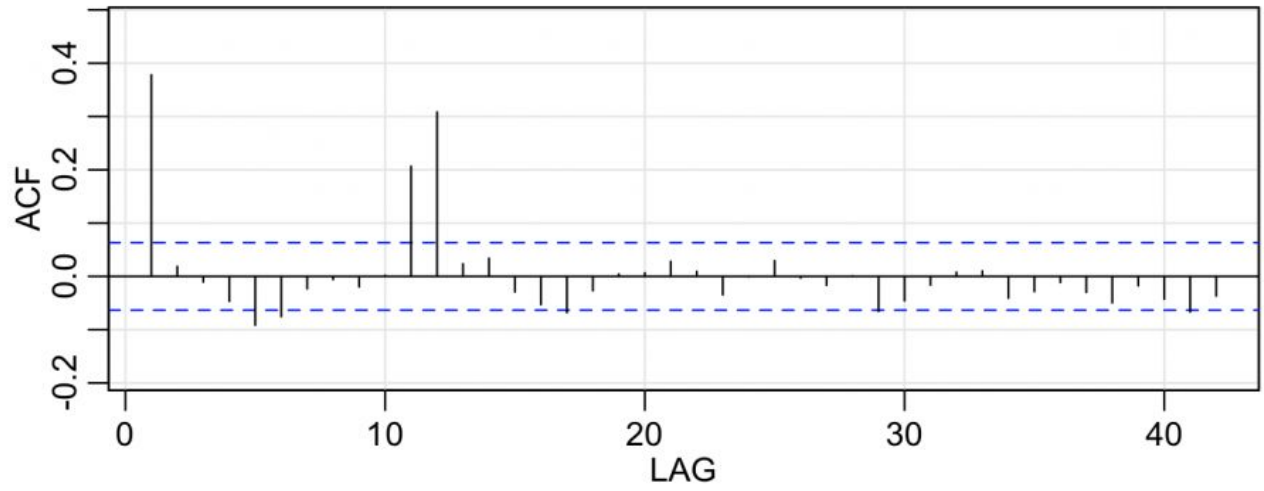
An ARIMA model can be understood by outlining each of its components as follows:

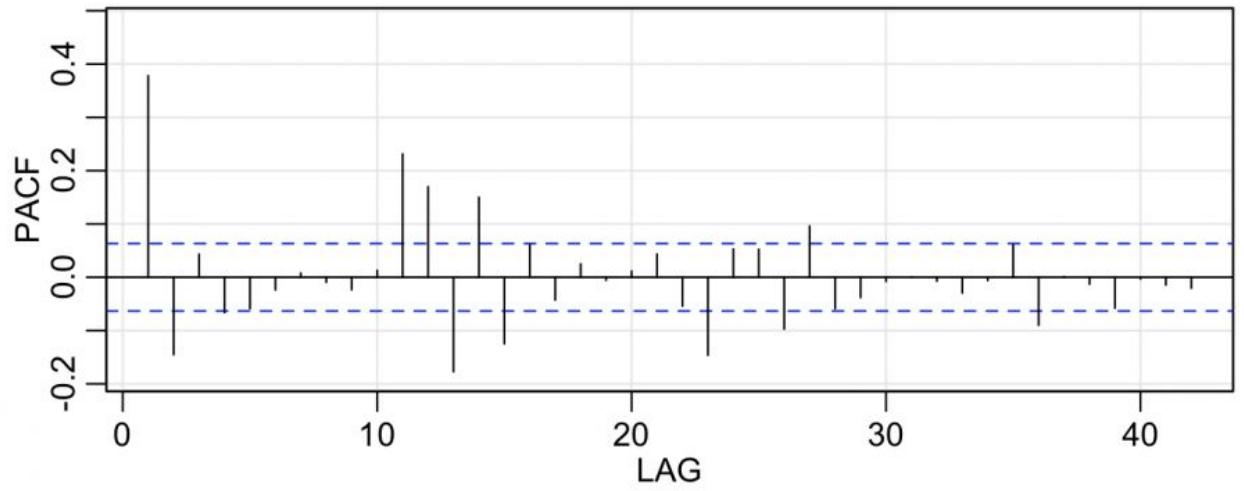
- *Autoregression (AR)* refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- *Integrated (I)* represents the differencing of raw observations to allow for the time series to become stationary, i.e., data values are replaced by the difference between the data values and the previous values.
- *Moving average (MA)* incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The ARIMA model has three important parameters p (AR term), q (MA term), d (order of differencing). The value of p and q are obtained from pacf and acf plots respectively.

Question 10

Identify the p, q values from the following acf pacf plots-





Answer 10

The value of p and q are taken from as the point where the dotted line is intersected the first time. ACF plot is used to determine the value of q while PACF plot is used to find the parameter p.

In the ACF plot shown above, the first intersection is at the value one, so the value of q must be taken as 1. Similarly, the value of p is also 1.

Question 11

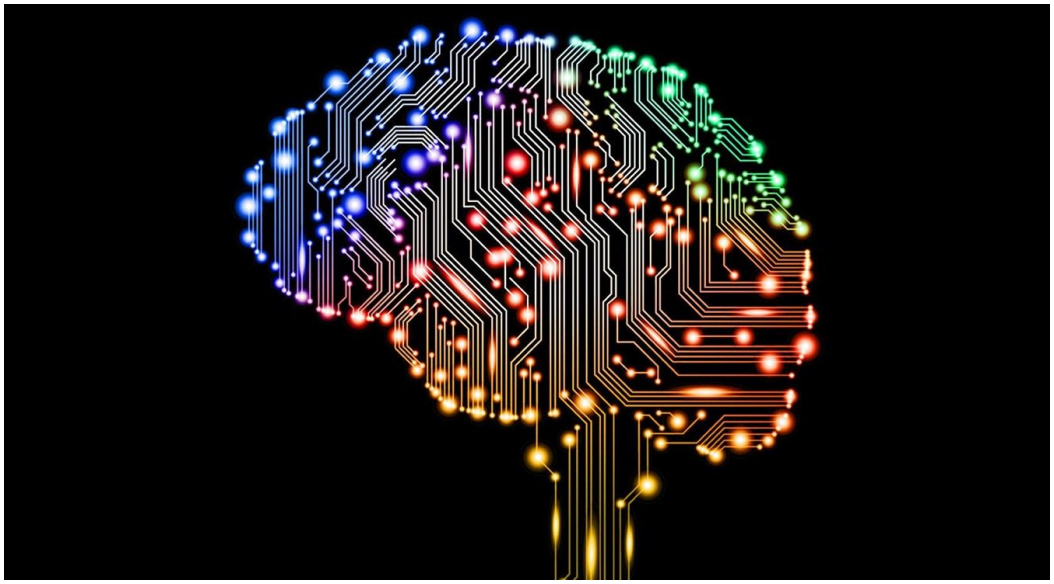
How can a validation set be created from the time series data to check the performance of the model? Can we simply use Cross-validation in this case? why/why not?

Answer 11

The time series data has an order or sequence, which is important for forecasting values. If we use cross validation, then the sequence in the series is destroyed. Instead the data should be split without changing the order in the points.

For instance using first 70% to train the model and next 30% for validation. This link explains a number of methods to backtest machine learning models.

Deep Learning



Question 1

Suppose we have the following architecture for a Multi-Layer Perceptron

a. Number of nodes at input layer = 10
b. Nodes at hidden layer = 5
c. Number of output node= 1

What are the total number of connections?

Answer 1

Each node in the input layer has a connection to each node in the hidden layer. So this makes the total number of connections between input and hidden = 50. Then we will have 5 connections between the hidden and output layer. So total number of connections in the above architecture is 55.

Question 2

How will the structure of a neural network be different for binary and multiclass classification problem?

Answer 2

For binary classification, only one (at most 2) node is required in the last/output layer and “sigmoid activation function” is used.
Whereas in Multiclass classification with k classes, we use k nodes in the last layer of the neural net with “softmax Activation function”

Question 3

here are many different activation functions like relu, sigmoid, softmax or tanh. How can one decide which activation function to use? How will you decide the activation functions for input and output layer?

Answer 3

Sigmoid is an activation function that returns a value between 0 and 1, it is primarily used for the output node of the Binary Classification, but can also be used in the hidden layers.

Tanh is an activation function that returns a value between -1 and 1, it is primarily used in hidden layers of Neural Networks.

Relu is an activation function that returns a linear value when the input is positive and returns 0 when the input is 0 or negative. It is primarily used in hidden layers of Neural Networks.

Softmax is an activation function which is primarily used in the last/output layer of multi-class classification problem.

Question 4

How are weights initialized in a neural network? Are the hidden weights updated during back propagation?

Answer 4

Usually weights can be randomly initialised to a random small value, but that can lead to vanishing gradients and exploding gradient in case of Deep Neural Network. Therefore it is always a good practice to initialise weights using ‘He initialisation” or “Xavier’s Initialisation”.
Yes, the hidden weights are updated during the Backpropagation.

Question 5

The terms ‘cost function’ and ‘loss function’ are often used interchangeably. Is there any difference between the two?

Answer 5

The loss function and cost function are synonyms and there is no difference.

Question 6

It is said that deep networks learns better than shallow ones. Is this statement correct? Why do you think that is the case?

Answer 6

In every consecutive hidden layer, more and more complex permutations of patterns are captured by the neural network based on the previous layers. As Deep Neural network tend to have more such layers, it can learn Very complex patterns as compared to a shallow Neural Network.

But, theoretically speaking even a single hidden layered neural network works as a universal function approximator.

Question 7

What the problems with deep networks?

Answer 7

In case of very deep neural networks, the weights of the hidden layer might experience either vanishing and explosive gradient problem.

They are also prone to “overfitting” the data.

It is difficult to train a neural network and they take a very long time to train.

Question 8

What are the factors to select the depth of neural network?

Answer 8

A. Type of neural network (eg. MLP, CNN etc)
B. Input data

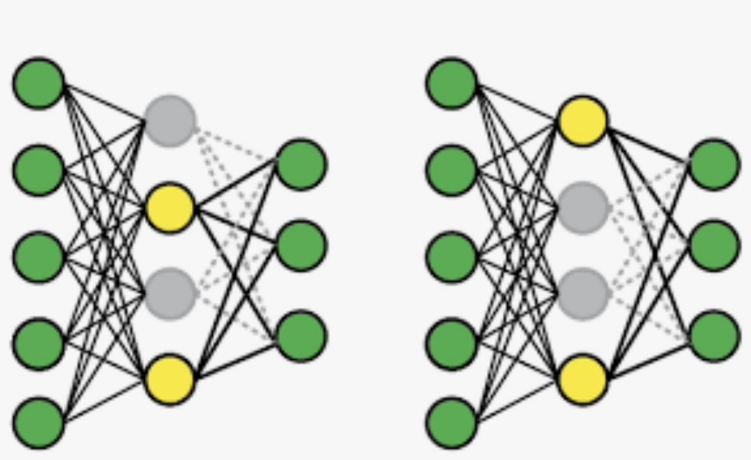
C. Computation power, i.e. Hardware capabilities and software capabilities
D. Learning Rate
E. The output function to map

Question 9

Dropout and DropConnect are both regularization techniques for Neural Network. Is there a difference between these two? How is setting dropout =0.3 different from drop connect =0.3?

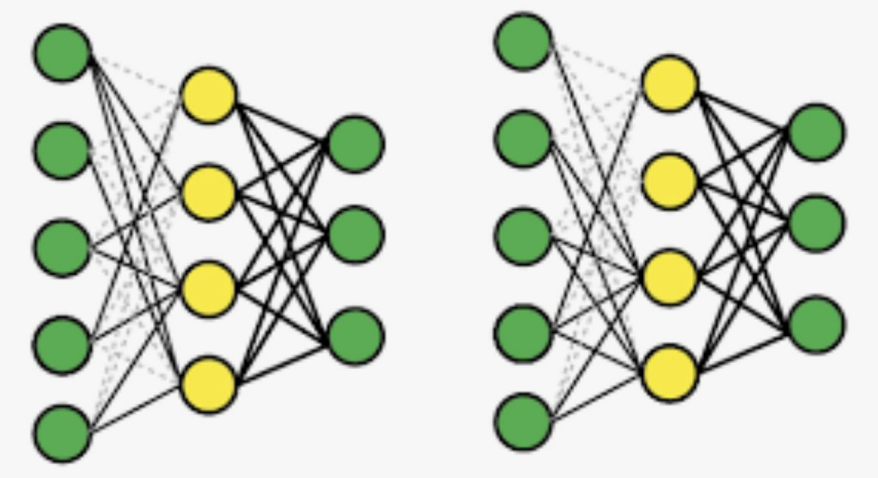
Answer 9

Drop-Out



The function drop-out in a layer assigns a probability p to every node in that layer such that, that node will not be included in the computation during the runtime with respect to probability p. (0.3 in question)

Drop-Connect



The function drop-connection in a layer is a probability p for every node in that layer such that, there is a chance p such that , that node will skip a connection to consecutive layer by the probability p.

Question 10

Suppose a neural network we built, takes time t1 to train and time t2 for predictions. Now we add a dropout=0.3 to all hidden layers. How will this affect t1 and t2?

Answer 10

As we have added 0.3 drop-out to all hidden layers, the actual number of neurons in the network will we reduced by ⅓ to that of without drop-out. Therefore the training time is expected to be 0.7*t1.
There will be no significant difference in the prediction time t2.

Question 11

Neural networks need a large amount of data for training. Suppose you are working on a dataset to classify images of clothes, shoes, watches, etc. You wish to increase the data in this case. List down some techniques for doing so.

Answer 11

We use the technique of data augmentation, we generate new data on the basis of data that we already have by the means of:

- Tilting
- Cropping
- Taking mirror images

We can also scrape the web for more images in order to have more data.

Question 12

In training a neural network, you notice that the loss does not decrease in the few starting epochs. What could be the possible reasons?

Answer 12

There are 4 possible scenarios where this is possible,

- Learning rate is very low
- Regularisation parameter is very high
- Optimisation is stuck in the local optima
- Optimisation started on a plateau

Question 13

what is covariate shift and how is the problem solved?

Answer 13

Covariate shift is a condition when the neurons in deep neural networks stop learning or learn extremely slow due to the vanishing gradient.

It can be solved by (or combination of) following

- Batch normalisation
- Careful Initialisation
- Slow learning rate
- dropout

Question 14

How are CNNs different from Multi layer perceptron?

Answer 14

CNN uses the convolution layer in its network, with the help of which it can utilise the structure of data, whereas Multi Layer Perceptron does not leverage any structural information from data.

Question 15

For an image recognition problem (recognizing a cat in a photo), which architecture of neural network would be better suited to solve the problem

A) MLP
B) CNN
C) RNN

Answer 15

CNN is most preferable as it emphasis on the structure of the data. Therefore it can identify the cat by its features such as whiskers, ears etc.

Question 16

What do you mean by Learning rate?

Answer 16

Learning Rate is a factor by which the weights of neural network are updated in each cycle.

Question 17

Explain the concept of dead unit?

Answer 17

A dead unit in a deep neural network is a neuron which is experiencing the vanishing gradient problem and covariate shift. In this state, the neuron learns extremely slowly or apparently stops learning altogether.

Question 18

Do we need feature engineering in Deep neural networks?

Answer 18

No, we do not necessarily need feature engineering for Deep Learning (Assuming we have ample data), the primary function of the hidden layers in the Deep Neural Network is to form complex features by itself, therefore “handcrafting” of features is not required.

Question 19

What is transfer learning?

Answer 19

Transfer learning is a concept in which we use a pre-trained model instead of training from scratch, the model used must be from similar problem.

Question 20

What happens when we use one-dimensional convolutional neural network on sequential data?

Answer 20

Convolutional neural networks are known to leverage information from the structure of the data, when it is applied on the sequential data, it can learn the short term fixed relation on the sequential data.
BUT, it will fail to capture the variable and long term dependencies in the sequence data, which in-fact is observed in the natural language problems.

Question 21

What is object detection and object localisation ?

Answer 21

Object detection is a process in which the model predicts whether the object is present in the image or not.
In object localisation, the model outputs the coordinate values where the object is present within the image, given the object was present in image.

Question 22

What is the significance of 1 X 1 convolutions?

Answer 22

1x1 convolutions are primarily used as a dimensionality reduction technique, it is primarily used to vary the number of filters in the convolution layers, it can be used to either increase or decrease the number of filters.

Question 23

What is the difference between Stochastic Gradient Descent and Batch Gradient Descent?

Answer 23

Stochastic Gradient Descent uses only one instance to compute the loss and update the parameters. As a result, it converges faster but often yields a sub-optimal solution. Batch gradient Descent on the other hand uses the whole data to compute the loss and update the parameters. As a result, it converges at the slowest rate but guarantees an almost optimal solution.

Question 24

Can we use a multi layer perceptron for regression purpose, how?

Answer 24

A multi layer perceptron can be used for a regression problem, it can be done by removing the activation function at the output node and using a suitable cost function.

Question 25

Why can we not use Multi layer perceptron for text data?

Answer 25

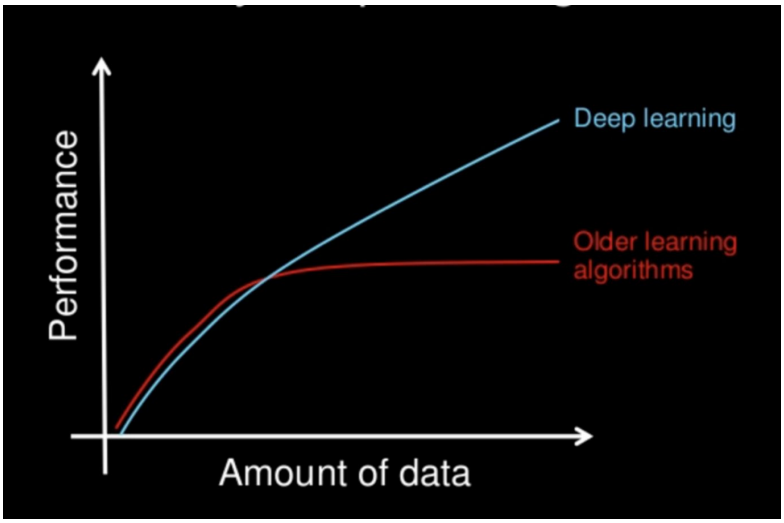
A multi layer perceptron assumes that the input is of constant size. In natural language, the length of the sentences can vary. There multi layer perceptron is not applicable.

Question 26

Why do we even deep learning when we have the traditional machine learning algorithms?

Answer 26

Machine Learning algorithms tend to saturate very early on very large datasets, where as Deep Learning can keep on learning the patterns in the large datasets.



Natural Language Processing

Question 1

N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from given sentence: “Analytics Vidhya is a great source to learn data science”

Answer 1

Analytics Vidhya, Vidhya is, is a, a great, great source, source to, To learn, learn data, data science. So 9

Question 2

Let’s say we have removed the stop words from the above sentence, and now a list of bigrams is created. How many bigrams would you say can be created in this case?

Answer 2

New sentence- Analytics vidhya great source learn data science. So 6.

Question 3

Suppose you are given the data of tweets to classify as ‘positive’, ‘neutral’ and ‘negative’. This data has twitter handles and emojis. How would you work with this data?

Answer 3

The twitter handles do not add any important information to classify, so the best idea would be to clean the tweets data and remove the twitter handles.
Emojis, on the other hand might prove useful in determining the sentiment since people use different emojis to express happiness and anger like :) or :/ :(. so removing emojis will be a loss of information

Question 4

There is a chance that out of thousands of emojis in the data, only certain are useful. Is there any way to determine the which are actually useful?

Answer 4

We can! The term frequency or tf is the number of times a word appears in the document by the total number of words in the document. So a document about data science will have a high number of words such as data science, machine learning, and other words like training, particular, path etc only a few times. So using TF, I can get a good sense of what the document is about. Then why do we need to calculate the inverse document frequency?

Question 5

What are stop words? Why do we remove stop words while using Count Vectoriser?

Answer 5

Parts of speech like “a, an, the, he , she etc.” are considered as stop words, they do not add any significance to the count-vectoriser and classification of text data, they are considered redundant and are removed.

Question 6

What is stemming? Why do we use stemming?

Answer 6

Stemming is the process of extracting the root word from the variants of the same word and these variants make no significant difference in the classification process.
Eg: lover, lovely, loving, loved all have the same root word love.

Question 7

Why do we remove punctuations from the text?

Answer 7

While using the count vectoriser, the punctuations are just a redundancy which add no significance to the classification process.

Question 8

What is bag of words, what does it do?

Answer 8

Bag of words is a count vectoriser which is using in the process of text classification, Bag vectoriser builds a dictionary of all the words present in the input data, this dictionary is a sparse matrix. Now all the instances of the text data are vectorised (sparse) using this dictionary.

Question 9

What are the drawbacks of bag of words?

Answer 9

Bag of words by itself can not handle the the stop words.
Also, Bag of words considers no association between the words and therefore semantic meaning of the input data is completely lost.

Question 10

What is tf-idf? What is its working principle?

Answer 10

Term Frequency - Inverse Document Frequency is a count vectoriser which is an improvement over the bag of words, it can easily handle the stop words.
Tf-Idf works of the principle of Information Theory, which states that frequent words are less significant and vice versa.

Question 11

Which machine Learning algorithms work very well with count vectorised text? Why?

Answer 11

Naive Bayes tend to work best for the count-vectoriser because Words in a count vectoriser have no association between them, which is perfect according to the naive bayes assumption that the features should be perfectly independent.

Question 12

How can we analyse the text data.?

Answer 12

Text data can be analysed by the following ways

- Finding mean length of sentences.
- Finding most frequent words (excluding stop words)
- Doing the above two step with respect to classes.

Question 13

Why are word embeddings more powerful than bag of words and tf-idf vectoriser?

Answer 13

Bag-of-words and tf-idf are are also known as sparse vectors. In this, the vector representation only concerns which words are “present in the text data”. They fail to represent any relation between words.

In word vectors , also known as “dense” vectors, comprises of the properties associated to each word. Therefore in this, the words can be associated to each other and therefore yields a better results compared to the sparse vectors.

Question 14

Why do we need feature transformation?

Answer 14

Machine Learning algorithms only understand numbers and no other format.
Therefore categories are label encoded or Dummy Encoded.
Text is converted to word vectors.

Question 15

Both RNN and LSTM take the sequence of data into account, which is essential when it comes to time series or NLP problems. How are LSTMs different from RNN?

Answer 15

Deep RNN are very prone to the Vanishing gradients, therefore they are not good at retaining long term dependencies, Whereas the LSTM comprises of gates that help in capture the long term dependencies in deep networks.

Question 16

Why are word embeddings more powerful than bag of words and tf-idf vectoriser?

Answer 16

Bag-of-words and tf-idf are are also known as sparse vectors. In this, the vector representation only concerns which words are “present in the text data”. They fail to represent any relation between words.

In word vectors , also known as “dense” vectors, comprises of the properties associated to each word. Therefore in this, the words can be associated to each other and therefore yields a better results compared to the sparse vectors.

Question 17

Why do we need feature transformation?

Answer 17

Machine Learning algorithms only understand numbers and no other format.
Therefore categories are label encoded or Dummy Encoded.
Text is converted to word vectors.