

Where you have used Hypothesis Testing in your Machine learning Solution.

Suppose you are working on a machine learning project, for which you want to predict if a set of patients have or not a mortal disease, based on several features on your dataset as blood pressure, heart rate, pulse and others.

It sounds like a serious project, for which you'll need to trust your model and predictions, right? That's why you got hundreds of samples that your local hospital very gently allowed you to collect, given the importance and the seriousness of the topic. But how do you know if your sample is representative of the whole population? And how can we know how much difference might be reasonable? For example, assume that thanks to some previous studies, we know that the actual probability for any given patient of not having this particular disease is 99%. Now suppose that our sample says that 95% of the patients don't have the condition. Well, 4% difference doesn't sound like a significant difference that may lead us to SUCH bad modelling, right? It might not be the same, but it kind of sounds like it may be representative. To confirm this, we need to build a better understanding of the theoretical background.

Let's start by what we know...the real probability of not having the disease:

$$P(\text{not having the disease}) = 99\%$$

Now let's assume that we find a new group of 100 people, and we test all of them to check if any has this disease we're studying. Can we be sure that 99 of these folks won't have the condition? Maybe, but there's also a possibility that none of them has the disease or even that several may have it. What we have here is a binomial probability problem. The objective of this story is not to talk about probabilities, and however, in simple words, the binomial probability is no more than a given chance of something happening a fixed number of times, given a prior probability for each independent event. We can find it by just applying the following equation:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)! x!} p^x q^{n-x}$$

Where:

- n = the number of trials (or the number being sampled)
- x = the number of successes desired
- p = probability of getting a success in one trial
- $q = 1 - p$ = the probability of getting a failure in one trial

So if we want to know what is the probability that in our sample of 100 guys, we don't have any of them infected with the disease, we may just fill in the blanks to find that the probability is of 36.6%. And if we want to know the probability of having 99% folks NOT infected, we fill in the blanks again, to find out it is approximately 37.0%. And this sounds reasonable: getting 100 out of 100 not infected doesn't sound very unlikely if every single case has 99% of not being infected. And in this line, it also sounds reasonable that having 99 not infected out 100 folks might be a little more likely.

We could keep going and find the probability of having even less people not infected in our sample of 100 people:

- $P(\text{not infected } 98 \text{ out of } 100) = 18.5\%$

- $P(\text{not infected } 97 \text{ out of } 100) = 6.0\%$
- $P(\text{not infected } 96 \text{ out of } 100) = 1.5\%$
- $P(\text{not infected } 95 \text{ out of } 100) = 0.3\%$

Now, let's go back to the sample we had from our friendly local hospital, which says that 95% of the guys in our sample are NOT infected by this horrible mortal disease. Well, even though it might sound like the difference in between 95% and 99% is not relevant, given that we would not be working with a random sample of folks, but instead these guys belong to the same population that we know has a 99% probability of not being infected, we'd be setting a hypothesis that our sample is representative when in reality we would have only 0.3% chance of obtaining a sample with 95 out of 100 people not infected. Therefore we should reject our hypothesis and not proceed.

What kind of statistical tests you have performed in your ML Application

In statistics we have a lot of tests like t-tests, Z-test, anova test, Welch's test etc. we can't say in every problem statement and every project we can use same test. No. It depends on our dataset, problem statements and goal. we can use and perform.

What do you understand by P Value? And what is use of it in ML?

P-value helps us determine how likely it is to get a particular result when the null hypothesis is assumed to be true. It is the probability of getting a sample like ours or more extreme than ours if the null hypothesis is correct. Therefore, if the null hypothesis is assumed to be true, the p-value gives us an estimate of how "strange" our sample is.

If the p-value is very small (<0.05 is considered generally), then our sample is “strange,” and this means that our assumption that the null hypothesis is correct is most likely to be false. Thus, we reject it.

When and how is p-value is used?

P-values are often reported whenever you perform a statistical significance test (like t-test, chi-square test etc). These tests typically return a computed test statistic and the associated p-value. This reported value is used to establish the statistical significance of the relationships being tested.

So, whenever you see a p-value, there is an associated statistical test.

That means there is Hypothesis testing being conducted with a defined Null Hypothesis (H_0) and a corresponding Alternate hypothesis (H_A).

The p-value reported is used to make a decision on whether the null hypothesis being tested can be rejected or not.

Let's understand a little bit more about the null and alternate hypothesis.

Now, how to frame a Null hypothesis in general?

While the null hypothesis itself changes with every statistical test, there is a general principle to frame it:

The null hypothesis assumes there is ‘no effect’ or ‘relationship’ by default.

Significance of P-values:

- If $p > 0.10$: the observed difference is “not significant”
- If $p \leq 0.10$: the observed difference is “marginally significant”
- If $p \leq 0.05$: the observed difference is “significant”
- If $p \leq 0.01$: the observed difference is “highly significant.”

For example: if you are testing if a drug treatment is effective or not, then the null hypothesis will assume there is not difference in outcome between the treated and

untreated groups. Likewise, if you are testing if one variable influence another (say, car weight influences the mileage), then null hypothesis will postulate there is no relationship between the two.

Which type of error is severe Error, Type 1 or Type 2? And why with example.

Ans: It's depends, based on the researcher and instructor Type 1 (false positive) is worse than a Type 2 (false negative) error. The rationale boils down to the idea that if you stick to the status quo or default assumption, at least you're not making things worse. And in many cases, that's true. But like so much in statistics, in application it's not really so black or white. The analogy of the defendant is great for teaching the concept, but when we try to make it a rule of thumb for which type of error is worse in practice, it falls.

In one instance, the Type I error may have consequences that are less acceptable than those from a Type II error. In another, the Type II error could be less costly than a Type I error. And sometimes, as Dan Smith pointed out in Significance a few years back with respect to Six Sigma and quality improvement, "neither" is the only answer to which error is worse:

Most Six Sigma students are going to use the skills they learn in the context of business. In business, whether we cost a company \$3 million by suggesting an alternative process when there is nothing wrong with the current process or we fail to realize \$3 million in gains when we should switch to a new process but fail to do so, the end result is the same. The company failed to capture \$3 million in additional revenue.

POTENTIAL CONSEQUENCES

Since there's not a clear rule of thumb about whether Type 1 or Type 2 errors are worse, our best option when using data to test a hypothesis is to look very carefully at the fallout that might follow both kinds of errors. Several experts suggest using a table like the one below to detail the consequences for a Type 1 and a Type 2 error in your particular analysis.

| Null | Type 1 Error: H_0 true, but rejected | Type 2 Error: H_0 false, but not rejected |
|--|--|--|
| Medicine A does not relieve Condition B. | Medicine A does not relieve Condition B, but is not eliminated as a treatment option. | Medicine A relieves Condition B, but is eliminated as a treatment option. |
| Consequences | Patients with Condition B who receive Medicine A get no relief. They may experience worsening condition and/or side effects, up to and including death. Litigation possible. | A viable treatment remains unavailable to patients with Condition B. Development costs are lost. Profit potential is eliminated. |

Where we can use chi square and have used this test anywhere in your application

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying. Therefore, a chi-square test is an excellent choice to help us better understand and interpret the relationship between our two categorical variables.

Use cases example:

A research scholar is interested in the relationship between the placement of students in the statistics department of a reputed University and their C.G.P.A (their final assessment score).

He obtains the placement records of the past five years from the placement cell database (at random). He records how many students who got placed fell into each of the following C.G.P.A. categories - 9-10, 8-9, 7-8, 6-7, and below 6.

If there is no relationship between the placement rate and the C.G.P.A., then the placed students should be equally spread across the different C.G.P.A. categories (i.e. there should be similar numbers of placed students in each category).

However, if students having C.G.P.A more than 8 are more likely to get placed, then there would be a large number of placed students in the higher C.G.P.A. categories as compared to the lower C.G.P.A. categories. In this case, the data collected would make up the observed frequencies.

So the question is, are these frequencies being observed by chance or do they follow some pattern?

Here enters the chi-square test! The chi-square test helps us answer the above question by comparing the observed frequencies to the frequencies that we might expect to obtain purely by chance.

Reference: <https://www.analyticsvidhya.com/blog/2019/11/what-is-chi-square-test-how-it-works/>

Can we use Chi square with Numerical dataset? If yes, give example. If no, give Reason?

What do you understand by ANOVA Testing?

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not.

In simpler and general terms, it can be stated that the ANOVA test is used to identify which process, among all the other processes, is better. The fundamental concept behind the Analysis of Variance is the “Linear Model”.

Example of ANOVA

An example to understand this can be prescribing medicines.

- Suppose, there is a group of patients who are suffering from fever.
- They are being given three different medicines that have the same functionality i.e., to cure fever.

- To understand the effectiveness of each medicine and choose the best among them, the ANOVA test is used.

You may wonder that a t-test can also be used instead of using the ANOVA test. You are probably right, but, since t-tests are used to compare only two things, you will have to run multiple t-tests to come up with an outcome. While that is not the case with the ANOVA test.

That is why the ANOVA test is also reckoned as an extension of t-test and z-tests.

Types of ANOVA Test

The ANOVA test is generally done in three ways depending on the number of Independent Variables (IVs) included in the test. Sometimes the test includes one IV, sometimes it has two IVs, and sometimes the test may include multiple IVs.

We have three known types of ANOVA test:

1. One-Way ANOVA
2. Two-Way ANOVA
3. N-Way ANOVA (MANOVA)

Example: Suppose medical researchers want to find the best diabetes medicine and choose from four medicines. They can choose 20 patients and give them each of the four medicines for four months.

The researchers can take note of the sugar levels before and after medication for each medicine and then to understand whether there is a statistically significant difference in the mean results from the medications, they can use one-way ANOVA.

The type of medicine can be a factor and reduction in sugar level can be considered the response. Researchers can then calculate the p-value and compare if they are lower than the significance level.

If the results reveal that there is a statistically significant difference in mean sugar level reductions caused by the four medicines, the post hoc tests can be run further to determine which medicine led to this result.

Give me a scenario where you can use Z test and T test.

z-tests are used when we have large sample sizes ($n > 30$), whereas t-tests are most helpful with a smaller sample size ($n < 30$). Both methods assume a normal distribution of the data, but the z-tests are most useful when the standard deviation is known.

Z-test is the statistical test, used to analyze whether two population means are different or not when the variances are known and the sample size is large.

This test statistic is assumed to have a normal distribution, and standard deviation must be known to perform an accurate z-test.

A z-statistic, or z-score, is a number representing the value's relationship to the mean of a group of values, it is measured with population parameters such as population standard deviation and used to validate a hypothesis.

For example, the null hypothesis is "sample mean is the same as the population mean", and the alternative hypothesis is "the sample mean is not the same as the population mean".

T-test:

In order to know how significant the difference between two groups are, a T-test is used; basically, it tells that difference (measured in means) between two separate groups could have occurred by chance.

This test assumes to have a normal distribution while based on t-distribution, and population parameters such as mean, or standard deviation are unknown.

The ratio between the difference between two groups and the difference within the group is known as T-score. Greater is the t-score, more is the difference between groups, and smaller is the t-score, more similarities are there among groups.

For example, a t-score value of 2 indicates that the groups are two times as different from each other as they are with each other.

Also, after running t-test, if the larger t-value is obtained, it is highly likely that the outcomes are more repeatable, such that

- A larger t-score states that groups are different
- A smaller t-score states that groups are similar.

Mainly, there are three types of t-test:

1. An Independent Sample t-test, compare the means for two groups.
2. A Paired Sample t-test, compare means from the same group but at different times, such as six months apart.
3. A One Sample t-test, test a mean of a group against the known mean.

What do you understand by inferential Statistics?

· Inferential statistics is work with a random sample of data taken from a population to illustrate and make inferences about the population.

· Inferential statistics are valuable when working with of each member of an entire population is not convenient or possible.

- It's help us get to the conclusions and make predictions based on our data.
- Inferential statistics understands the whole population from sample taken from it.
- In Inferential statistics we use a random sample, so we can generalize outcome from the sample to the large population.
- In Inferential statistics, we can calculate the mean, standard deviation, and proportion for our random sample data from population.

The following types of inferential statistics are mostly used and quite easy to interpret:

- Conditional Probability
- Probability Distribution and Distribution function
- Probability
- Regression Analysis
- Central Limit Theorem
- Hypothesis Testing
- T- Test
- Z- Test
- Sampling Distribution
- Chi-square test

- Confidence Interval

- ANOVA (Analysis of variance)

When you are trying to calculate Std Deviation or Variance, why you used N-1 in Denominator?

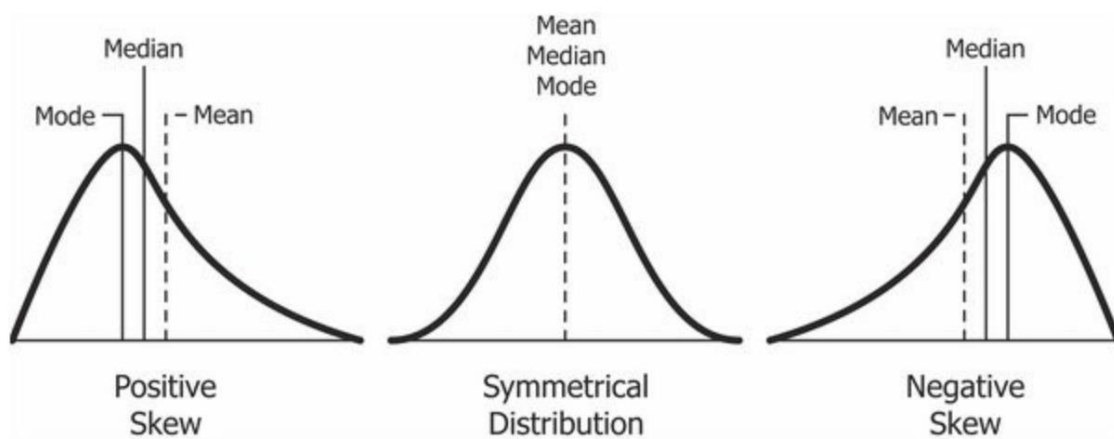
The n-1 equation is used in the common situation where you are analyzing a sample of data and wish to make more general conclusions. The SD computed this way (with n-1 in the denominator) is your best guess for the value of the SD in the overall population.

If you simply want to quantify the variation in a particular set of data, and don't plan to extrapolate to make wider conclusions, then you can compute the SD using n in the denominator. The resulting SD is the SD of those particular values. It makes no sense to compute the SD this way if you want to estimate the SD of the population from which those points were drawn. It only makes sense to use n in the denominator when there is no sampling from a population, there is no desire to make general conclusions.

What do you understand by right skewness, Give example?

Ans: skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution. Now, you might be thinking - why am I talking about normal distribution here?

Well, the normal distribution is the probability distribution without any skewness. You can look at the image below which shows symmetrical distribution that's basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. Apart from this, there are two types of skewness:



The probability distribution with its tail on the right side is a positively skewed distribution and the one with its tail on the left side is a negatively skewed distribution. If you're finding the above figures confusing, that's alright. We'll understand this in more detail later.

What is difference between Normal distribution and Std Normal Distribution and Uniform Distribution?

The only thing similar about the two is that they are both continuous distributions with two parameters. Differences include:

1. Normal has infinite support, uniform has finite support
2. Normal has a single most likely value, uniform has every allowable value equally likely
3. Uniform has a piecewise constant density, normal has a continuous bell-shaped density
4. Normal distributions arise from the central limit theorem, uniforms do not.

What is different kind of Probabilistic distributions you heard of?

Probability: Simply put, probability is an intuitive concept. We use it on a daily basis without necessarily realising that we are speaking and applying probability to work.

Life is full of uncertainties. We don't know the outcomes of a particular situation until it happens. Will it rain today? Will I pass the next math test? Will my favourite team win the toss? Will I get a promotion in next 6 months? All these questions are examples of uncertain situations we live in. Let us map them to few common terminologies which we will use going forward.

Experiment - are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.

Outcome is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained"

Event is one or more outcome from an experiment. "It rained" is one of the possible events for this experiment.

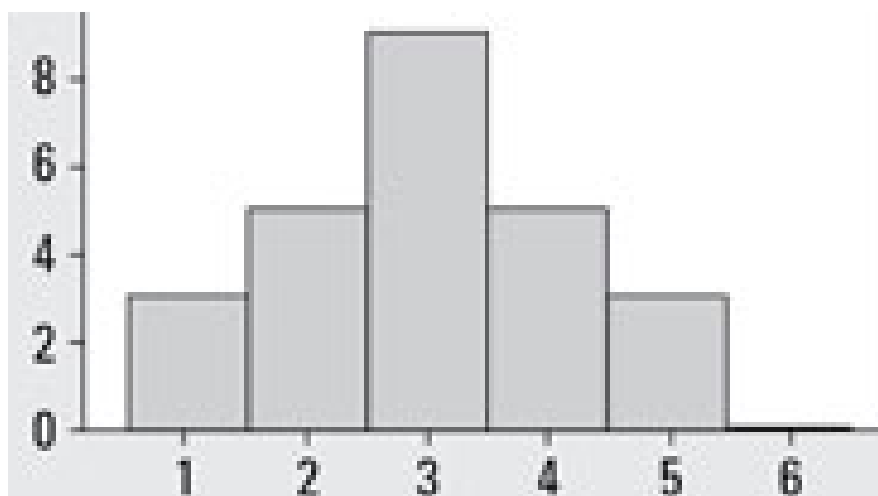
Probability is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6

Types of Distributions

1. Bernoulli Distribution
2. Uniform Distribution
3. Binomial Distribution
4. Normal Distribution
5. Poisson Distribution
6. Exponential Distribution

What do you understand by symmetric dataset?

If the data are symmetric, they have about the same shape on either side of the middle. In other words, if you fold the histogram in half, it looks about the same on both sides. Below figure shows an example of symmetric data. With symmetric data, the mean and median are close together.



In your last project, were you using symmetric data or Asymmetric Data, if its asymmetric, what kind of EDA

you have performed?

Ans: Recently I have done one Chatbot project, here I have dataset in the from of CSV. During the EDA we have some lots of steps because our dataset is a imbalanced and in our dataset we have lots of outliers.

1. Here, we are plotting a Box plot and scatter plot for checking a outlier.
2. We are plotting a count plot for checking a how many total values available in our every features.
3. As we know our dataset is imbalanced so here we are used a under sampling for handling a imbalanced datasets.
4. Before handling a imbalanced data and outliers we are getting 43% accuracy but after handling this we are getting 83% using DistilBERT.

Can you please tell me formula for skewness?

Skewness = $(3 * (\text{Mean} - \text{Median})) / \text{Standard Deviation}$

Have you applied student T distribution Anywhere?

Student T distribution is a type of normal distribution wherein it can be used for smaller sample sizes and is approximately normal distributed.

I have used it in a situation where I was supposed to measure the average of salary of just 10 people which was very low. But I found that there is some bell curve kind of a structure approximately. So to measure the average of salary I used T distribution

What do you understand by statistical analysis of data, Give me scenario where you have used statistical analysis in last projects?

In one of my projects I had a dataset wherein there were lot of columns. At that time as from domain I made an assumption that these two features looks same so I looked up for the correlation between the columns vs the dependent column and it was multicollinear so I removed one of the columns. So this helped me in feature selection step. Similarly I have used Chi square based approach also in one of my projects. I use describe function in most of my projects to get to know about the data better.

Can you please tell me criterion to apply binomial distribution, with example?

The criteria are :

The observations must be independent of each other,

The number of observations must be fixed,

The probability of success is same for each outcome

A real time example is Lottery ticket where there is only two ways either you reach success or failure

lets suppose I have appeared in 3 interviews, what is the probability that I am able to crack at least 1 interview?

$$n(S) = 3$$

$$\begin{aligned} P(A \leq 1) &= 1 - P(A=1) * P(A=2) * P(A=3) \\ &= 1 - \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = 1 - \frac{1}{8} = \frac{7}{8} \end{aligned}$$

Explain Gaussian Distribution in your own way.

Gaussian distribution is also called as normal distribution which has a bell shaped curve structured data distribution. This is denoted as

$$N(\mu, \sigma^2)$$

Here always the skewness and kurtosis is 0

What do you understand by 1st, 2nd and 3rd Standard Deviation from Mean?

According to Empirical Formula,

If the data is normally distributed

The data between the mean and 1st standard deviation is 68% approx.

The data between the mean and 2nd standard deviation is 95% approx.

The data between the mean and 3rd standard deviation is 99.7% approx.

If not normal, then we can use Chebyshev's inequality to find how the data is distributed between mean and 1st, 2nd and 3rd deviation.

What do you understand by variance in data in simple words?

Variance is spread of dataset. It is a statistical measure which will say how the data is distributed. It shows how much a data is deviated from the mean. It is denoted by sigma

Formula: $\text{Variance} = (\text{Summation of } (X - \mu)^2) / N$

X = Data value

μ = mean

N = Total Population

If variance of dataset is too high, in that case How you will be able to handle it or decrease it?

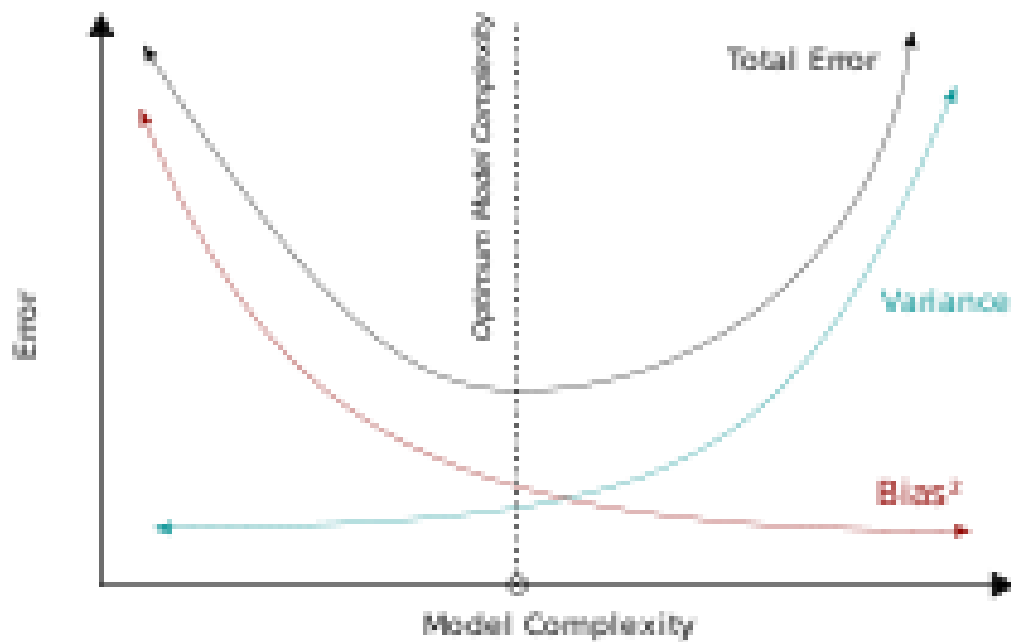
- Ensemble models

- Increasing the train dataset size

- Hyperparameter Tuning

Explain the relationship between Variance and Bias.

Bias tells us the difference between the average difference between the actual value and predicted value and Variance says how the data is spread. Variance and Bias are inversely proportional which means if bias increase variance will decrease or vice versa.



Overfitting - Low Bias High Variance

Underfitting - High Bias High Variance

Perfect - Low Bias, Low Variance

What do you understand by Z Value given in Z Table?

A Z Value in a Z Table will give us how much percentage of data is available to the left of the given Z score

For example for a specific z score 1.25, 89.44% of value lies behind the given z score (left of it)

Do you know a Standard Normal Distribution Formula?

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

To convert a data into standard normal distributed data the following formula can be used:

$z = (x - \text{mean}) / \text{standard deviation}.$

Can you please explain critical region in your way?

It is a region we use in hypothesis testing. We used make a null hypothesis and if the statistic tests like T test Anova test outputs a value in this critical region then we will reject the null hypothesis and accept the alternate hypothesis. So it is also called the rejection region.

Have you used AB testing in your project So far? If yes, Explain. If not, Tell me about AB testing.

If yes:

Yes we have used a/b testing in our data science project for Analyzing the Results of two models. It is one of the most effective methods in making conclusions about any hypothesis one may have. We create the A and B version of our model and calculate the success rate of the that based on the comparison

Tell me about AB testing

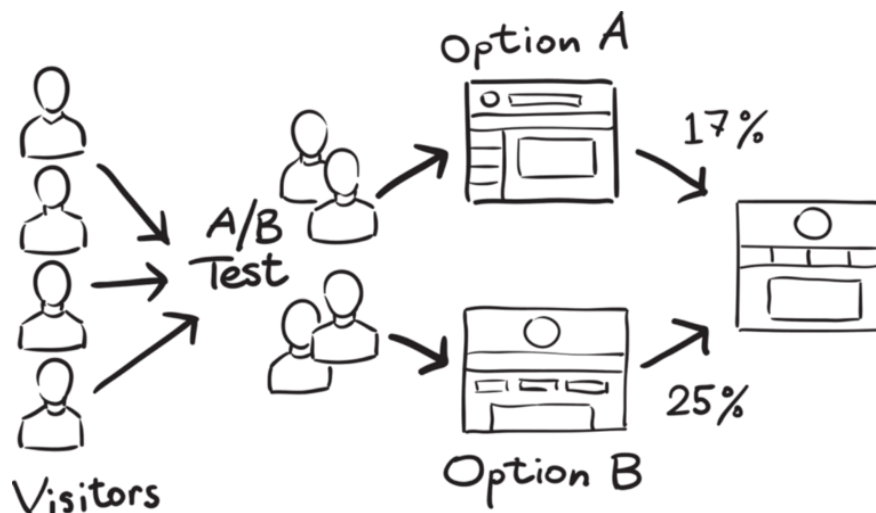
A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better.

Or

(definition from wikipedia)

A/B testing is a method of comparing two versions of a product or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a product are shown to users at random, and statistical analysis is used to determine which variation performs better.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.



In the above scenario, you may divide the products into two parts - A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

Can we use the Alternate hypothesis as a null Hypothesis?

Hypothesis is a statement, assumption or claim about the value of the parameter (mean, variance, median etc).

Like, if we make a statement that "Dhoni is the best Indian Captain ever." This is an assumption that we are making based on the average wins and loses team had under his captaincy. We can test this statement based on all the match data.

Null Hypothesis

The null hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true. The concept of the null is similar to innocent until proven guilty. We assume innocence until we have enough evidence to prove that a suspect is guilty.

It is denoted by H_0 .

Alternate Hypothesis

The alternative hypothesis complements the Null hypothesis. It is opposite of the null hypothesis such that both Alternate and null hypothesis together cover all the possible values of the population parameter.

It is denoted by H_1 .

Let's understand this with an example:

A soap company claims that its product kills on an average 99% of the germs. To test the claim of this company we will formulate the null and alternate hypothesis.

Null Hypothesis(H_0): Average = 99%

Alternate Hypothesis(H_1): Average is not equal to 99%.

Note: The thumb rule is that a statement containing equality is the null hypothesis.

Hypothesis Testing

When we test a hypothesis, we assume the null hypothesis to be true until there is sufficient evidence in the sample to prove it false. In that case we reject the null hypothesis and support the alternate hypothesis.

If the sample fails to provide sufficient evidence for us to reject the null hypothesis, we cannot say that the null hypothesis is true because it is based on just the sample data. For saying the null hypothesis is true we will have to study the whole population data.

So the main question is: Can we use the Alternate hypothesis as a null Hypothesis?

No, We can't use it based on the above explanation. The alternate hypothesis is the opposite of the null hypothesis.

Can you please explain the confusion matrix for more than 2 variables?

Confusion matrix is a performance measurement for machine learning classification problems where output can be two or more classes.





For 2 variables:

It is a table with 4 different combinations of predicted and actual values

| | | Actual Values | |
|------------------|--------------|---------------|--------------|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.

Let's understand TP, FP, FN, TN in terms of pregnancy analogy.

| | | Actual Values | |
|------------------|---|--|--|
| | | 1 | 0 |
| Predicted Values | 1 | TRUE POSITIVE  | FALSE POSITIVE  |
| | 0 | FALSE NEGATIVE  | TRUE NEGATIVE  |

Confusion matrix for a 3 class classification:

Let's try to answer the above question with a popular dataset - IRIS DATASET.

The dataset has 3 flowers as outputs or classes, Versicolor, Virginia, Setosa.



Source: Google

With the help of petal length, petal width, sepal length, sepal width the model has to classify the given instance as Versicolor or Virginia or Setosa flower.

Let's apply a classifier model here: decision Tree classifier is applied on the above dataset. The dataset has 3 classes hence we get a 3 X 3 confusion matrix.

But how to know TP, TN, FP, FN values !!!!!

In the multi-class classification problem, we won't get TP, TN, FP, FN values directly as in the binary classification problem. We need to calculate for each class.

How to calculate FN, FP, TN, TP :

FN: The False-negative value for a class will be the sum of values of corresponding rows except for the TP value.

FP: The False-positive value for a class will be the sum of values of the corresponding column except for the TP value.

TN: The True Negative value for a class will be the sum of values of all columns and rows except the values of that class that we are calculating the values for.

TP: The True positive value is where the actual value and predicted value are the same.

The confusion matrix for the IRIS dataset is as below:

| | | Predicted Values | | |
|---------------|------------|-----------------------|-----------------------|-----------------------|
| | | Setosa | Versicolor | Virginica |
| Actual Values | Setosa | 16 (cell 1) | 0 (cell 2) | 0 (cell 3) |
| | Versicolor | 0 (cell 4) | 17 (cell 5) | 1 (cell 6) |
| | Virginica | 0 (cell 7) | 0 (cell 8) | 11 (cell 9) |

1. Let us calculate the TP, TN, FP, FN values for the class Setosa using the Above tricks:

TP: The actual value and predicted value should be the same. So concerning the Setosa class, the value of cell 1 is the TP value.

FN: The sum of values of corresponding rows except the TP value

$$\text{FN} = (\text{cell 2} + \text{cell 3})$$

$$= (0 + 0)$$

$$= 0$$

FP : The sum of values of the corresponding column except the TP value.

$$\text{FP} = (\text{cell 4} + \text{cell 7})$$

$$= (0 + 0)$$

$$= 0$$

TN: The sum of values of all columns and rows except the values of that class that we are calculating the values for.

$$\text{TN} = (\text{cell 5} + \text{cell 6} + \text{cell 8} + \text{cell 9})$$

$$= 17 + 1 + 0 + 11$$

$$= 29$$

Similarly, for Versicolor class the values/ metrics are calculated as below:

TP : 17 (cell 5)

FN : 0 + 1 = 1 (cell 4 +cell 6)

FP : 0 + 0 = 0 (cell 2 + cell 8)

TN : 16 +0 +0 + 11 =27 (cell 1 + cell 3 + cell 7 + cell 9).

I hope the concept is clear and you can try for the Virginia class.

Give me an example of False Negative From this interview?

A false negative error, or false negative, is a test result which wrongly indicates that a condition does not hold. For example, when a pregnancy test indicates a woman is not pregnant, but she is, or when a person guilty of a crime is acquitted, these are false negatives.



What do you understand by Precision, Recall and F1 Score with example?

Confusion Matrix

A typical confusion matrix looks like the figure shown.

Where the terms have the meaning:

True Positive(TP): A result that was predicted as positive by the classification model and also is positive

True Negative(TN): A result that was predicted as negative by the classification model and also is negative

False Positive(FP): A result that was predicted as positive by the classification model but actually is negative

False Negative(FN): A result that was predicted as negative by the classification model but actually is positive.

The Credibility of the model is based on how many correct predictions the model did.

What is the accuracy of the machine learning model for this classification task?

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Accuracy represents the number of correctly classified data instances over the total number of data instances.

In this example, Accuracy = $(55 + 30)/(55 + 5 + 30 + 10) = 0.85$ and in percentage the accuracy will be 85%.

Is accuracy the best measure?

Accuracy may not be a good measure if the dataset is not balanced (both negative and positive classes have different numbers of data instances). We will explain this with an example.

Consider the following scenario: There are 90 people who are healthy (negative) and 10 people who have some disease (positive). Now let's say our machine learning model perfectly classified the 90 people as healthy but it also classified the unhealthy people as healthy. What will happen in this scenario? Let us see the confusion matrix and find out the accuracy?

In this example, TN = 90, FP = 0, FN = 10 and TP = 0. The confusion matrix is as follows.

| | | PREDICTED LABEL | |
|------------|----------|----------------------|---------------------|
| | | NEGATIVE | POSITIVE |
| TRUE LABEL | NEGATIVE | 90 TRUE NEGATIVE | 0 FALSE POSITIVE |
| | POSITIVE | 10 FALSE NEGATIVE | 0 TRUE POSITIVE |

Figure 7: Confusion matrix for healthy vs unhealthy people classification task.

Accuracy in this case will be $(90 + 0)/(100) = 0.9$ and in percentage the accuracy is 90 %.

Is there anything fishy?

The accuracy, in this case, is 90 % but this model is very poor because all the 10 people who are unhealthy are classified as healthy. By this example what we are trying to say is that accuracy is not a good metric when the data set is unbalanced. Using accuracy in such scenarios can result in misleading interpretation of results.

So now we move further to find out another metric for classification. Again we go back to the pregnancy classification example.

Now we will find the precision (positive predictive value) in classifying the data instances. Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

What does precision mean?

Precision should ideally be 1 (high) for a good classifier. Precision becomes 1 only when the numerator and denominator are equal i.e $TP = TP + FP$, this also means FP is zero. As FP increases the value of the denominator becomes greater than the numerator and precision value decreases (which we don't want).

So in the pregnancy example, precision = $30/(30 + 5) = 0.857$

Now we will introduce another important metric called recall. Recall is also known as sensitivity or true positive rate and is defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

Recall should ideally be 1 (high) for a good classifier. Recall becomes 1 only when the numerator and denominator are equal i.e $TP = TP + FN$, this also means FN is zero. As FN increases the value of the denominator becomes greater than the numerator and recall value decreases (which we don't want).

So in the pregnancy example let us see what the recall will be.

Recall = $30/(30 + 10) = 0.75$

So ideally in a good classifier, we want both precision and recall to be one which also means FP and FN are zero. Therefore we need a metric that takes into account both precision and recall. F1-score is a metric which takes into account both precision and recall and is defined as follows:

$$F1 \text{ Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 Score becomes 1 only when precision and recall are both 1. F1 score becomes high only when both precision and recall are high. F1 score is the harmonic mean of precision and recall and is a better measure than accuracy.

In the pregnancy example, F1 Score = $2 * (0.857 * 0.75)/(0.857 + 0.75) = 0.799$.

What kind of questions do you ask your client if they give you a dataset?

- How was the data compiled? Was it aggregated from multiple sources? ...
- Is the data accurate? ...
- Is the data clean? ...
- How much data should you have? ...
- Remember why: what problem do you want to tackle?
- Dimension of the dataset
- Type of the attributes in the dataset
- For predictive analytics, target attribute
- Missing values in the data set
- How to fill missing values?

Have you ever done an F test on your dataset, if yes, give an example. If No, then explain F distribution?

F Distribution

F-Test (variance ratio test)

When we run a regression analysis, we get f value to find out the means between two populations. It's similar to a T statistic from a T-Test. A T-test will tell you if a single variable is related statistically, and an F test will tell you if a group of variables is jointly significant.

- F-test is used to test the two independent estimations of population variances(S_1^2 & S_2^2).
- F-test is used by comparing the ratio of the two variances S_1^2 & S_2^2 .
- The samples must be independent.
- F-test is a small sample test.
- $F = (\text{Larger estimate of population variance}) / (\text{Smaller estimate Of population variance})$
- The variance ratio = S_1^2 & S_2^2
- F-test never is -ve because the upper value is greater than lower.
- Degree of freedom for larger population[vS1] variance is $V1[vS2]$ and smaller $V2$
- The null hypothesis of two population variance are equal, i.e., $H_0: S_1^2 = S_2^2$

Determining the Values of F

F Distribution using Python

#import scipy, numpy and matplotlib

`x=np.linspace(-10, 10, 100)`

`dfn = 29`

dfd = 18

```
mean, var, skew, kurt = scipy.stats.f.stats(dfn, dfd, moments='mvsk')
```

```
print('mean: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, skew, kurt))
```

```
plt.plot(x, scipy.stats.f.pdf(x, dfn, dfd))
```

```
plt.show()
```

mean: 1.12, skewness: 0.28, kurtosis: 1.81

Note:

- The Student 't' distribution is robust, which means that if the population is non-normal, the results of the t-test and confidence interval estimate are still valid provided that the population is not extremely non-normal.
- To check this requirement, draw a histogram of the data and see how bell-shaped the resulting figure is. If a histogram is extremely skewed (say in that case of an exponential distribution), that could be considered “extremely non-normal,” and hence, t-statistics would not be valid in this case.

Example

Question: From a population of women, suppose you randomly select 7 women, and from the population of men, 12 men are selected.

| Population | Population standard deviation | Sample deviation | standard |
|------------|-------------------------------|------------------|----------|
| Women | 30 | 35 | |
| Men | 50 | 45 | |

To calculate f statistics.

Answer: The f statistic can be calculated from the sample standard deviations and population, using the following equation:

$$f = [s_1^2 / \sigma_1^2] / [s_2^2 / \sigma_2^2]$$

where Standard deviation of the sample drawn from population 1 is s_1 and s_2 in the denominator is the standard deviation of the sample drawn from population 2, σ_1 is the standard deviation of population 1, Population 2's standard deviation is σ_2 .

As we can see from the equation, there are two ways to compute an f statistic from these data. If the data of women appears in the numerator, we can compute f statistic as follows:

$$f = (55^2 / 20^2) / (45^2 / 50^2)$$

$$f = (3025 / 400) / (2025 / 2500).$$

$$f = 1.361 / 0.81 = 1.68$$

For calculations, the numerator degrees of freedom v_1 are 7 - 1 or 6; and the degrees of freedom for denominator v_2 are 12 - 1 or 11.

On the other hand, if the men's data appears in the numerator, we can calculate the f statistic as follows:

$$f = (45^2 / 50^2) / (55^2 / 20^2)$$

$$f = (2025 / 2500) / (3025 / 400)$$

$$f = 0.812 / 1.3610 = 0.5955$$

For this calculation, the denominator degrees of freedom v_2 is 7 - 1 or 6 and the numerator degrees of freedom v_1 is 12 - 1 or 11

When we are trying to find the cumulative probability associated with an f statistic, you need to know v_1 and v_2 .

Find the cumulative probability related to each of the f statistics from the above example:

Answer: First, we need to find the degrees of freedom for each sample. Then, probabilities can be found.

- The sample of women's degrees of freedom is equal to $n - 1 = 7 - 1 = 6$.
- The sample of men's degrees of freedom is equal to $n - 1 = 12 - 1 = 11$.

Therefore, when data of women appear in the numerator, then v_1 is equal to 6; and then v_2 is equal to 11. And, the f statistic is equal to 1.68. So, 0.78 is the cumulative probability.

When data of men appear in the numerator, then v_1 is equal to 11; and then v_2 is equal to 6. And, the f statistic is equal to 0.595. Thus the cumulative probability is 0.22.

What is AUC & ROC Curve? Explain with uses.

We know that the classification algorithms work on the concept of probability of occurrence of the possible outcomes. A probability value lies between 0 and 1. Zero means that there is no probability of occurrence and one means that the occurrence is certain.

But while working with real-time data, it has been observed that we seldom get a perfect 0 or 1 value. Instead of that, we get different decimal values lying between 0 and 1. Now the question is if we are not getting binary probability values how are we actually determining the class in our classification problem?

There comes the concept of Threshold. A threshold is set, any probability value below the threshold is a negative outcome, and anything more than the threshold is a favourable or the positive outcome. For Example, if the threshold is 0.5, any probability value below 0.5 means a negative or an unfavourable outcome and any value above 0.5 indicates a positive or favourable outcome.

Now, the question is, what should be an ideal threshold?

The horizontal lines represent the various values of thresholds ranging from 0 to 1.

- * Let's suppose our classification problem was to identify the obese people from the given data.

- * The green markers represent obese people and the red markers represent the non-obese people.

- * Our confusion matrix will depend on the value of the threshold chosen by us.

- * For Example, if 0.25 is the threshold then

TP(actually obese)=3

TN(Not obese)=2

FP(Not obese but predicted obese)=2(the two red squares above the 0.25 line)

FN(Obese but predicted as not obese)=1(Green circle below 0.25line)

A typical ROC curve looks like the following figure.

- * Mathematically, it represents the various confusion matrices for various thresholds. Each black dot is one confusion matrix.

- * The green dotted line represents the scenario when the true positive rate equals the false positive rate.

- * As evident from the curve, as we move from the rightmost dot towards left, after a certain threshold, the false positive rate decreases.

- * After some time, the false positive rate becomes zero.

- * The point encircled in green is the best point as it predicts all the values correctly and keeps the False positive as a minimum.

- * But that is not a rule of thumb. Based on the requirement, we need to select the point of a threshold.

- * The ROC curve answers our question of which threshold to choose.

But we are confused!!

Let's suppose that we used different classification algorithms, and different ROCs for the corresponding algorithms have been plotted.

The question is: which algorithm to choose now?

The answer is to calculate the area under each ROC curve.

AUC(Area Under Curve)

- * It helps us to choose the best model amongst the models for which we have plotted the ROC curves
- * The best model is the one that encompasses the maximum area under it.
- * In the adjacent diagram, amongst the two curves, the model that resulted in the red one should be chosen as it clearly covers more area than the blue one

Who decided in your last project, what will be the accuracy of your model & what was the criterion to make the decision.

Whether i was doing a classification problem so i have chosen parameters for classification model evaluation

For classification model evaluation, we have different parameters like performance matrix, pr curve, roc-auc curve in performance matrix also we have different different evaluation parameters like accuracy, error rate, precision, recall so based on our class distribution, we can choose any of them, first i have checked accuracy of the model and also i have gone through with roc-auc curve inside that i have checked auc score of the given model.

In the auc score I had the criterion 0.5 or 50% based on that i have filtered the model then i have compared auc score between the models as well so whatever auc score i have found greater that i have model i have chosen finally.

What do you understand by 1 tail test & 2 tail test? Give an example.

If the alternate hypothesis gives the alternate in both directions (less than and greater than) of the value of the parameter specified in the null hypothesis, it is called a Two tailed test.

If the alternate hypothesis gives the alternate in only one direction (either less than or greater than) of the value of the parameter specified in the null hypothesis, it is called One tailed test.

e.g. if H_0 : mean = 100 H_1 : mean not equal to 100

Here according to H_1 , the mean can be greater than or less than 100. This is an example of Two tailed test

Similarly, if $H_0: \text{mean} \geq 100$ then $H_1: \text{mean} < 100$

Here, the mean is less than 100, it is called One tailed test.

What do you understand by the power of a test?

The statistical power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis H_0 when a specific alternative hypothesis H_1 is true. It is commonly denoted by $1 - \beta$, and represents the chances of a "true positive" detection conditional on the actual existence of an effect to detect. Statistical power ranges from 0 to 1, and as the power of a test increases,

the probability β of making a type II error by wrongly failing to reject the null hypothesis decreases.

How do you set the level of significance for your dataset?

In normal English, "significant" means important, while in Statistics "significant" means probably true (not due to chance). A research finding may be true without being important. When statisticians say a result is "highly significant" they mean it is very probably true. They do not (necessarily) mean it is highly important.

we are determining the significance level(α). This refers to the likelihood of rejecting the null hypothesis even when it's true. A common α is 0.05 or 5 per cent.(We can choose 1% or 10% as well)

Have you ever used a T table in any of your projects so far? If No, then why is statistics important for data scientists? If yes, explain the scenario.

It is the science of conducting studies to collect, organize, summarize, analyze, and draw a conclusion out of data. It deals with collective informative data, interpreting those data, and drawing a conclusion from that data. It is used in many disciplines like marketing, business, healthcare, telecom, etc.

In any data science project, data helps us to analyze the initial level of insight.

Building models using popular statistical methods such as Regression, Classification, Time Series Analysis and Hypothesis Testing which is core of data science. Data Scientists run suitable experiments and interpret the results with the help of these statistical methods.

So we have a data and based on that data we are going to create a statistical model which will be able to learn from data itself above some of the statics techniques has been given based on this scenario you can understand statistics is important with respect to datascience.

Can we productionise statistical model?

What is productionize

It means testing and deploying an application to production such that it uses real data on a frequent basis to produce output for use by the business. When data scientists build and test models, it is often a very manual process.

Optimizing data science across the entire enterprise requires more than just cool tools for wrangling and analyzing data. Obviously, we can simply hardcode a data science model or rent a pre-trained predictive model in the cloud, embed it into an application in-house and we are done.yes so we can productionise our statistical model.

How frequently do you build the model and test it?

If a model's predictive performance has fallen due to changes in the environment, the solution is to retrain the model on a new training set, which reflects the current reality. How often should you retrain your model? And how do you determine your new training set? The answer is that it depends. But what does it depend on?

Sometimes the problem setting itself will suggest when to retrain your model. For instance, suppose you're working for a university admissions department and are tasked with building a student attrition model that predicts whether a student will return the following semester. This model will be used to generate predictions on the current cohort of students directly after midterms. Students identified as being at risk of churning will automatically be enrolled in tutoring or some other such intervention.

Let's think about the time horizon of such a model. Since we're generating predictions in batches once a semester, it doesn't make sense to retrain the model any more often than this because we won't have access to any new training data. Therefore we might choose to retrain our model at the start of each semester after we've observed which students from the previous semester dropped out.

This is an example of a periodic retraining schedule. It's often a good idea to start with this simple strategy but you'll need to determine based on your business problem exactly how frequently you'll need to retrain.

Quickly changing training sets might require you to train as often as daily or weekly. Slower varying distributions might require monthly or annual retraining.

What are the testing techniques that you use for model testing, name some of those?

Answer:-

For making the good/Quality of model we need to perform Model Testing some of them are

In machine Learning-

1. Hypothesis testing
2. Cross validation
3. Regularization
4. Accuracy
5. Precision
6. Recall
7. F1 score , F2 score
8. Confusion matrix
9. R2 , adjusted R2

What do you understand by sensitivity in dataset? Give example.

Sensitivity:- it is the basically a measure of the proportion/percentage of actual positive cases predicted out of total Positive cases (true positive) present . Sensitivity is also termed as Recall.

$\text{Recall/Sensitivity} = (\text{true positives} / \text{all actual positives}) \text{ or } (TP)/(TP+FN)*100$

TP = how many positively predicted out of all actual Positive

FN= how many negatively predict our model but they are Positive In reality.

For eg in hospital 100 patient is present out of them 50 are really pregnant.

For eg ,Our model predict

For this TP = 45

FN =5

| | | |
|------------------|-----------|----|
| R e a l | Predicted | |
| | TP | FN |
| | FP | TN |

Answer :- Sensitivity = $(45/45+5)*100 = 90\%$ is our model sensitive .

Let's suppose you are trying to solve the classification problem; how do you decide which algorithm to use?

So for Choosing any of the classification model you need think for some important points like,

1. What is the Size of the training data. It is usually recommended to gather a good amount of data to get reliable
2. predictions Results
3. Accuracy of the output. ...
4. Speed or Training time.
5. Checking is Data is having Linearity or Non-linearity.
6. Number of features.

You need to try with all necessary algorithm that can full fill these points.

For eg. If you need to perform some Regression Problems over your data

You need look for the dataset is linear or not if linear then you can use Linear Regression over it and if not then you can use Random Forest algorithm, which is the Robust algo. It will learn the non linear relationship between data and similarly, you need to look for model which can withstand your Requirements.

Can we use Logistic regression for classification if my no. of classes are 5?

Answer:- Yes we can use but Logistic Regression is a simple but very effective classification algorithm so it is commonly used for many binary classification tasks. Logistic regression model takes a linear equation as input and use logistic function and log odds to perform a binary classification task but you want to perform Multi class Classification(class=5) then you can use One Vs Rest logistic Regression method which will Divide the whole dataset in Two part one for single class and another it will consider rest as a one class and in the same way it will do classification. For multi class classification, there are lots of algorithms are the Robust.

Let's suppose there is a company like OLA or UBER who provides service to many customers, then how will they make sure that car availability in particular region and what kind of dataset is required?

Ans:- So first of all They are using google maps api for getting the exact co-ordinates of peoples

And also the exact co-ordinate of the cab so that after getting the request for booking cab there are finding the less displacement between the cab and people by using the hamming distance formula so that they can provide the quick booking, and after reaching the destination of people the cab got vacant and the same way they are trying to compare the distance between people and cab and based on this they are doing this things.

AI Solution for architecture -- Let's suppose there is agricultural field in diff areas in India, and we know soil & weather condition is different over India, So I am trying to build system which helps me understanding what kind of treatments I will be able to apply on my crops, which crop I can grow in particular month so I can be able to maximize the benefit form the soil. Then what kind of algorithm you will use whether its ML,DL, Vision? What will be your approach and what kind of solution design you will provide?

Ans:- We can use Machine learning then this process will perform good ,first Question is to Classify the of Different region of india ,based on soil types , and after that you need to predict the weather for that specific regions of india and also you need to classify the crops for based on soil and based on time and based on weather and after you can use any Classification algorithm which will try to learn from the data model and if you enter the input data then it will be able to classify that which crop you should farm.

And if you want to use deep learning, then you need to use some ANN which is very good capacity to learn the non-linear relationship between the feature so that it can give you the best prediction

Why do we need neural networks instead of straightforward traditional computing?

Answer

Neural networks offer a different way to analyze data, and to recognize patterns within that data, than traditional computing methods. However, they are not a solution for all computing problems. Traditional computing methods work well for problems that can be well characterized. Balancing checkbooks, keeping ledgers, and keeping tabs of inventory are well defined and do not require the special characteristics of neural networks.

Traditional computers are ideal for many applications. They can process data, track inventories, network results, and protect equipment. These applications do not need the special characteristics of neural networks.

Expert systems are an extension of traditional computing and are sometimes called the fifth generation of computing. (First generation computing used switches and wires. The second generation occurred because of the development of the transistor. The third generation involved solid-state technology, the use of integrated circuits, and higher level languages like COBOL, Fortran, and "C". End user tools, "code generators," are known as the fourth generation.) The fifth-generation involves artificial intelligence.

| CHARACTERISTICS | TRADITIONAL COMPUTING (including Expert Systems) | ARTIFICIAL NEURAL NETWORKS |
|---------------------------------|--|--|
| Processing style Functions | Sequential Logically (left brained) via Rules Concepts Calculations | Parallel Gestalt (right brained) via Images Pictures Controls |
| Learning Method Applications | by rules (didactically) Accounting word processing math inventory digital communications | by example (Socratically) Sensor processing speech recognition pattern recognition text recognition |

Table 2.6.1 Comparison of Computing Approaches.

What are the different weight initialization techniques you have used?

Answer

1. Zero initialization

2. Random initialization
3. He initialization
4. Xavier initialization

Can you visualize a neural network? if yes provide name of the software we can use?

Answer

Yes, using Netron

How will you explain the training of neural networks?

Answer

To build a good Artificial Neural Network (ANN) you will need the following ingredients

Artificial Neurons (processing node) are composed of:

- (many) input neuron(s) connection(s)
- a computation unit composed of:
 - a linear function ($ax+b$)
 - an activation function
- an output

All Neurons of a given Layer are generating an Output, but they don't have the same Weight for the next Neurons Layer. This means that if a Neuron on a layer observes a given pattern it might mean less for the overall picture and will be partially or completely muted. This is what we call Weighting: a big weight means that the Input is important and of course a small weight means that we should ignore it.

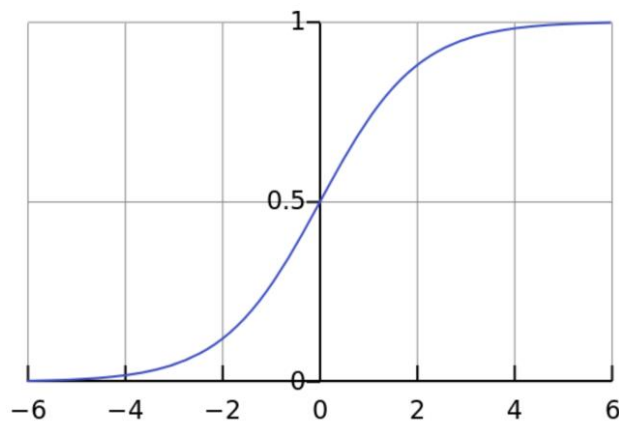
Every Neural Connection between Neurons will have an associated Weight. And this is the magic of Neural Network Adaptability: Weights will be adjusted over the training to fit the objectives we have set (recognize that a dog is a dog and that a cat is a cat). In simple terms: Training a Neural Network means finding the appropriate Weights of the Neural Connections thanks to a feedback loop called Gradient Backward propagation.

Can you please explain difference between sigmoid & tanh function.

Answer

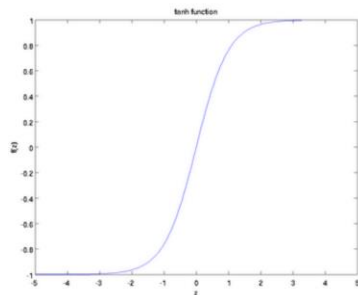
Sigmoid function and tanh function are two activation functions used in deep learning. Also, they look very similar to each other. In this article, I'd like to have a quick comparison.

$$A = \frac{1}{1+e^{-x}}$$



Sigmoid function

tanh function



$$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

Differences between them

The difference can be seen from the picture below. Sigmoid function has a range of 0 to 1, while tanh function has a range of -1 to 1. "In fact, tanh function is a scaled sigmoid function!"

$$\tanh(x) = 2 \operatorname{sigmoid}(2x) - 1$$

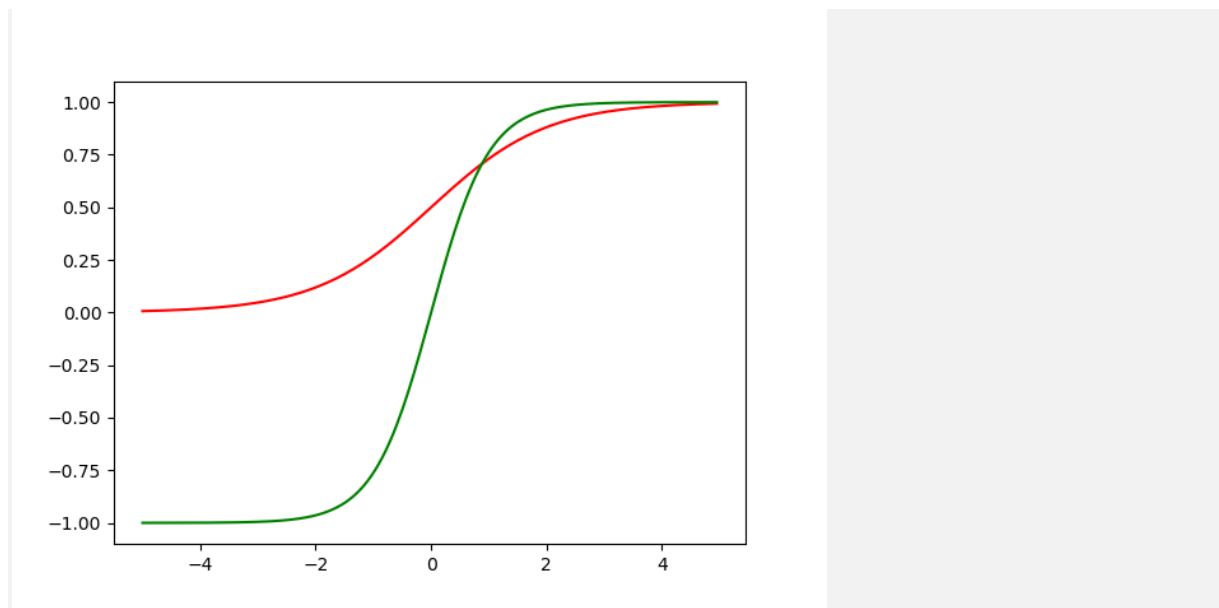
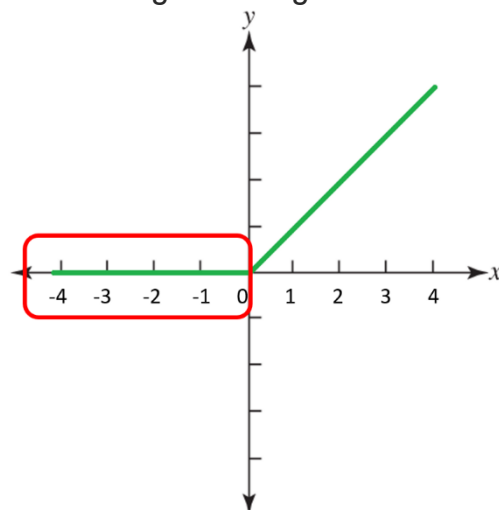


Figure :- The red one is sigmoid and the green one is the tanh function

Explain disadvantage of using ReLU function.

The dying ReLU Problem

The dying ReLU problem refers to the scenario when a large number of ReLU neurons only output values of 0. From the red outline below, we can see that this happens when the inputs are in the negative range.



Red outline (in the negative x range) demarcating the horizontal segment where ReLU outputs 0

While this characteristic is what gives ReLU its strengths (through network sparsity), it becomes a problem when a majority of the inputs to these ReLU neurons is in the

negative range. The worst case scenario is when the entire network dies, meaning that it becomes just a constant function.

When most of these neurons return output zero, the gradients fail to flow during backpropagation and the weights do not get updated. Ultimately a large part of the network becomes inactive and it is unable to learn further.

Because the slope of ReLU in the negative input range is also zero, once it becomes dead (i.e. stuck in negative range and giving output 0), it is likely to remain unrecoverable.

However, the dying ReLU problem does not happen all the time, since the optimizer (e.g. stochastic gradient descent) considers multiple input values each time. As long as NOT all the inputs push ReLU to the negative segment (i.e. some inputs are in positive range), the neurons can get to stay active, the weights can get updated, and the network can continue learning.

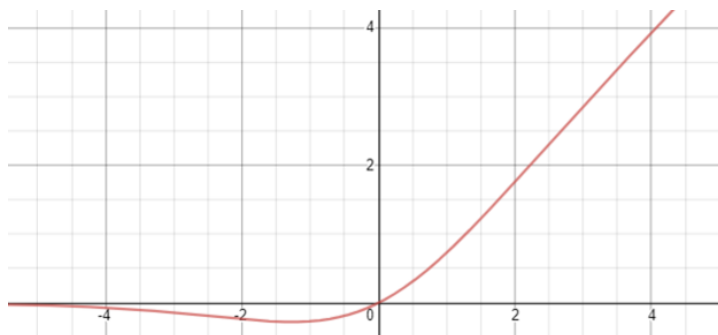
How do you select no. of layers & no. of neurons in neural network?

These are Hyperparameters so the exact the number is not defined. We take references from different research papers.

Have you ever designed any Neural network architecture by yourself?

Yes

Can you please explain SWISS Function?

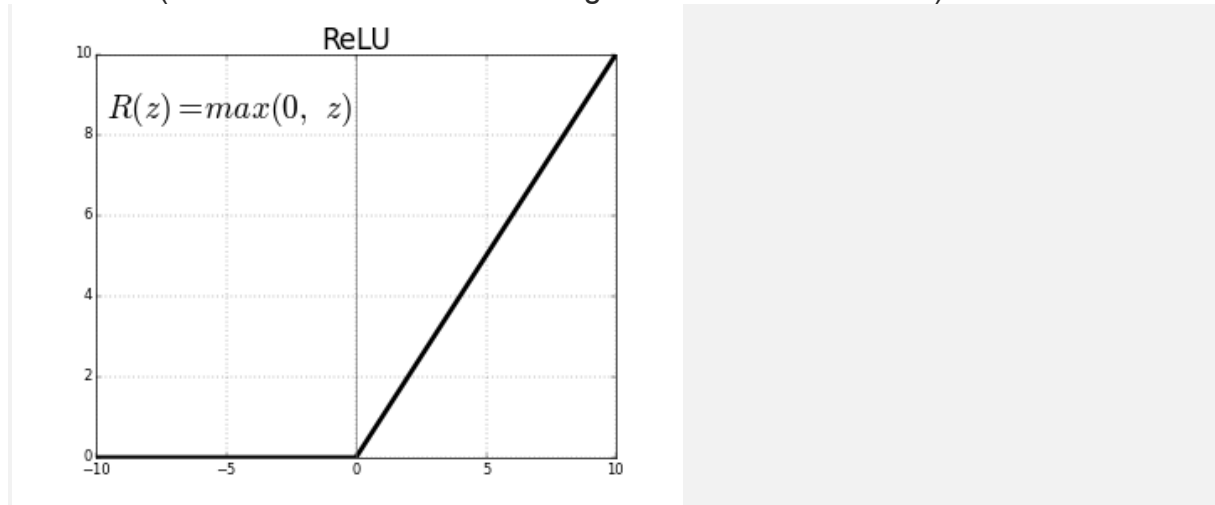


The Swish activation function

Formally stated, the Swish activation function is...

$$f(x) = x * (1 + \exp(-x))^{-1}$$

Like ReLU, Swish is bounded below (meaning as x approaches negative infinity, y approaches some constant value) but unbounded above (meaning as x approaches positive infinity, y approaches infinity). However, unlike ReLU, Swish is smooth (it does not have sudden changes of motion or a vertex):

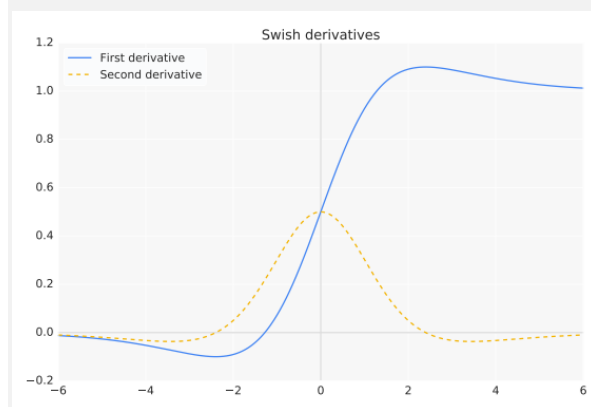


Additionally, Swish is non-monotonic, meaning that there is not always a singularly and continually positive (or negative) derivative throughout the entire function. (Restated, the Swish function has a negative derivative at certain points and a positive derivative at other points, instead of only a positive derivative at all points, like Softplus or Sigmoid.)

The derivative of the Swish function is...

$$f'(x) = f(x) + \sigma(x)(1 - f(x))$$

The first and second derivatives of Swish, plotted:



For inputs less than about 1.25, the derivative has a magnitude of less than 1.

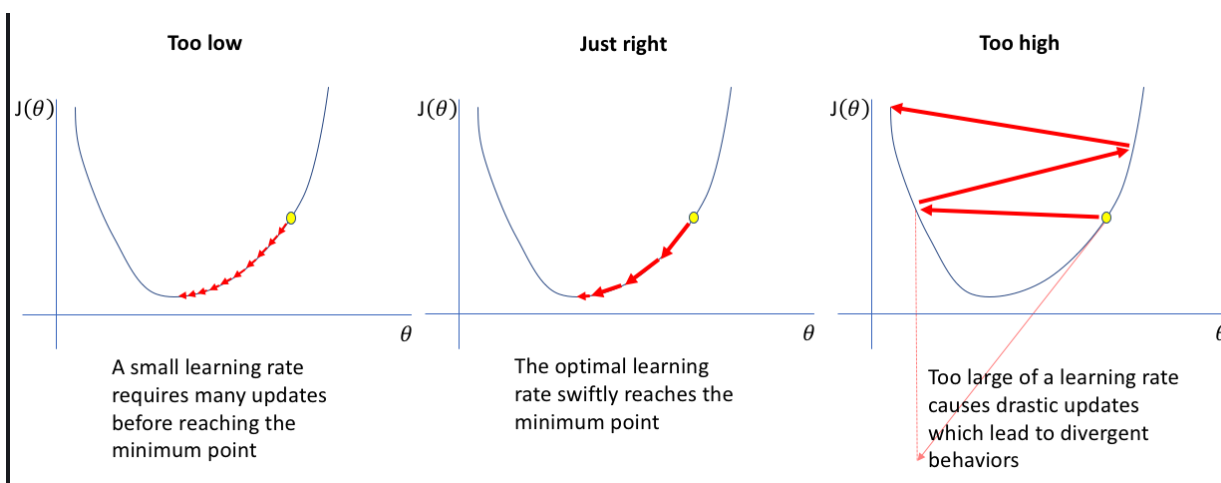
What is learning rate in laymen way and how do you control learning rate?

Answer

The learning rate is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function

One of the key hyperparameters to set in order to train a neural network is the learning rate for gradient descent. As a reminder, this parameter scales the magnitude of our weight updates in order to minimize the network's loss function.

If your learning rate is set too low, training will progress very slowly as you are making very tiny updates to the weights in your network. However, if your learning rate is set too high, it can cause undesirable divergent behavior in your loss function. I'll visualize these cases below - if you find these visuals hard to interpret, I'd recommend reading (at least) the first section in my post on gradient descent.



What is diff between batch, minibatch & stochastic gradient decent.

Answer

Batch Gradient Descent

In Batch Gradient Descent, all the training data is taken into consideration to take a single step. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters. So that's just one step of gradient descent in one epoch.

Batch Gradient Descent is great for convex or relatively smooth error manifolds. In this case, we move somewhat directly towards an optimum solution.

Stochastic Gradient Descent

In Batch Gradient Descent we were considering all the examples for every step of Gradient Descent. But what if our dataset is very huge. Deep learning models crave for data. The more the data the more chances of a model to be good. Suppose our dataset has 5 million examples, then just to take one step the model will have to calculate the gradients of all the 5 million examples. This does not seem an efficient way. To tackle this problem we have Stochastic Gradient Descent. In Stochastic Gradient Descent (SGD), we consider just one example at a time to take a single step.

SGD can be used for larger datasets. It converges faster when the dataset is large as it causes updates to the parameters more frequently.

Mini Batch

Neither we use all the dataset all at once nor we use the single example at a time. We use a batch of a fixed number of training examples which is less than the actual dataset and call it a mini-batch. Doing this helps us achieve the advantages of both the former variants we saw.

So, when we are using the mini-batch gradient descent we are updating our parameters frequently as well as we can use vectorized implementation for faster computations.

What do you understand by batch size while training Neural N/w with example

Answer

The batch size is a hyperparameter that defines the number of samples to work through before updating the internal model parameters.

Think of a batch as a for-loop iterating over one or more samples and making predictions. At the end of the batch, the predictions are compared to the expected output variables and an error is calculated. From this error, the update algorithm is used to improve the model, e.g. move down along the error gradient.

A training dataset can be divided into one or more batches.

When all training samples are used to create one batch, the learning algorithm is called batch gradient descent. When the batch is the size of one sample, the learning algorithm is called stochastic gradient descent. When the batch size is more than one sample and less than the size of the training dataset, the learning algorithm is called mini-batch gradient descent.

- Batch Gradient Descent. Batch Size = Size of Training Set

- Stochastic Gradient Descent. Batch Size = 1
- Mini-Batch Gradient Descent. $1 < \text{Batch Size} < \text{Size of Training Set}$
In the case of mini-batch gradient descent, popular batch sizes include 32, 64, and 128 samples.

Explain 5 best optimizer you know with mathematical explanation.

Answer

Stochastic Gradient Descent

It's a variant of Gradient Descent. It tries to update the model's parameters more frequently. In this, the model parameters are altered after computation of loss on each training example. So, if the dataset contains 1000 rows SGD will update the model parameters 1000 times in one cycle of dataset instead of one time as in Gradient Descent.

$\theta = \theta - \alpha \cdot \nabla J(\theta; x(i); y(i))$, where $\{x(i), y(i)\}$ are the training examples.

As the model parameters are frequently updated parameters have high variance and fluctuations in loss functions at different intensities.

Advantages:

1. Frequent updates of model parameters hence, converges in less time.
2. Requires less memory as no need to store values of loss functions.
3. May get new minima's.

Disadvantages:

1. High variance in model parameters.
2. May shoot even after achieving global minima.
3. To get the same convergence as gradient descent needs to slowly reduce the value of learning rate.

Mini-Batch Gradient Descent

It's best among all the variations of gradient descent algorithms. It is an improvement on both SGD and standard gradient descent. It updates the model parameters after every batch. So, the dataset is divided into various batches and after every batch, the parameters are updated.

$\theta = \theta - \alpha \cdot \nabla J(\theta; B(i))$, where $\{B(i)\}$ are the batches of training examples.

Advantages:

1. Frequently updates the model parameters and also has less variance.
2. Requires medium amount of memory.

All types of Gradient Descent have some challenges:

1. Choosing an optimum value of the learning rate. If the learning rate is too small than gradient descent may take ages to converge.
2. Have a constant learning rate for all the parameters. There may be some parameters which we may not want to change at the same rate.
3. May get trapped at local minima.

Adagrad

One of the disadvantages of all the optimizers explained is that the learning rate is constant for all parameters and for each cycle. This optimizer changes the learning rate. It changes the learning rate ' η ' for each parameter and at every time step ' t '. It's a type second order optimization algorithm. It works on the derivative of an error function.

$$g_{t,i} = \nabla_{\theta} J(\theta_{t,i}),$$

A derivative of loss function for given parameters at a given time t .

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i}.$$

Update parameters for given input i and at time/iteration t

η is a learning rate which is modified for given parameter $\theta(i)$ at a given time based on previous gradients calculated for given parameter $\theta(i)$.

We store the sum of the squares of the gradients w.r.t. $\theta(i)$ up to time step t , while ϵ is a smoothing term that avoids division by zero (usually on the order of $1e-8$). Interestingly, without the square root operation, the algorithm performs much worse.

It makes big updates for less frequent parameters and a small step for frequent parameters.

Advantages:

1. Learning rate changes for each training parameter.
2. Don't need to manually tune the learning rate.
3. Able to train on sparse data.

Disadvantages:

1. Computationally expensive as a need to calculate the second order derivative.
2. The learning rate is always decreasing results in slow training.

AdaDelta

It is an extension of AdaGrad which tends to remove the decaying learning Rate problem of it. Instead of accumulating all previously squared gradients, Adadelta limits the window of accumulated past gradients to some fixed size w . In this exponentially moving average is used rather than the sum of all the gradients.

$$\mathbb{E}[g^2](t) = \gamma \cdot \mathbb{E}[g^2](t-1) + (1-\gamma) \cdot g^2(t)$$

We set γ to a similar value as the momentum term, around 0.9.

$$\mathbb{E}[g^2]_t = \gamma \mathbb{E}[g^2]_{t-1} + (1 - \gamma) g_t^2,$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\mathbb{E}[g^2]_t + \epsilon}} \cdot g_t.$$

Update the parameters

Advantages:

1. Now the learning rate does not decay and the training does not stop.

Disadvantages:

1. Computationally expensive.

Adam

Adam (Adaptive Moment Estimation) works with momentums of first and second order. The intuition behind the Adam is that we don't want to roll so fast just because we can jump over the minimum, we want to decrease the velocity a little bit for a careful search. In addition to storing an exponentially decaying average of past squared gradients like AdaDelta, Adam also keeps an exponentially decaying average of past gradients $M(t)$.

$M(t)$ and $V(t)$ are values of the first moment which is the Mean and the second moment which is the uncentered variance of the gradients respectively.

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

First and second order of momentum

Here, we are taking mean of $M(t)$ and $V(t)$ so that $E[m(t)]$ can be equal to $E[g(t)]$ where, $E[f(x)]$ is an expected value of $f(x)$.

To update the parameter:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

Update the parameters

The values for β_1 is 0.9 , 0.999 for β_2 , and $(10 \times \exp(-8))$ for ' ϵ '.

Advantages:

1. The method is too fast and converges rapidly.
2. Rectifies vanishing learning rate, high variance.

Disadvantages:

Computationally costly.

Can you build Neural network without using any library? If yes, prove it.

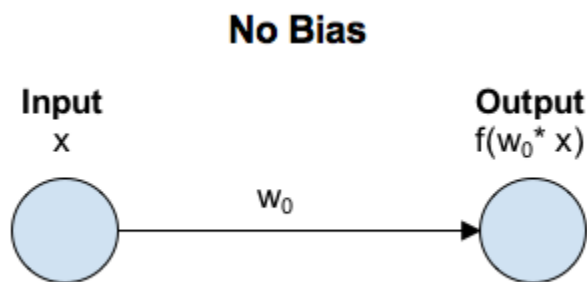
Answer

Notebook Link: - https://colab.research.google.com/drive/1iNt4eKU6PjDG_ygv-_FhcuBhumel-v3B

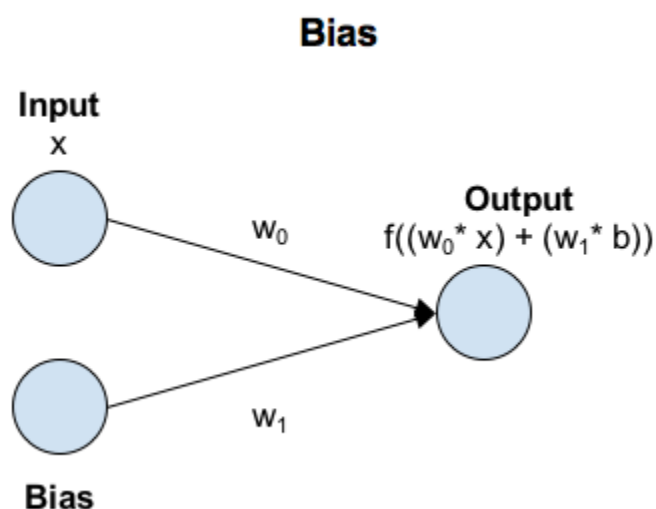
What is use of biases in neural network?

Answer

The activation function in Neural Networks takes an input 'x' multiplied by a weight 'w'. Bias allows you to shift the activation function by adding a constant (i.e. the given bias) to the input. Bias in Neural Networks can be thought of as analogous to the role of a constant in a linear function, whereby the line is effectively transposed by the constant value.



In a scenario with no bias, the input to the activation function is 'x' multiplied by the connection weight 'w₀'.



In a scenario with bias, the input to the activation function is 'x' times the connection weight 'w₀' plus the bias times the connection weight for the bias 'w₁'. This has the effect of shifting the activation function by a constant amount (b * w₁).

How do you do hyper-parameter tuning for neural network

Answer

1. Step 1 – Deciding on the network topology
2. Step 2 – Adjusting the learning rate. ...
3. Step 3 – Choosing an optimizer and a loss function. ...
4. Step 4 – Deciding on the batch size and number of epochs. ...
5. Step 5 – Random restarts.
6. Step 6 - Define the Input Shape
7. Step 7- Choose the Right activation function
8. Step 8 - Choosing the no of kernels and layers in CNN

What kind of regularization you used wrt neural network.

L2 Parameter Regularization

This regularization is popularly known as weight decay. This strategy drives the weights closer to the origin by adding the regularization term omega which is defined as:

$$\Omega(\theta) = \frac{1}{2} \|w\|_2^2$$

This technique is also known as ridge regression or Tikhonov regularization.

L1 Regularization

Here the regularization term is defined as:

$$\Omega(\theta) = \|w\|_1 = \sum_i |w_i|,$$

Dataset Augmentation

The best and easiest way to make a model generalize is to train it on a large amount of data but mostly we are provided with limited data. One way is to create fake data and add it to our training dataset, for some domains this is fairly straightforward and easy.

Noise Robustness

Noise is often introduced to the inputs as a dataset augmentation strategy. the addition of noise with infinitesimal variance at the input of the model is equivalent to imposing a penalty on the norm of the weights. Noise injection is much more powerful than simply shrinking the parameters, especially when the noise is added to the hidden units.

Early Stopping of Training

When training a large model on a sufficiently large dataset, if the training is done for a long amount of time rather than increasing the generalization capability of the model, it increases the overfitting. As in the training process, the training error keeps on reducing but after a certain point, the validation error starts to increase hence signifying that our model has started to overfit.

Dropout

Dropout is a computationally inexpensive but powerful regularization method, dropout can be thought of as a method of making bagging practical for ensembles of very many large neural networks. The method of bagging cannot be directly applied to large neural networks as it involves training multiple models, and evaluating multiple models on each test example. since training and evaluating such networks is costly in terms of runtime and memory, this method is impractical for neural networks.

Bagging

Bagging or bootstrap aggregating is a technique for reducing generalization error by combining several models. The idea is to train several different models separately, then have all of the models vote on the output for test examples. This is an example of a general strategy in machine learning called model averaging. Techniques employing this strategy are known as ensemble methods. This is an efficient method as different models don't make the same types of errors.

What are the libraries you have used for neural network implementation?

Answer

Keras Tuner, Optuna, HyperOPT, Tune

What do you understand by custom layer and a custom model?

Answer

Reference Link :-
https://keras.io/guides/making_new_layers_and_models_via_subclassing

How do you implement differentiation using TensorFlow or Pytorch library?

Answer

Tensorflow :- <https://jonathan-hui.medium.com/tensorflow-automatic-differentiation-autodiff-1a70763285cb>

Notebook Link :-
https://colab.research.google.com/github/tensorflow/tensorflow/blob/r1.9/tensorflow/contrib/eager/python/examples/notebooks/automatic_differentiation.ipynb

Pytorch

Using autograd to Find and Solve a Derivative

First, it should be obvious that we have to represent our original function in Python as such:

$$y = 5x^4 + 3x^3 + 7x^2 + 9x - 5$$

```
import torch

x = torch.autograd.Variable(torch.Tensor([2]),requires_grad=True)
y = 5*x**4 + 3*x**3 + 7*x**2 + 9*x - 5

y.backward()
x.grad
```

Line by line, the above code:

- imports the torch library
- defines the function we want to compute the derivative of
- defines the value (2) we want to compute the derivative with regard to as a PyTorch Variable object and specifies that it should be instantiated in such a way that it tracks where in the computation graph it connects to in order to perform differentiation by the chain rule (requires_grad)
- uses autograd's `backward()` to compute the sum of gradients, using the chain rule
- outputs the value stored in the x tensor's grad attribute, which, as shown below `tensor([233.])`

This value, 233, matches what we calculated by hand, above.

What is meaning of epoch in simple terms?

Answer

An epoch is a term used in machine learning and indicates the number of passes of the entire training dataset the machine learning algorithm has completed. Datasets are usually grouped into batches (especially when the amount of data is very large).

The number of epochs is a hyperparameter that defines the number times that the learning algorithm will work through the entire training dataset.

One epoch means that each sample in the training dataset has had an opportunity to update the internal model parameters. An epoch is comprised of one or more batches. For example, as above, an epoch that has one batch is called the batch gradient descent learning algorithm.

You can think of a for-loop over the number of epochs where each loop proceeds over the training dataset. Within this for-loop is another nested for-loop that iterates over each batch of samples, where one batch has the specified "batch size" number of samples.

What do you understand by a TensorFlow record?

Answer

The TFRecord format is a simple format for storing a sequence of binary records. Protocol buffers are a cross-platform, cross-language library for efficient serialization of structured data. Protocol messages are defined by .proto files, these are often the easiest way to understand a message type.

More Depth :- https://www.tensorflow.org/tutorials/load_data/tfrecord

Explain the technique for doing data augmentation in deep learning

Answer

Data augmentation is the technique of increasing the size of data used for training a model. For reliable predictions, the deep learning models often require a lot of training data, which is not always available. Therefore, the existing data is augmented in order to make a better generalized model.

Although data augmentation can be applied in various domains, it's commonly used in computer vision. Some of the most common data augmentation techniques used for images are:

- Position augmentation
- Scaling

- Cropping
- Flipping
- Padding
- Rotation
- Translation
- Affine transformation
- Color augmentation
- Brightness
- Contrast
- Saturation
- Hue

List down diff CNN network you heard of.

1. LeNet
2. AlexNet
3. ZFNet
4. Inception (GoogLeNet)
5. VGG
6. ResNet (MSRA)
7. ResNeXt
8. SEnet
9. PNASNet
10. EfficientNet
11. DenseNet

List down a names of object detection algorithm you know

1. RCNN
2. Fasster RCNN
3. Faster RCnn
4. Yolo
5. SSD
6. CenterNet

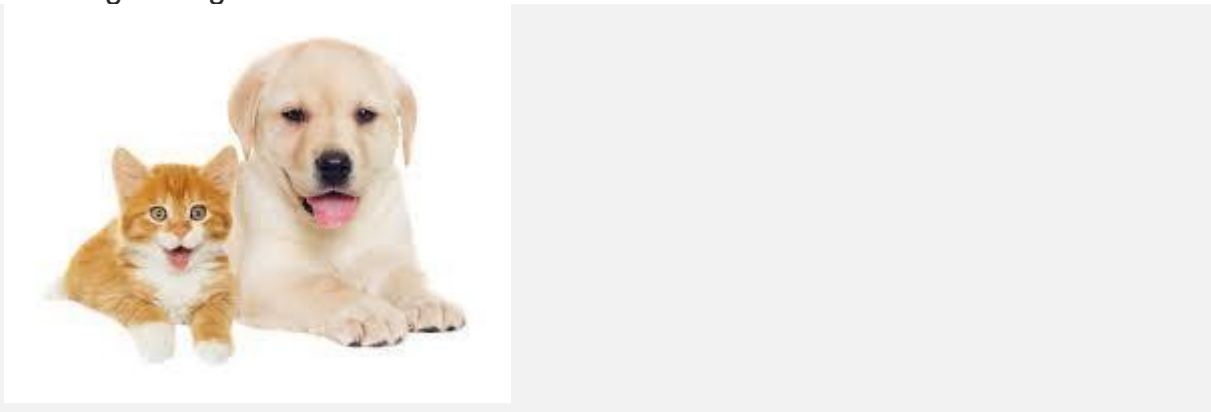
What is difference between object detection and classification?

Consider the below image:



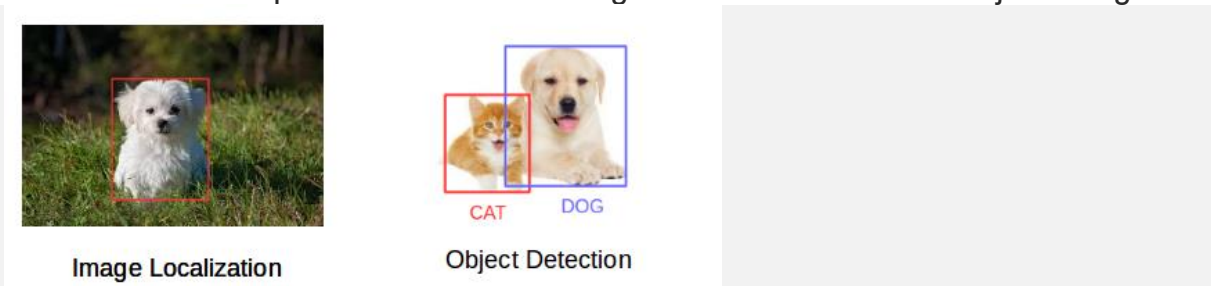
You will have instantly recognized it. It's a dog.

There's only one object here: a dog. We can easily use image classification model and predict that there's a dog in the given image. But what if we have both a cat and a dog in a single image?



We can train a multi-label classifier, in that instance. Now, there's another caveat - we won't know the location of either animal/object in the image.

Image Localization helps us to identify the location of a single object in the given image. In case we have multiple objects present, we then rely on the concept of [Object Detection](#). We can predict the location along with the class for each object using OD.



List down major tasks we perform in CNN.

1. Image Classification
2. Object Detection

3. Image Segmentaion
4. Image Captioning
5. Visual Question Answering
6. Image Generation

List down algorithms for segmentation

1. Region-Based Segmentation
 1. Threshold Segmentation
 2. Regional Growth Segmentation
2. Edge Detection Segmentation
 1. Sobel Operator
 2. Laplacian Operator
3. Segmentation based on Clustering

Reference: - <https://arxiv.org/ftp/arxiv/papers/1707/1707.02051.pdf>

What was kind of evaluation you were doing in the production environment?

What you should consider more often in a production scenario is revenue for your model, and A/B test is a must. Besides, you can check if the distribution of your prediction is consistent with that of ground truth concerning accuracy and stability for your model.

What was no. of requests (hits) your model was receiving on daily basis?

40K requests on daily basis. We have checked the log in the past and found that we usually receive 2.5k to 3.5k requests per hour for the model. Job is scheduled to process the bulk of requests from a file. The file size is usually about 370 MB to 440MB. Job is scheduled for the fixed time at night.

How you have implemented logging in the project for any failure cases?

We have implemented logging in our system. It captures each request and its status whether it was successful or failed. Even the exception of the system is captured properly. None of the detail can be missed.

We have the policy to keep a log of the previous 6 months and the remaining log we archive into backup tables.

We only keep a 13-month log including an archival log.

How you have integrated a notification (or Alarm) system for your project?

We have implemented our custom notification program. It sends notification emails to our team.

We have an application where an incident gets created automatically if any job failed or any exception occurs in our system. We have an SLA of 99.9%. Our application has

been operational it's repose time has to be less than 1.2 seconds. Else we get a notification then we investigate.

How you have implemented model monitoring?

Few of model monitoring we have implemented:

Model Input Monitoring: the set of expected values for an input feature, we can check that

a) the input values fall within an allowed set (for categorical inputs) or range (for numerical inputs) and

b) that the frequencies of each respective value within the set align with what we have seen in the past

Depending on our model configuration, we will allow certain input features to be null or not. This is something we can monitor. If features we expect generally to not be the null start to change, that could indicate a data skew or change in consumer behavior, both of which would be cause for further investigation.

Model Response time: The number of requests that can be processed data validation and cleaning was also included.

Model Versioning: Usually we retrain our model every weekend concerning new data we receive.

How you have derived the final KPI (Key Performance Indicator) for your client?

How many dashboards were there in your project?

We have 2 dashboards one is concerning application and one is for our model performance.

We only use it to work model performance-related dashboard.

Successful train rate. In our system, we do out-of-core learning because of the huge dataset. Our team is working to transform training steps in the production environment. We have started adopting mlops currently we have not fully transformed our application with MLOps practices within the next 6 months our team will do it. We have started the planning.

On which platform do you have productions your model?

We have deployed our model at the AKS cluster. Sometimes we need to scale our API capabilities so we are using AKS.

What kind of API you have exposed to receive data for the model?.

We have a user interface to upload a file. Where user uploads their file in the backend file will be uploaded to s3 bucket. And we did the configuration in such a way that it can pick the file from s3 buck and it will start the validation file. If the file will be validated successfully. We transform data for prediction and again we upload the file at the s3 bucket. We have a UI where users can see their prediction file generate for the

uploaded file. Users will get a notification that the file is available you can download the file.

What was the size of your final production environment (system configuration)?

We have 64 GB RAM.

We have GPU and CPU in our system. We have 5 TB of Hard disk.

What and all Databases you have used in the project?

The application uses a Microsoft SQL Server. But logging related to our machine learning operation we used MongoDB.

What kind of optimization you have done in your project, to what depth & explain the example.

Can you please talk about complete team structure and team size?

Our team was divided into MLOps, developers, data engineers, data scientists. 24 members were there in my team.

What was the duration of your complete project?

My project was all about 9 months.

What was your day-to-day responsibility in the last 2 months?

We usually try various random experiments and trying to analyze changes in data and estimate whether we should incorporate another mechanism to transform our data to get a more generalized prediction. Our meeting is scheduled with the client on weekly basis. If we want to highlight anything.

Previous it was daily as we were working proactively client decided to switch on weekly basis.

What kind of change request you have been receiving after your productions project

What kind of testing you have done in development, UAT, pre-pod, and prod?

Stress testing. Integration testing. Of course model evaluation is the core part of testing.

We tested API. We also simulated malicious requests to check robustness.

Have you used some of the predefined AI-OPS pipelines if yes explain.

We implement AIOps not completed but you can 60%. We have done the data version by ourself. We develop logic for that. We have deployed our application AKS.

We did Data Version COntr0, CI and CD.

CML was not entirely implementing. We have to manually trigger training if it's required or Schedule the training at a certain time.

Who has implemented AI-OPS in your project?

AIOps team has done. 2 person was in AIOps team who designed the whole workflow. We have used github and dvc and mlflow and circle ci actions.

What was the OPS stack you have been using?

GITHUB, GIT, DVC MLFLOW, CIRCLE CI , DOCKER HUB, elastic container service

What do you understand by CI-CD & have you implemented those in your project? If yes, what was the tech

stack you used for the CI-CD pipeline?

CI: Frequent changes have to be integrated with the whole application on daily basis. We should have automated test cases to check against every checkin of code. a notification has to be sent if test cases pass or fail.

CD: Our application can be deployed at any anytime so if we want to release a new version we should not discuss what date would be best because with help of a CD we can do it on daily basis but you should be careful if it's required to meet demand

GITHUB, GIT, CIRCLECI, DOCKERHUB, and HEROKU.

What was the biggest challenge you faced in the project and how you have resolved it?

Data was huge we can not train our model on whole data at once we impletd incremental learning.

Where we have created mini-batches of the dataset and then trained our model.

Give me one scenario where you worked as a team player?

What was your overall learning from the current project?

I have seen end-to-end real work machine learning application. I have seen how team work is important to success for project delivery timeline. One of the most important like project planning and defining milestones to achieve within specific days and ensuring overtime if everything is going as per the plan if not what is the reason. What is another alternative to resolve the problem for time being. Evaluating risk and develop a strategy to encounter the worst-case scenarios.

Technically: I got familiar with cloud infrastructure and MLOps practices. Different types of orchestration toll-like airflow

How do you keep yourself updated on new technology?

I devote few hours to learn new technology on daily basis. I don't push myself for long hour continuous learning. I usually decide what tech I am interested in then I use to start exploring those.

Have you designed an architecture for this project? If yes, define a strategy wrt to your current project.

No, I haven't designed architecture for the project.

How many images you have taken to train your DL model?

80K

What is the size of the model that you have in your production system?

It was around some KB. But I am not sure what was the exact size of my model. It may be around 150 KB to 250 KB

Have you tried optimizing this Vision or DL model?

Where you have hosted your Computer Vision model?

What was your frame per second?

What is the data filtration strategy you have defined for the CV project in production?

Have you used any edge device in this project, if yes, why?

What was the name of the camera & camera quality?

What was the outcome you were generating from these devices?

Have you processed the data in the local system or the cloud? Give reason.

How many number devices do you have productions (camera, edge devices, etc.)

Let's suppose I am trying to build a solution to count the no. of the vehicle or to detect their no. plate or track their

speed. Then what is the dependency of distance, position & angle of the camera on your final model? What will

happen to your model? if we change position angle.

What was your data collection strategy in the CV project, have you received data from the client or you have

created the data? And how you have implemented it?

What is difference between Euclidian distance and Manhattan distance. Explain in simple words.

Manhattan Distance: It is also called the Taxicab distance or the City Block distance. It is the distance between two points measured along axes at right angles.

- $\text{ManhattanDistance} = \sum_{i=1}^N |x[i] - y[i]|$

Euclidean Distance: It is the straight line distance between two data points in a plane. It is calculated using Minkowski Distance formula by setting 'p' value to 2, also known as the 'L2' norm distance metric.

- EuclideanDistance = $\sqrt{\sum_{i=1}^N (x[i] - y[i])^2}$

What do you understand by feature selection, transformation, engineering and EDA & What are the steps that you have performed in each of these in detail with example.

Feature Selection: Selection of features with the highest influence on the target variable, from a set of existing features.

This can be done with various techniques: Linear Regression, Decision Trees, calculation of "importance" weights

Feature Transformation: Transformation of features in order to create new ones based on the old ones to improve the accuracy of the algorithm

A popular technique is Principal Component Analysis

Feature Engineering: Generation of features which is in a format that is difficult to analyse directly and are not directly comparable. Ex: images, time-series, etc.

What is difference between single values decomposition (SVD) and PCA? (hint: SVD is one of the way to do

PCA)

PCA: Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables. PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models.

SVD: The Singular-Value Decomposition, is a matrix decomposition method for reducing a matrix to its constituent parts in order to make certain subsequent matrix calculations simpler.

$$A = U \cdot \text{Sigma} \cdot V^T$$

Where A is the real $m \times n$ matrix that we wish to decompose, U is an $m \times m$ matrix, Sigma (often represented by the uppercase Greek letter Sigma) is an $m \times n$ diagonal matrix, and V^T is the transpose of an $n \times n$ matrix where T is a superscript.

What kind of feature transformations have you done in your last project?

Feature transformation is the process of modifying data but keeping the information.

Few thing we have done is:

Data Smoothing

Data Aggregation

Generalization

Normalization

Have you taken any external features in any of the projects from any 3rd party data?
If yes, explain that scenario.

Yes, while predicting the covid cases from different states, we had to measure the oxygen production capacity of that state. If not then the nearest oxygen producer state. For this purpose, we had used 3rd party data.

If your model is overfitted, what will you do next?

Overfitting of the model means Low bias with High variance

There are the number of techniques available to handle overfitting:

Cross-validation: This is done by splitting your dataset into 'test' data and 'train' data. Build the model using the 'train' set. The 'test' set is used for in-time validation.

Regularization: It regularizes or shrinks the coefficient estimates towards zero.

Early stopping: This prevents the model from memorizing the dataset.

Pruning: This technique applies to decision trees.

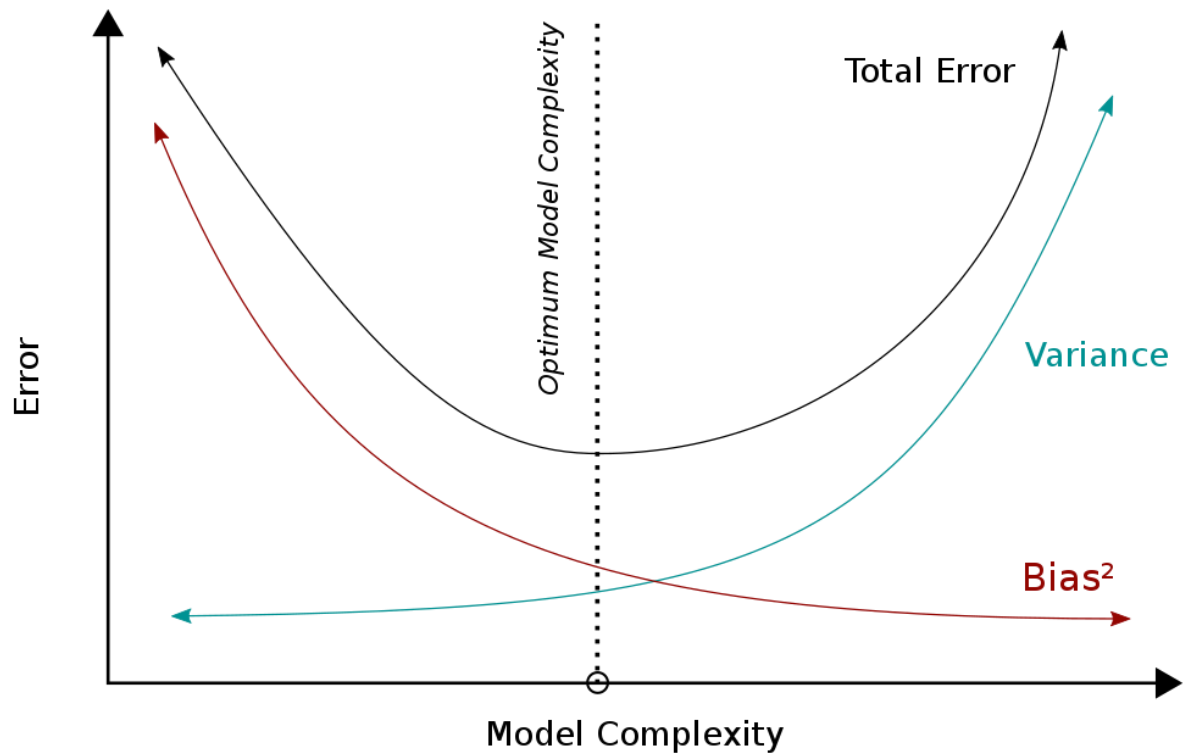
Pre-pruning: Stop 'growing' the tree earlier before it perfectly classifies the training set.

Post-pruning: This allows the tree to 'grow', perfectly classify the training set and then post prune the tree.

Dropout: This is a technique where randomly selected neurons are ignored during training.

Explain me the bias-variance trade-off.

If our model is too simple and has very few parameters, then it may have high bias and low variance. But, on the other hand, if our model has a large number of parameters, it will have high variance and low bias. So we need to find an excellent balance without overfitting and underfitting the data.



What steps would you take to improve accuracy of your model? At-least mention 5 approach. And justify

why would you choose those approach

Few steps to improve the accuracy of a model:

Add more data: Adding more data will help to learn the model in a good way instead of relying on assumptions and weak correlations.

Handle missing and outlier values: The presence of missing and outlier values often reduces the accuracy of a model or leads to a biased model because we don't analyse the relationship with another variable correctly. So it is essential to handle missing and outlier value.

Feature Engineering: It will help to extract more information from existing data. It may have high ability to explain the variance in the training data.

Feature Selection: It will help to find the best attribute which better explains the relationship of independent variable with target variable.

Multiple algorithms: Some algorithm is better suited to a particular type of dataset than other. Hence we should apply all relevant models and check the performance.

Algorithm Tuning: The objective of parameter tuning is to find the optimum value for each parameter to improve the model's accuracy.

Explain process of feature engineering in context of text categorization.

- Language detection: Understand which natural language data is in.
- Text preprocessing: Preparing raw data to make it suitable for machine learning model. Ex: text cleaning, stopword removal, stemming and lemmatization
- Length analysis: It's important to have a look at the length of the text because it's an easy calculation that can give a lot of insights.
- Sentiment analysis: determine whether a text is positive or negative.
- Named-Entity recognition: It is a process to tag text with pre-defined categories such as person names, organizations, locations.
- Word frequency: find the importance of single word by computing the n-gram frequency.
- Word vectors: transform a word into numbers.

- Topic modeling: extract the main topics from corpus.

Explain vectorization and hamming distance.

Vectorization: It is a technique by which we can make our code execute very fast. It takes multiple iterative operations among data points and turn them into matrix operation. Matrix operations are fast, they can be parallelize to some extent

Hamming distance: Hamming distance is a metric for comparing two binary data strings. While comparing two binary strings of equal length, Hamming distance is the number of bit positions in which the two bits are different.

The Hamming distance between two strings, a and b is denoted as $d(a,b)$.

Can you please explain chain rule and its use?

Suppose cost is calculated as follows, the input is x and the target value is y,

$$f' = f(x)$$

$$g' = g(x)$$

$$y' = k(g')$$

$$\text{cost} = \text{criterion}(y, y')$$

If you want to calculate $d(\text{cost}) / d(x)$, x can be a number, a vector, or a matrix. You can calculate $d(f') / d(x) \times d(g') / d(f') \times d(\text{cost}) / y'$ to get $d(\text{cost}) / d(x)$. In machine learning, the three functions here, f, g, k represent different mappings, and the criterion is also understood as a mapping, except that the input here adds the target value y. The x here represents the input data, but the meaning of the input value is not significant because we can't change the data to make our target cost smaller. The

actual situation is to change the variable contained in each map. The variables are derived. For example, if you use wf to represent the variable in function f , you can now calculate the derivative of cost to wf . You can calculate it as follows. Before displaying the calculation method, rewrite the previous expression here. Include wf ,

$$f' = f(x, wf)$$

$$g' = g(f')$$

$$y' = k(g')$$

$$\text{cost} = \text{criterion}(y, y')$$

$$d(\text{cost})/d(wf) = d\{f\}/d(wf) * d(g')/d(g') * d(y')/d(g') * d(\text{cost})/y'$$

This is from the chain rule of calculus.

What is the difference between correlation and covariance?

| Covariance | Correlation |
|--|--|
| Covariance is a measure to indicate the extent to which two random variables change in tandem. | Correlation is a measure used to represent how strongly two random variables are related to each other. |
| Covariance is nothing but a measure of correlation. | Correlation refers to the scaled form of covariance. |
| Covariance indicates the direction of the linear relationship between variables. | Correlation on the other hand measures both the strength and direction of the linear relationship between two variables. |
| Covariance can vary between $-\infty$ and $+\infty$ | Correlation ranges between -1 and +1 |
| Covariance is affected by the change in scale. | Correlation is not influenced by the change in scale. |
| Covariance assumes the units from the product of the units of the two variables. | Correlation is dimensionless, i.e. It's a unit-free measure of the relationship |

| | |
|---|--|
| | between variables. |
| Covariance of two dependent variables measures how much in real quantity (i.e. cm, kg, liters) on average they co-vary. | The correlation of two dependent variables measures the proportion of how much on average, these variables vary w.r.t one another. |
| Covariance is zero in case of independent variables | Independent movements do not contribute to the total correlation. |

What are the sampling techniques you have used in your project?

Probability Sampling:

Simple random sampling

Stratified sampling

Systematic sampling

Cluster sampling

Multi stage sampling

Non-Probability Sampling:

Convenience Sampling

Purposive Sampling

Quota sampling

Referral/Snowball sampling

Have you ever used Hypothesis testing in your last project, if yes, explain How?

Hypothesis testing is a statistical method that is used in making statistical decision using experimental data. It is basically an assumption that we make about population parameter.

Ex: In a heart disease prediction project we took the hypothesis that Depression increases the risk for coronary heart disease in established diabetes.

In which case you will use naïve Bayes classifier and decision tree separately?

What is the adv & disadvantage of naïve Bayes classifier, explain

In case of numerical data what is naïve Bayes classification equation you will use?

Naive Bayes is used a lot in robotics and computer vision, and does quite well with those tasks. Decision trees perform very poorly in those situations.

Decision trees are neat because they tell you what inputs are the best predictors of the outputs. Decision trees can often guide you to find a statistical relationship between a given input to the output and how strong that relationship is.

Advantages

- This algorithm works quickly and can save a lot of time.
- Naive Bayes is suitable for solving multi-class prediction problems.
- If its assumption of the independence of features holds true, it can perform better than other models and requires much less training data.
- Naive Bayes is better suited for categorical input variables than numerical variables.

Disadvantages

- Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. This limits the applicability of this algorithm in real-world use cases.
- This algorithm faces the 'zero-frequency problem' where it assigns zero probability to a categorical variable whose category in the test data set wasn't available in the training dataset. It would be best if you used a smoothing technique to overcome this issue.

- Its estimations can be wrong in some cases, so you shouldn't take its probability outputs very seriously.

Give me scenario where I will be able to use a boosting classifier and regressor?

Boosting can be used for regression as well as for classification problems. However, mainly focusing on reducing bias, the base models often considered for boosting are models with low variance but high bias.

Regression analysis is used to predict a continuous dependent variable from several independent variables.

In case of Bayesian classifier what exactly it tries to learn. Define its learning procedure.

Working of Naïve Bayes' Classifier can be understood with the help of the below example:

Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posterior probability.

Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$.

$$P(c|x) = P(x|c)P(c)/P(x)$$

Above,

- $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of the predictor given class.

- $P(x)$ is the prior probability of the predictor.

Give me a situation where I will be able to use SVM instead of Logistic regression.

If the use case is we have to find if the student has passed or not,
or in a supermarket, if a customer will purchase a product or not,
in such cases we can use Logistic regression over SVM.

What do you understand by rbf kernel in SVM?

RBF kernels are the most generalized form of kernelization and is one of the most widely used kernels due to its similarity to the Gaussian distribution. The RBF kernel function for two points X_1 and X_2 computes the similarity or how close they are to each other. This kernel can be mathematically represented as follows:

$$K(X_1, X_2) = \exp\left(-\frac{\|X_1 - X_2\|^2}{2\sigma^2}\right)$$

Where,

' σ ' is the variance and our hyperparameter

$\|X_1 - X_2\|$ is the Euclidean (L_2 -norm) Distance between two points X_1 and X_2

Give me 2 scenarios where AI can be used to increase revenue of travel industry.

1. Smarter traffic light algorithms & real-time tracking can control higher and lower traffic patterns effectively, This can be applied to public transport for optimal scheduling & routing.
2. Autonomous Rail Rapid Transit is a train system that runs without rails, operating instead on a virtual painted track which the train's computer system detects and follows

What do you understand by leaf node in decision tree?

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label.

What is information gain & Entropy in decision tree?

Entropy is the measures of impurity, disorder or uncertainty in a bunch of examples.

$$\text{Entropy} = -\sum p(X) \log p(x)$$

Information gain (IG) measures how much "information" a feature gives us about the class.

$$\text{IG} = \text{entropy}(\text{parent}) - [\text{weighted average}] * \text{entropy}(\text{children})$$

Give disadvantages of using Decision tree

1. A small change in the data can cause a large change in the decision tree structure, causing instability.
2. For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
3. Decision tree often involves higher time to train the model.
4. Decision tree training is relatively expensive as the complexity and time has taken are more.
5. The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

List some of the features of random forest.

- It reduces overfitting in decision trees and helps to improve the accuracy.
- It is flexible to both classification and regression problems.
- It works well with both categorical and continuous values.
- It automates missing values present in the data.

How can you avoid overfitting in decision tree?

Pruning: Pruning refers to a technique to remove the parts of the decision tree to prevent growing to its full depth. By tuning the hyperparameters of the decision tree model one can prune the trees and prevent them from overfitting.

Pre-Pruning: The pre-pruning technique refers to the early stopping of the growth of the decision tree.

Post-Pruning: The Post-pruning technique allows the decision tree model to grow to its full depth, then removes the tree branches to prevent the model from overfitting.

Explain polynomial regression in your own way.

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + \dots + b_nx_1^n$$

It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

Explain learning mechanism of linear regression.

- Regression is a supervised machine learning technique which is used to predict continuous values.

- The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data.
- The three main metrics that are used for evaluating the trained regression model are variance, bias and error. If the variance is high, it leads to overfitting and when the bias is high, it leads to underfitting.
- Based on the number of input features and output labels, regression is classified as linear (one input and one output), multiple (many inputs and one output) and multivariate (many outputs).
- Linear regression allows us to plot a linear equation, i.e., a straight line. We need to tune the coefficient and bias of the linear equation over the training data for accurate predictions.
- The tuning of coefficient and bias is achieved through gradient descent or a cost function – least squares method.

What is the cost function in logistic regression?

The cost function in Logistic Regression is Log Loss:

Log Loss is the most important classification metric based on probabilities. It's hard to interpret raw log-loss values, but log-loss is still a good metric for comparing models. For any given problem, a lower log loss value means better predictions.

$$\text{Log Loss} = \sum (x, y) \in D - y \log [f_0](y') - (1 - y) \log [f_0](1 - y')$$

What is the error function in linear regression?

Mean squared error (MSE) is the most commonly used loss function for regression. The loss is the mean over the data of the squared differences between true and predicted values

MSE is calculated by: measuring the distance of the observed y-values from the predicted y-values at each value of x; squaring each of these distances; calculating the mean of each of the squared distances.

What is the use of implementing OLS technique wrt dataset?

In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable in the given dataset

Explain the dendrogram in your own way.

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

How do you measure quality of clusters in DBSCAN?

There are two methods to measure quality of clusters in DBSCAN which are Silhouette Method and Visual Cluster Interpretation.

Silhouette Method: This method determines the separability of clusters. To begin, an average distance between each point and all other points in a cluster is calculated. The distance between each point and each point in other clusters is then calculated. We divide by whichever average is bigger after subtracting the two averages.

In the end, we desire a high (i.e., near to 1) score, which indicates a short intra-cluster average distance (tight clusters) and a large inter-cluster average distance (clusters well separated).

Visual Cluster Interpretation: It's important to understand each cluster when you've gotten your clusters. This is usually accomplished by merging the original dataset with the clusters and viewing each cluster separately. The more prominent and obvious each cluster is, the better.

How do you evaluate DBSCAN algorithm?

We can evaluate DBSCAN algorithm with Mean Silhouette Coefficient.

The Silhouette Coefficient has bounded a range of 1 to -1. 1 is the best value, and -1 is the worst. A higher score suggests that the model's clusters are more defined and denser. Negative values typically indicate that data points have been assigned to the wrong clusters. Values close to 0 imply overlapping clusters, whereas negative values usually indicate that data points have been assigned to the wrong clusters.

The silhouette coefficient is calculated using two scores:

- a: The average distance between one data point and the rest of the data points in the same cluster.
- b: The average distance between one data point and the next closest cluster's other points.

$$s = \frac{(b - a)}{\max(b - a)}$$

Formula to calculate the Silhouette Coefficient

What do you understand by market basket analysis?

Market basket analysis is a data mining approach used by merchants to better understand customer purchase patterns and thereby enhance revenue. It entails evaluating huge data sets, such as purchase histories, to identify product groups and products that are likely to be bought together.

The introduction of electronic point-of-sale (POS) systems boosted the implementation of market basket analysis. The digital records generated by POS systems made it easier for apps to process and analyse massive volumes of purchase data when compared to handwritten records held by store owners.

Explain centroid formation technique in K Means algorithm.

A centroid is an imaginary or real location that represents the cluster's centre. By lowering the in-cluster sum of squares, each data point is assigned to one of the clusters.

To put it another way, the K-means algorithm finds k centroids and then assigns each data point to the closest cluster while keeping the centroids as small as possible. The average of the data, or determining the centroid, is what the 'means' in K-means refers to.

Have you ever used SVM regression in any of your project, If yes, why?

Yes, I had used in one of my projects because Support Vector Regression (SVR) recognises the existence of non-linearity in the data and offers a reliable prediction model.

Explain the concept of GINI Impurity.

We need to understand Entropy first for better understanding of Gini Impurity.

Entropy: Entropy helps us to build an appropriate decision tree for selecting the best splitter. The entropy of a sub split can be defined as a measure of its purity. Entropy is always between 0 and 1. This formula can be used to compute the entropy of any split.

$$H(s) = -P_{(+)} \log_2 P_{(+)} - P_{(-)} \log_2 P_{(-)}$$

Here $P_{(+)} / P_{(-)} = \% \text{ of } +ve \text{ class} / \% \text{ of } -ve \text{ class}$

Gini Impurity: Gini Impurity is a measure of the likelihood of a new instance of a random variable being incorrectly categorised if it were randomly classified using the distribution of class labels from the data set.

If the data set comprises only one class, the Gini impurity is lower bounded by 0.

$$GI = 1 - \sum_{i=1}^n (p)^2$$
$$GI = 1 - \left[(P_{(+)})^2 + (P_{(-)})^2 \right]$$

Let's suppose I have given you dataset with 100 columns how you will be able to control growth of decision tree?

First, I'll apply a dimensionality reduction on the dataset by using PCA, Interactive binning (IB). Methods. Then applying Decision tree on it and check via tree chart.

In the decision tree chart, each internal node has a decision rule that splits the data. Gini referred to as the Gini ratio, which measures the impurity of the node. You can say a node is pure when all of its records belong to the same class, such nodes known as the leaf node.

If you are using Ada-boost algorithm & if it is giving you underfitted result What is the hyperparameter tuning you will do?

Will use different parameters in hyperparameter tuning which are `base_estimator`, `estimators`, `learning_rate` and `random_state`. The variables we supply to a model before we start the modelling process are known as hyper-parameters. Let's have a look at them all.

`base_estimator`: The ensemble's model; by default, a decision tree is used.

`n_estimators`: The number of models that will be created.

`learning_rate`: reduces each classifier's contribution by this amount.

`random_state`: The seed for the random number generator, which ensures that the same random numbers are created each time.

Will do some twitching in these parameters and getting the good result.

Explain gradient boosting algorithm.

Gradient boosting is a sort of boosting used in machine learning. It is based on the assumption that when the best potential next model is coupled with prior models, the overall prediction error is minimised. To decrease error, the fundamental notion is to specify the target outcomes for the next model. How are the goals determined? The goal outcome for each case in the data is determined by how much modifying the prediction for that case affects the overall prediction error:

- If a slight change in a case's prediction results in a big reduction in error, the case's next target outcome is a high value. The error will be reduced if the new model's predictions are near to its targets.
- If a slight modification in a case's prediction generates no change in error, the case's next target result is zero. Changes to this prediction have no effect on the error.

Can we use PCA to reduce dimensionality of highly non-linear data.

No, PCA cannot handle non-linear data.

How do you evaluate performance of PCA.

Reconstruction error is one way to measure performance. Indeed, one way to think of PCA is that it reduces the amount of data on the training set. To project the points into a low-dimensional space, use PCA. Then, by projecting the low-dimensional representations back into the original, high-dimensional space, recreate the original points. The distance between the original points and their reconstructions is inversely proportional to the model's ability to capture the data's structure. This is related to PCA's reputation as a lossy data compression method. The original points can be recreated more precisely when the low-dimensional representation contains more information. The commonly used performance measure R^2 can also be calculated using reconstruction error (fraction of variance accounted for).

Have you ever used multiple dimensionality techniques in any project? if yes, give reason. If no, where can we use it?

In One of my projects was to estimate churn using the enormous data set. The enormous dimensionality of this data set, with 15K data columns, is its unique feature. Most data mining techniques are implemented column-by-column, which makes them slower and slower as the number of data columns grows. The project's first milestone was to lower the number of columns in the data collection while sacrificing the least amount of information possible. So, I used PCA (Principal Component analysis) for dimensionality reduction.

What do you understand by curse of dimensionality explain.

When working with high-dimensional data, the "Curse of Dimensionality" refers to a set of issues. The number of attributes/features in a dataset corresponds to the dataset's dimension. High dimensional data is a dataset containing a large number of attributes, usually on the order of a hundred or more. Some of the challenges that come with high-dimensional data show up while analysing or displaying the data to look for trends, and others show up when training machine learning models. The 'Curse of Dimensionality' refers to the difficulty in training machine learning models due to high dimensional data. 'Data sparsity' and 'distance concentration' are two well-known characteristics of the curse of dimensionality.

What is the difference between anomaly detection and novelty detection?

Anomaly detection (also known as outlier analysis) is a data mining step that detects data points, events, and/or observations that differ from the expected behaviour of a dataset. Atypical data might reveal significant situations, such as a technical fault, or prospective possibilities, such as a shift in consumer behaviour. Anomaly detection is increasingly automated thanks to machine learning.

Novelty detection, as the name implies, is the process of identifying new or uncommon data inside a dataset. Outliers, also known as anomalies, are frequently detected as a result of their variances from the rest of the data. However, novelty detection algorithms may need to be tweaked to look for groups or bursts of uncommon data rather than single incidences of odd data. Cluster analysis is an approach that is commonly used in bank fraud algorithms to monitor suspicious activity patterns.

Explain gaussian mixture model.

A Gaussian mixture model (GMM) is a type of probabilistic model in which all data points are generated from a mixture of finite Gaussian distributions with unknown parameters. The parameters for Gaussian mixture models are produced from a well-trained prior model using either maximum a posteriori estimation or an iterative expectation-maximization approach. When it comes to modelling data, especially data from multiple groups, Gaussian mixture models are quite effective.

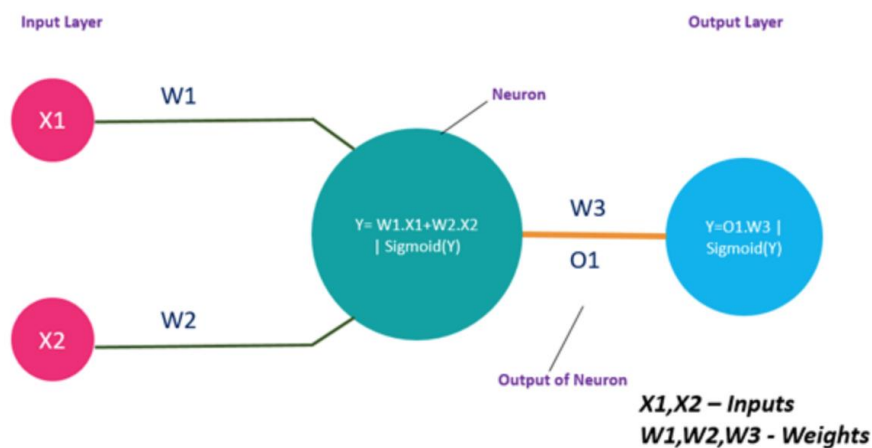
Explain neural network in terms of mathematical function.

Two principles govern the operation of a neural network.

- Forward Propagation
- Backward Propagation

Let's use an example to better comprehend these building blocks. To make the understanding obvious, I am considering a single input layer, hidden layer, and output layer.

Forward Propagation



1. Given that we have data, we'd like to use binary classification to obtain the desired result.
2. Consider a sample with features such as $X1$ and $X2$, which will be used to forecast the outcome using a series of processes.
3. Each feature is assigned a weight, with $X1$, $X2$ representing features and $W1$, $W2$ representing weights. These are fed into a neuron as input.
4. Both functions are carried out by a neuron. a) Activation b) Summation
5. All features are multiplied by their weights in the summing, and bias is totalled. ($Y = W1X1 + W2X2 + b$).
6. This summing function is used in conjunction with an Activation function. The output of this neuron is multiplied by the weight $W3$ and fed to the output layer as input.
7. Each neuron goes through the same procedure, although the activation functions in hidden layer neurons differ from those in the output layer.

We just initialised the weights at random and carried on with the process. Initializing the weights can be done in a variety of ways. But you might be wondering how these weights are updated, right??? Back propagation will be used to answer this question.

Backward Propagation

Let us return to our calculus basics, and we will update the weights using the chain rule that we learnt in school.

Chain Rule

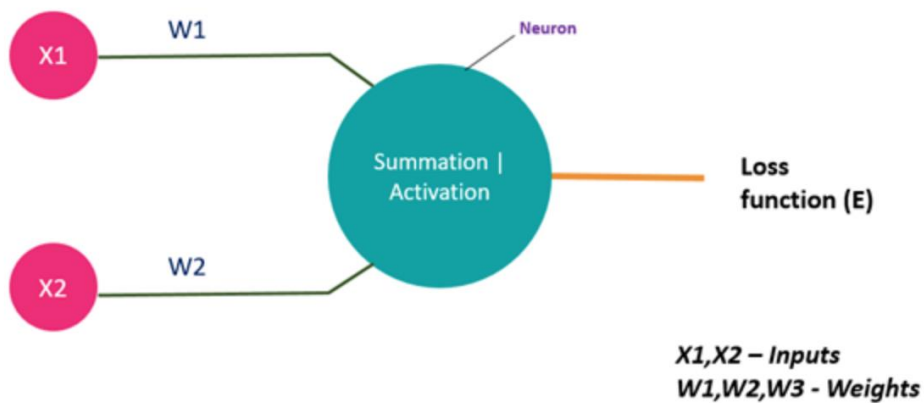
The chain rule is a method for computing the derivative of composite functions, with the number of functions in the composition influencing the number of differentiation steps required. If a composite function $f(x)$ is defined as, for example,

$$f(x) = (g \circ h)(x) = g[h(x)]$$

$$\text{then } f'(x) = g'[h(x)] \cdot h'(x)$$

Chain rule

Let's take a single neuron and apply the chain rule to it.



Our primary goal in neural networks will be to reduce error, which will necessitate updating all weights via backpropagation. We need to determine a change in weights that will result in the least amount of inaccuracy. To do so, we use the dE/dw_1 and dE/dw_2 formulas.

Considering S (Summation) = $x_1W_1 + x_2W_2$

A (Activation) = sigmoid = $e^x / (1 + e^x)$

Using Chain Rule

$$\frac{dE}{dW_1} = \frac{dE}{dA} \times \frac{dA}{dS} \times \frac{dS}{dW_1}$$

$$\frac{dE}{dW_2} = \frac{dE}{dA} \times \frac{dA}{dS} \times \frac{dS}{dW_2}$$

Change in Weights

**Forward
Propagation**



**Backward
Propagation**



Backward Propagation

After you've calculated the changes in weights concern mistake, we'll use the gradient descent process to update the weights.

$$W_{1new} = W_{1old} - \eta \frac{dE}{dW_1}$$

$$W_{2new} = W_{2old} - \eta \frac{dE}{dW_2}$$

New weights

For all samples, forward and backward propagation will continue until the error hits a minimum value.

Can you please correlate a biological neuron and artificial neuron?

The complexity of biological neural networks far exceeds that of DNNs, making understanding the representations they learn even more difficult. As a result, both machine learning and computational neuroscience face a similar problem: how can we evaluate their representations to learn how they handle complex problems? We look at how computational neuroscientists' data-analysis concepts and methodologies may be applied to analysing representations in DNNs, and how recently discovered DNN-analysis approaches can be applied to understanding representations in biological neural networks.

By figure you can analyse the difference:

| Biological Neuron | Artificial Neuron |
|--------------------|-------------------|
| Dendrites | Input |
| Cell Nucleus(Soma) | Node |
| Axon | Output |
| Synapse | Interconnections |

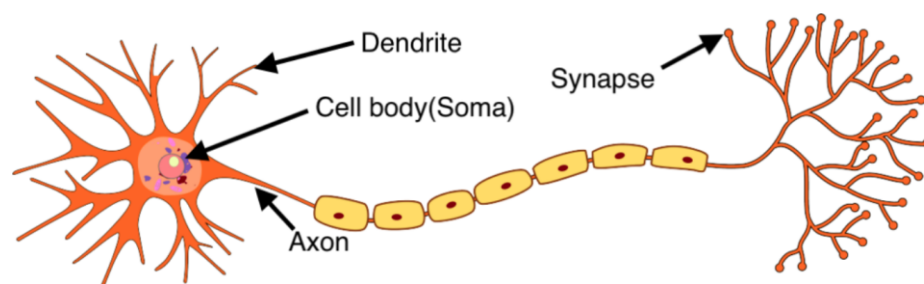
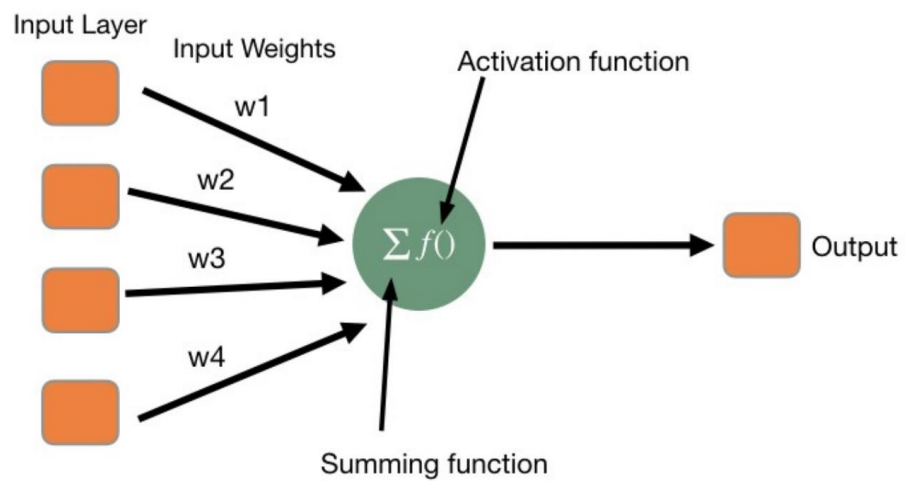


Image source: Wikimedia Commons

Perceptron



Give list of cost functions you heard of.

List of cost functions used in Neural Networks.

A cost function is a quantitative assessment of a model's fit quality: how well it reproduces the data. A cost function is a single number that represents the sum of the model's divergence from the true value for all points in the dataset.

1. Quadratic Cost function: regression

$$C = \frac{1}{2} \sum_j (\hat{y}_j - y_j^{pred})^2$$

where \hat{y}_j and y_j^{pred} are the true target value of point j , and the predicted target value respectively.

2. Cross Entropy Cost: Classification

$$C = - \sum_j (\hat{y}_j \log(y_j) + (1 - \hat{y}_j) \log(1 - y_j))$$

3. Exponential Cost

$$C = \tau \exp \left[\frac{1}{\tau} \sum_j (\hat{y}_j - y_j^{pred})^2 \right]$$

where τ is a hyper-parameter.

4. Hellinger Distance

$$C = \frac{1}{\sqrt{2}} \sum_j \left(\sqrt{\hat{y}_j} - \sqrt{y_j^{pred}} \right)^2$$

it needs to have positive values in $[0, 1]$.

5. Kullback-Leibler Divergence

Kullback-Leibler Divergence is also known as : *Information Divergence, Information Gain, Relative entropy, KLIC divergence or KL Divergence*, and is defined as:

$$D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$$

where $D_{KL}(P||Q)$ is a measure of the information lost when Q is used to approximate P .

The **cost function** using *KL Divergence* is:

$$C = \sum_j \hat{y}_j \log \frac{\hat{y}_j}{y_j^{pred}}$$

6. Generalized Kullback-Leibler Divergence

$$C = \sum_j \hat{y}_j \log \frac{\hat{y}_j}{y_j^{pred}} - \sum_j \hat{y}_j - \sum_j y_j^{pred}$$

7. Itakura-Saito Distance

$$C = \sum_j \left(\frac{\hat{y}_j}{y_j^{pred}} - \log \frac{\hat{y}_j}{y_j^{pred}} - 1 \right)$$

Can I solve problem of classification with tabular data in neural network?

Yes, OfCourse We can solve tabular data classification problem in neural network.

What do you understand by backword propagation in neural network?

The central technique by which artificial neural networks learn is backpropagation. It is doing the messenger's job to inform the neural network whether it made a mistake when making a forecast.

To propagate something (light, sound, motion, or information) means to send it in a certain direction or through a specific medium. Backpropagate is to communicate anything in response, or to send information back upstream - in this case, to fix an

error. When we talk about backpropagation in deep learning, we're talking about information transfer, and that information is related to the error that the neural network produces when it makes a guess about data. Correction is synonymous with backpropagation.

List down, time series algorithms that you know?

- Autoregression (AR)
- Moving Average (MA)
- Autoregressive Moving Average (ARMA)
- Autoregressive Integrated Moving Average (ARIMA)
- Seasonal Autoregressive Integrated Moving-Average (SARIMA)
- Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX)

How to solve TS problems using deep learning?

LSTM - Long Term Short Memory algorithm which is a variant of RNN which is a deep neural network is widely used in TS problems because of its capability to remember information for a long time. By this if there is any seasonality or pattern in the data it can remember it and give some importance of it.

Give application of TS in weather, financial, healthcare & network analysis?

Weather - Weather forecasting

Financial - Stock Price Prediction

Healthcare - Covid Cases Forecasting

Network Analysis - Predicting the computer traffic

What is diff between uptrend and downtrend in TS?

The criteria are :

The observations must be independent of each other,

The number of observations must be fixed,

The probability of success is same for each outcome

A real time example is Lottery ticket where there is only two ways either you reach success or failure

What do you understand by seasonality in TS?

Seasonality refers to the presence of some variations in trends occurring at specific regular intervals of time in a nutshell like a seasonal pattern.

For example the price of mango is cheap at summer season

What do you understand by Cyclic pattern in your TS data?

A cyclic pattern is similar to seasonality it shows some kind of patterns in the data but not in a fixed period. On an average the data exhibits this kind of a pattern at a minimum of two years

How will you find Trend in TS Data?

- A specific time window is selected
- In that window see how the data is
- For example take a span of 3 months for the sales of a company. Here in 1st the sales seems to be increasing but not a peak and in the 2nd month the sales is at a peak and 3rd month there is a drop here in the first month there we can see an uptrend and in the 3rd month we can see a downtrend

Have you implemented ARCH model in TS? If yes, give scenario?

Yes I have implemented ARCH model for a time series data in one of my projects where I had a lot of variance in the data up and down.

What is VAR (vector autoregressive) model?

Vector Auto Regressive model is a time series forecasting algorithm that can be used when we have two or more time series data which will influence each other. So here we can say that the relationship is bidirectional

What do you understand by univariate and multivariate TS Analysis?

Univariate Analysis:

As the name suggests univariate analysis means analysing the time series data which has only feature to analyse

For eg: Analysing the stock's close price

Multivariate Analysis:

As the name suggests multivariate analysis means analysing the time series data which has more than one feature to analyse straight opposite to the univariate analysis

For eg: Analysing the stock's close price along with the volume

Give example where you have created a multivariate model?

I created a Vector AutoRegressive model which is a multivariate model in one of my projects where I was supposed to use the Air Quality Index data and forecast the value for each particulate. Some of the particulates are CO, CO₂, NO₂, etc..

What do you understand by p, d, & q in ARIMA model?

p(AR) - the number of autoregressive terms

d(I) - the number of non-seasonal differences needed for stationarity

q(MA) - the number of lagged forecast errors

These are some characteristics of ARIMA(p, d, q) model

Tell me mechanism by which I can find p, d, q in ARIMA model?

p - We should make a partial autocorrelation plot and see till where there is an exponential decrease which is the shutoff point that value is taken for p

q - A similar kind of process like p is used to calculate q but the difference is here we will be using autocorrelation plot

d - It is found by seeing the seasonal difference shift

What is SARIMA and how it's different from ARIMA?

SARIMA which is known as seasonal ARIMA model and is an extension of ARIMA model. The difference between ARIMA and SARIMA models is the seasonality of the dataset. If the data is not seasonal we can use ARIMA or if the data is seasonal we can use SARIMA

What is meaning of AR, MA and I in ARIMA model?

AR - Auto Regressive

MA - Moving Average

I - Integration

ARIMA model is a model which combines both AR and MA models along with a differencing preprocessing step of sequence in purpose to make it stationary called Integration

Can we solve TS problems with transformers? What is your thought on that? why do you think in that way?

Yes we can solve Time Series problem using Transformers. But it should be the last resort at any point of time because Transformers is a heavy model so the output generating time at real time will be very slow and even for training the model is computationally intensive compared to the conventional Time Series models like ARIMA, SARIMAX or else NN's like LSTM or RNN. But if we are you going to use it we should start from just a single attention layer with a considerable learning rate

Have you ever productionised TS Based Model using LSTM? What are advantages and disadvantages?

Can we solve TS problem using Regressive algorithm, if yes, why, if no, give a reason?

Yes we can solve Time Series problem using regressive algorithms like XGBoost, Linear Regression, etc.. but this wont work for a realtime time series data because our model needs to understand the patterns in the data and capture it and give some importance it that's where the power Time Series models comes into picture. So though we can convert TS problem into a regressive problem it is not advisable to do

Explain ACF and PACF plots

A correlogram (also called Auto Correlation Function ACF Plot or Autocorrelation plot) is a visual way to show serial correlation in data that changes over time (i.e. time series data). Serial correlation (also called autocorrelation) is where an error at one point in time travels to a subsequent point in time.

The PACF plot is a plot of the partial correlation coefficients between the series and lags of itself.

Arima and Sarima, which to use when?

ARIMA is an acronym for "autoregressive integrated moving average." It's a model used in statistics and econometrics to measure events that happen over a period of time. The model is used to understand past data or predict future data in a series.

ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity of the mean function (i.e., the trend).

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA.

It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

How do you check the stationarity of the time series?

The simplest way to check for stationarity is to split your total timeseries into 2, 4, or 10 (say N) sections (the more the better), and compute the mean and variance within each section. If there is an obvious trend in either the mean or variance over the N sections, then your series is not stationary.

Hypothesis testing?

The Hypothesis Testing is a statistical test used to determine whether the hypothesis assumed for the sample of data stands true for the entire population or not. Simply, the hypothesis is an assumption which is tested to determine the relationship between two data sets.

The types of hypotheses testing:

- Simple Hypothesis.
- Complex Hypothesis.
- Working or Research Hypothesis.
- Null Hypothesis.
- Alternative Hypothesis.
- Logical Hypothesis.

- Statistical Hypothesis.

Data normalization and data standardization? And which one is prone to outliers?

Normalization typically means rescales the values into a range of $[0,1]$. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalizing the data is sensitive to outliers, so if there are outliers in the data set it is a bad practice. Standardization creates a new data not bounded (unlike normalization).

How back propagation works?

Back-propagation is just a way of propagating the total loss back into the neural network to know how much of the loss every node is responsible for, and subsequently updating the weights in such a way that minimizes the loss by giving the nodes with higher error rates lower weights and vice versa.

Backpropagation algorithm works by computing the gradient of the loss function with respect to each weight by the chain rule, computing the gradient one layer at a time.

What is Vanishing and Exploding gradients?

Vanishing: As the backpropagation algorithm advances downwards (or backward) from the output layer towards the input layer, the gradients often get smaller and smaller and approach zero which eventually leaves the weights of the initial or lower layers nearly unchanged. As a result, the gradient descent never converges to the optimum. This is known as the vanishing gradients problem.

Exploding: On the contrary, in some cases, the gradients keep on getting larger and larger as the backpropagation algorithm progresses. This, in turn, causes very large weight updates and causes the gradient descent to diverge. This is known as the exploding gradients problem.

How do you overcome vanishing and Exploding gradients?

Vanishing Gradients: The simplest solution is to use other activation functions, such as ReLU, which doesn't cause a small derivative. Residual networks are another solution, as they provide residual connections straight to earlier layers.

Exploding Gradients: A common solution to exploding gradients is to change the error derivative before propagating it backward through the network and using it to update the weights. By rescaling the error derivative, the updates to the weights will also be rescaled, dramatically decreasing the likelihood of an overflow or underflow.

How do you initialise weights to NN? And explain Adagrad and Adam initializers?

Neural network models are fit using an optimization algorithm called stochastic gradient descent that incrementally changes the network weights to minimize a loss function, hopefully resulting in a set of weights for the model that is capable of making useful predictions.

This optimization algorithm requires a starting point in the space of possible weight values from which to begin the optimization process. Weight initialization is a procedure to set the weights of a neural network to small random values that define the starting point for the optimization (learning or training) of the neural network model.

Each time, a neural network is initialized with a different set of weights, resulting in a different starting point for the optimization process, and potentially resulting in a different final set of weights with different performance characteristics.

We cannot initialize all weights to the value 0.0 as the optimization algorithm results in some asymmetry in the error gradient to begin searching effectively.

He-init and xavier Initialization differences?

The main difference for machine learning practitioners is the following:

- He initialization works better for layers with ReLU activation.
- Xavier initialization works better for layers with sigmoid activation.

Why CNN for images?

CNNs are used for image classification and recognition because of its high accuracy. The CNN follows a hierarchical model which works on building a network, like a funnel, and finally gives out a fully-connected layer where all the neurons are connected to each other and the output is processed.

How back propagation works in max pooling layer of cnn?

For the backward in a max pool layer, we pass of the gradient, we start with a zero matrix and fill the max index of this matrix with the gradient from above. On the other hand, if we tread it as an average pool layer, we need to fill each cell with the value of the gradient from above.

How do you train object detection model and deploy it?

Train Object Detection Model:

1. Collect your datasets
2. Annotate the custom images using 'labellmg'
3. Split them into train-test sets
4. Generate a TFRecord for the train-test split
5. Setup a config file
6. Train the actual model
7. Export the graph from the newly trained model
8. Bring in the frozen_inference_graph to classify in real-time

How do you check the accuracy of ocr output?

Measuring OCR accuracy is done by taking the output of an OCR run for an image and comparing it to the original version of the same text. You can then either count how many characters were detected correctly (character level accuracy), or count how many words were recognized correctly (word level accuracy).

Brief on Transformers architecture and Attention models.

Transformers are a type of neural network architecture that have been gaining popularity.

Transformers were recently used by OpenAI in their language models, and also used recently by DeepMind for AlphaStar – their program to defeat a top professional Starcraft player.

Transformers were developed to solve the problem of sequence transduction, or neural machine translation. That means any task that transforms an input sequence to an output sequence. This includes speech recognition, text-to-speech transformation, etc.

Attention models, or attention mechanisms, are input processing techniques for neural networks that allows the network to focus on specific aspects of a complex input, one at a time until the entire dataset is categorized. Attention models require continuous reinforcement or backpropagation training to be effective.

Decorators and Iterators in python

The python generators give an easy way of creating iterators. These generators instead of returning the function from the return statement use the "yield" keyword. These are the generator version of the list comprehensions.

If the function contains at least one "yield" statement, it becomes a generator function. Both the yield and return will return some value from the function.

we can implement decorators' concept in two ways: Class decorators. Function decorators. Usually, a decorator is any callable object that is used to modify the function (or) the class.

Iterators are used mostly to iterate or convert other objects to an iterator using iter() function. Generators are mostly used in loops to generate an iterator by returning all the values in the loop without affecting the iteration of the loop. Iterator uses iter() and next() functions.

Why can't we use traditional machine learning algorithms for Time series?

Time series forecasting is an important area of machine learning. It is important because there are so many prediction problems that involve a time component. However, while the time component adds additional information, it also makes time series problems more difficult to handle compared to many other prediction tasks. Time series data, as the name indicates, differ from other types of data in the sense that the temporal aspect is important. On a positive note, this gives us additional information that can be used when building our machine learning model – that not only

the input features contain useful information, but also the changes in input/output over time.

Comparing the performance of all methods, it was found that the machine learning methods were all out-performed by simple classical methods, where ETS and ARIMA models performed the best overall. This finding confirms the results from previous similar studies and competitions.

Why Vector Auto regression over LSTM's?

Vector autoregression (VAR) is a statistical model used to capture the relationship between multiple quantities as they change over time. VAR is a type of stochastic process model. VAR models generalize the single-variable (univariate) autoregressive model by allowing for multivariate time series. VAR models are often used in economics and the natural sciences.

VAR MODELING

With ARIMA we are using the past values of every variable to make the predictions for the future. When we have multiple time series at our disposal, we can also extract information from their relationships, in this way VAR is a multivariate generalization of ARIMA because it understands and uses the relationship between several inputs. This is useful for describing the dynamic behavior of the data and also provides better forecasting results.

To correctly develop a VAR model, the same classical assumptions encountered when fitting an ARIMA, have to be satisfied. We need to grant stationarity and leverage autocorrelation behaviors. These prerequisites enable us to develop a stable model. All our time series are stationary in mean and show a daily and weekly pattern.

COMBINE VAR AND LSTM

Now our scope is to use our fitted VAR to improve the training of our neural network. The VAR has learned the internal behavior of our multivariate data source adjusting the insane values, correcting the anomalous trends, and reconstructing properly the NaNs.

Our strategy involves applying a two-step training procedure. We start feeding our LSTM autoencoder, using the fitted values produced by VAR, for multi-step ahead forecasts of all the series at our disposal (multivariate output). Then we conclude the training with the raw data, in our case they are the same data we used before to fit the VAR. With our neural network, we can also combine external data sources, for example, the weather conditions or some time attributes like weekdays, hours, and months that we cyclically encode.

We hope that our neural network can learn from two different but similar data sources and perform better on our test data. When performing multiple-step training we have to take care of the Catastrophic Forgetting problem. Catastrophic forgetting is a problem faced by many models and algorithms. When trained on one task, then trained on a second task, many machine learning models “forget” how to perform the first task. This is widely believed to be a serious problem for neural networks.

To avoid this tedious problem, the structure of the entire network has to be properly tuned to provide a benefit in performance terms. From these observations, we preserve a final part of our previous training as validation.

Technically speaking the network is very simple. It's constituted by a seq2seq LSTM autoencoder which predicts the available sensors N steps ahead in the future. The training procedure is carried out using keras-hypetune. This framework provides hyperparameter optimization of the neural network structures in a very intuitive way. This is done for all three training involved (the fit on VAR fitted values, the fine-tuning fit with the raw data and the standard fit directly on the raw data)

1. What are the steps that you have followed in your last project to prepare the dataset?

Answer: The choice of data entirely depends on the problem you're trying to solve but, in my case, I have followed these 5 steps to prepare the dataset;

Step1: Gathering the data

The quantity of the data is important but not as important as the quality of it.

Step2: Handling missing data

This is one of the hardest step and handling missing data in the wrong way can cause disasters.

Step3: Taking data further with the feature extraction

Feature extraction can be a turning point. It is what makes a dataset unique. Getting insight by making relations between features is important thing.

Step4: Deciding which key factors are important

AI is able to decide which features truly affect the output and which doesn't. On the downside, The more data you give your model, it costs you money (computer power) & time. Both not always available. So, Giving your program a little help isn't always a bad idea. If you're sure that a certain feature is completely unrelated to the output, you should just disregard it altogether.

Step5: Splitting the data into training & testing sets

The data is 80-20 percent training & testing sets respectively. The 20% for the test set has engineered in a way that they're not just randomly cut out of the dataset.

2. In your last project what steps were involved in model selection procedure?

Answer: As considering the data-rich situation, the approach is selected where we randomly divide the dataset into three parts: training set, validation set, and, test set. Where training set was used to fit the models; the validation set was used to estimate the prediction error for model selection; and finally, the test set was used for assessment of the generalization error of the final chosen model.

We used k-fold cross-validation that splits the training dataset into k folds, where each example appears in a test set only once. At last, we have finalized the model based on test data score.

3. If I give you 2 columns of any dataset, what will be the steps will be involved to check the relationship between those 2 columns?

Answer: The first step will be to check the columns contain which kind of data type such as; Continuous or categorical. So that we can check correlation between categorical and numerical variables.

| | Categorical | Continuous |
|--------------------|--|--|
| Categorical | Lambda, Corrected Cramer's V | Point Biserial, Logistic Regression |
| Continuous | Point Biserial, Logistic Regression | Spearman, Kendall, Pearson |

Can you please explain 5 diff kind of strategies at least to handle missing values in dataset?

Answer:

1. Deleting Rows with missing values: Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.
2. Impute missing values with mean/median: Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.
3. Impute missing values for categorical variable: When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category.
4. Missing values imputation using k-NN:: The k nearest neighbours is an algorithm that is used for simple classification. The algorithm uses 'feature similarity' to predict the values of any new data points.
5. Imputation using Deep Learning Library (Datawig): This method works very well with categorical and non-numerical features. It is a library that learns Machine Learning models using Deep Neural Networks to impute missing values in a dataframe. It also supports both CPU and GPU for training.

What kind of diff. issues you have faced wrt your raw data? At least mention 5 issues.

Answer:

1. Getting data from multiple sources
2. Unlocking value out of Unstructured Text Data
3. Setting up the infrastructure and velocity of data
4. Adapting to different tools to collect unstructured data
5. Building a robust strategy before collecting data

What is your strategy to handle categorical dataset? Explain with example.

Answer: Categorical features have a lot to say about the dataset thus it should be converted to numerical to make it into a machine-readable format.

Two major types of categorical features are

- Nominal - These are variables which are not related to each other in any order such as colour (black, blue, green).
- Ordinal - These are variables where a certain order can be found between them such as student grades (A, B, C, D, Fail).

Encoding Categorical Variables is main approach to handle categorical dataset.

How do you define a model in terms of machine learning or in your own word?

Answer: Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.

What do you understand by k fold validation & in what situation you have used k fold cross validation?

Answer: We used k-fold cross-validation that splits the training dataset into k folds, where each example appears in a test set only once. At last, we have finalized the model based on test data score in model selection procedure.

What is meaning of bootstrap sampling? explain me in your own word.

Answer: Bootstrap Sampling is a method that involves drawing of sample data repeatedly with replacement from a data source to estimate a population parameter.

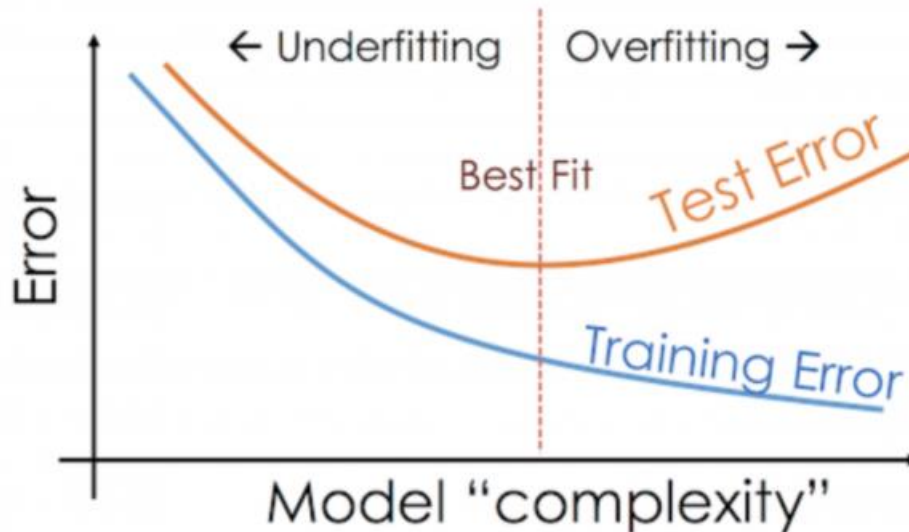
Bootstrap sampling is used in a machine learning ensemble algorithm called bootstrap aggregating (also called bagging). It helps in avoiding overfitting and improves the stability of machine learning algorithms.

In bagging, a certain number of equally sized subsets of a dataset are extracted with replacement. Then, a machine learning algorithm is applied to each of these subsets and the outputs are ensembled.

What do you understand by underfitting & overfitting of model with example?

Answer: The situation where any given model is performing too well on the training data but the performance drops significantly over the test set is called an overfitting model.

For example, non-parametric models like decision trees, KNN, and other tree-based algorithms are very prone to overfitting. These models can learn very complex relations which can result in overfitting.



On the other hand, if the model is performing poorly over the test and the train set, then we call that an underfitting model. An example of this situation would be building a linear regression model over non-linear data.

What is diff between cross validation and bootstrapping?

Answer: Bootstrapping is a technique that helps in many situations like validation of a predictive model performance, ensemble methods, estimation of bias and variance of the model. It works by sampling with replacement from the original data, and take the “not chosen” data points as test cases. We can make this several times and calculate the average score as estimation of our model performance.

In addition, Bootstrapping helps in ensemble methods as we may build a model (like a Decision tree) using each bootstrap data set and “bag” these models in an ensemble (like Random Forest) and take the majority voting for all of these models as our resulting classification.

On the other hand, cross validation is a technique for validating the model performance, and it's done by split the training data into k parts. We take $k-1$ parts as our training set and use the “held out” part as our test set. We repeat that k times differently (we hold out different part every time). Finally we take the average of the k scores as our performance estimation.

Cross validation can suffer bias or variance. if we increase the number of splits (k), the variance will increase and bias will decrease. On contrast, if we decrease (k), the bias will increase and variance will decrease. Generally 10-fold CV is used but of course it depends on the size of the training data.

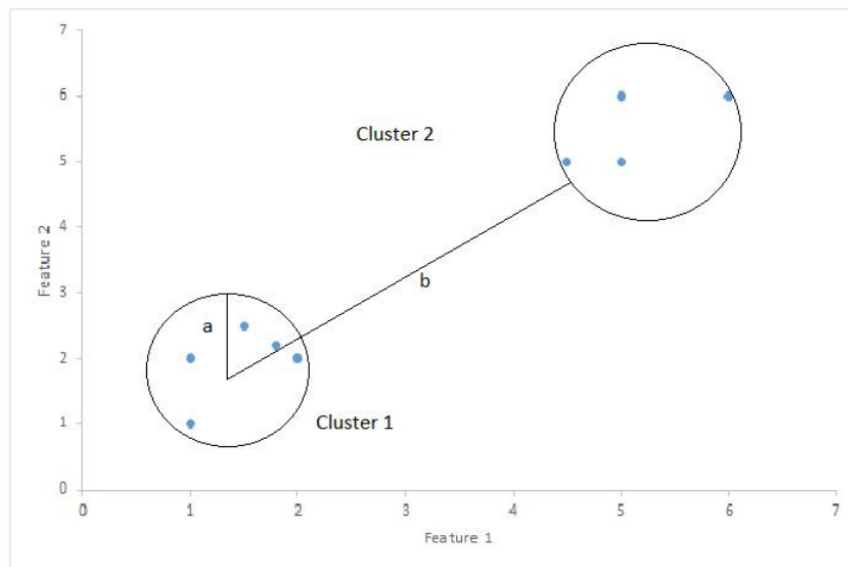
What do you understand by silhouette coefficient?

Answer: Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.



$$\text{Silhouette Score} = (b-a)/\max(a,b)$$

Where,

a= average intra-cluster distance i.e the average distance between each point within a cluster.

b= average inter-cluster distance i.e the average distance between all clusters.

What is the advantage of using ROC Score?

Answer:

- A simple graphical representation of the diagnostic accuracy of a test: the closer the apex of the curve toward the upper left corner, the greater the discriminatory ability of the test.
- Allows a simple graphical comparison between diagnostic tests
- Allows a simple method of determining the optimal cut-off values, based on what the practitioner thinks is a clinically appropriate (and diagnostically valuable) trade-off between sensitivity and false positive rate.
- Also, allows a more complex (and more exact) measure of the accuracy of a test, which is the AUC
 - The AUC in turn can be used as a simple numeric rating of diagnostic test accuracy, which simplifies comparison between diagnostic tests.
 - The AUC is non-parametric, which means it is unaffected by abnormal distributions in the population

Explain me complete approach to evaluate your regression model

Answer: There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square: R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Mean Square Error(MSE)/Root Mean Square Error(RMSE): While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

3. Mean Absolute Error(MAE): Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Give me example of lazy learner and eager learner algorithms example.

Answer:

Lazy learner:

1. Just store Data set without learning from it

2. Start classifying data when it receive Test data
3. So it takes less time learning and more time classifying data

Eager learner:

1. When it receive data set it starts classifying (learning)
2. Then it does not wait for test data to learn
3. So it takes long time learning and less time classifying data

In supervised learning Some examples are :

Lazy : K - Nearest Neighbour, Case - Based Reasoning

Eager : Decision Tree, Naive Bayes, Artificial Neural Networks

What do you understand by holdout method?

Answer: Holdout Method is the simplest sort of method to evaluate a classifier. In this method, the data set (a collection of data items or examples) is separated into two sets, called the Training set and Test set.

A classifier performs function of assigning data items in a given collection to a target category or class.

Example -E-mails in our inbox being classified into spam and non-spam.

Classifier should be evaluated to find out, it's accuracy, error rate, and error estimates. It can be done using various methods. One of most primitive methods in evaluation of classifier is 'Holdout Method'.

In the holdout method, data set is partitioned, such that - maximum data belongs to training set and remaining data belongs to test set.

What is diff between predictive modelling and descriptive modelling.

Answer:

| Basis for Comparison | Descriptive Analytics | Predictive Analytics |
|-------------------------|---|---|
| Describes | What happened in the past? By using the stored data. | What might happen in the future? By using the past data and analyzing it. |
| Process Involved | Involves Data Aggregation and Data Mining. | Involves Statistics and forecast techniques. |
| Definition | The process of finding useful and important information by analyzing the huge data. | This process involves in forecasting the future of the company, which are very useful. |
| Data Volume | It involves in processing huge data that are stored in data warehouses. Limited to past data. | It involves analyzing large past data and then predicts the future using advance techniques. |
| Examples | Sales report, revenue of a company, performance analysis, etc. | Sentimental analysis, credit score analysis, forecast reports for a company, etc. |
| Accuracy | It provides accurate data in the reports using past data. | Results are not accurate, it will not tell you exactly what will happen but it will tell you what might happen in the future. |
| Approach | It allows the reactive approach | While this a proactive approach |

How you have derived a feature for model building in your last project?

Answer: The great features that describe the structures inherent in your data.

Better features means flexibility and Better features means simpler models.

Tabular data is described in terms of observations or instances (rows) that are made up of variables or attributes (columns). An attribute could be a feature.

The idea of a feature, separate from an attribute, makes more sense in the context of a problem. A feature is an attribute that is useful or meaningful to your problem. It is an important part of an observation for learning about the structure of the problem that is being modeled.

I use “meaningful” to discriminate attributes from features. Some might not. I think there is no such thing as a non-meaningful feature. If a feature has no impact on the problem, it is not part of the problem.

Explain 5 different encoding techniques.

Answer: Since most machine learning models only accept numerical variables, preprocessing the categorical variables becomes a necessary step. We need to convert these categorical variables to numbers such that the model is able to understand and extract valuable information.

1. Label Encoding or Ordinal Encoding: We use this categorical data encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence. In Label encoding, each label is converted into an integer value. We will create a variable that contains the categories representing the education qualification of a person.
 2. One hot Encoding: We use this categorical data encoding technique when the features are nominal (do not have any order). In one hot encoding, for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category. These newly created binary features are known as Dummy variables.
 3. Dummy Encoding: Dummy coding scheme is similar to one-hot encoding. This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). In the case of one-hot encoding, for N categories in a variable, it uses N binary variables. The dummy encoding is a small improvement over one-hot-encoding. Dummy encoding uses N-1 features to represent N labels/categories.
 4. Binary Encoding: Binary encoding is a combination of Hash encoding and one-hot encoding. In this encoding scheme, the categorical feature is first converted into numerical using an ordinal encoder. Then the numbers are transformed in the binary number. After that binary value is split into different columns. Binary encoding works really well when there are a high number of categories. For example the cities in a country where a company supplies its products.
 5. Target Encoding: In target encoding, we calculate the mean of the target variable for each category and replace the category variable with the mean value. In the case of the categorical target variables, the posterior probability of the target replaces each category.
4. How do you define some features are not important for ML model? What strategy will you follow

Answer: Unnecessary features decrease training speed, decrease model interpretability, and, most importantly, decrease generalization performance on the test set.

The FeatureSelector library can be used to select important features.

The most common feature selection methods:

- Features with a high percentage of missing values: The first method for finding features to remove is straightforward: find features with a fraction of missing values above a specified threshold.

- Collinear (highly correlated) features: Collinear features are features that are highly correlated with one another. In machine learning, these lead to decreased generalization performance on the test set due to high variance and less model interpretability.
- Features with zero importance in a tree-based model: It finds features that have zero importance according to a gradient boosting machine (GBM) learning model.
- Features with low importance: The function `identify_low_importance` finds the lowest importance features that do not contribute to a specified total importance. For example, the call below finds the least important features that are not required for achieving 99% of the total importance:
- Features with a single unique value: A feature with only one unique value cannot be useful for machine learning because this feature has zero variance. For example, a tree-based model can never make a split on a feature with only one value (since there are no groups to divide the observations into).

List down at least 5 vectorization technique.

→ 5 vectorization techniques are-

1. Bag of Words.

→ BOW is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a “bag” of words because any information about the order or structure of words in the document is discarded.

2. TF-IDF

→ TF stands for Term Frequency.

$$Tf(w_a, d_n) = \frac{\text{number of times } w_a \text{ occurs in } d_n}{\text{total number of words in } d_n}$$

in any document

any word

IDF stands for Inverse Document Frequency. Where DF- Document Frequency.

$$DF = \frac{\text{Documents containing word } W}{\text{Total number of documents}}$$

$$IDF = \log\left(\frac{\text{Total number of documents}}{\text{Documents containing word } W}\right)$$

3. Word2Vec

→ Word2Vec used to group the vectors of similar words together in vector-space. It uses simple Neural-Networks for word-embeddings.

To implement Word2Vec we use two techniques-

I) Skip-Gram , II) CBOW(Continuous Bag of Words.)

4. GloVe

→ It is Similar to Word2Vec but GloVe is also creating contextual word embeddings but given the great performance of Word2Vec.

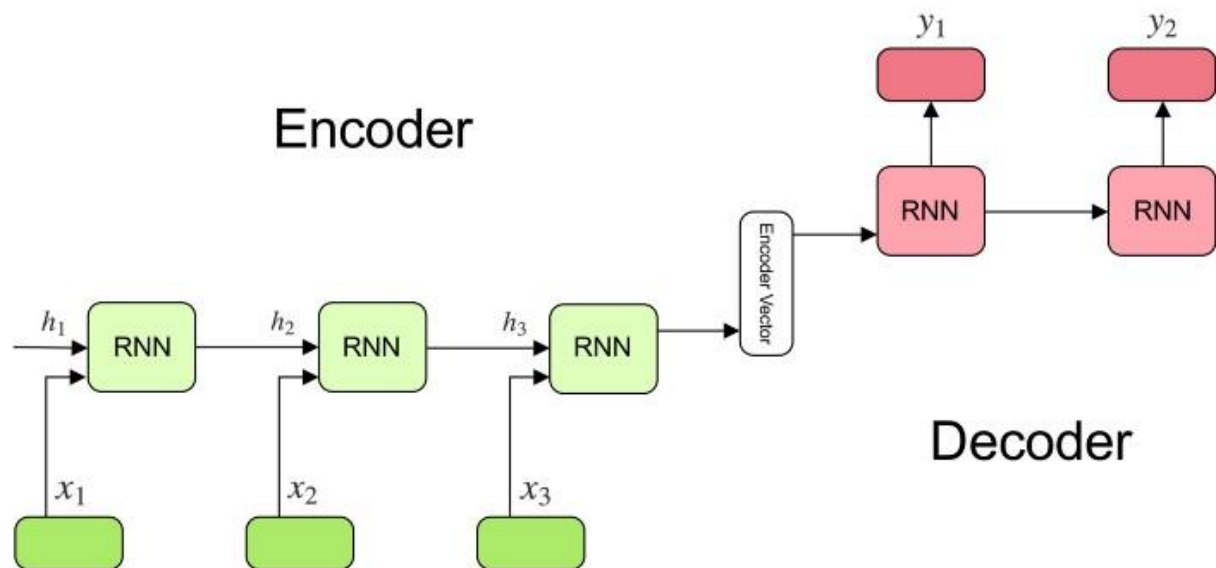
5. FastText

→ It was introduced by Facebook in 2016. It is also very much similar to Word2Vec. But it has the capability of generalizing the unknown words, which other methods can miss.

What is difference between RNN and Encoder-Decoder?

→

- A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. It can remember its input, due to an internal memory.
- Encoder is a stack of several recurrent units where each accepts a single element of the input sequence, collects information for that element and propagates it forward.
- Where Decoder is also a stack of several recurrent units where each predicts an output at a particular time step. each recurrent unit accepts a hidden state from the previous unit and produces an output as well as its own hidden state



What do you understand by attention mechanism and what is use of it ?

→ In deep learning attention mechanism is an attempt to implement perform a particular action by concentrating a few relevant things, while ignoring the other neural networks. The attention mechanism also brought an improvement over the encoder decoder-based on neural machine translation system in NLP. It helps to memorize long source sentences in neural machine translation (NMT). Rather than building a single context vector out of the encoder's last hidden state. It was used in other applications, including Computer Vision , speech processing, etc.

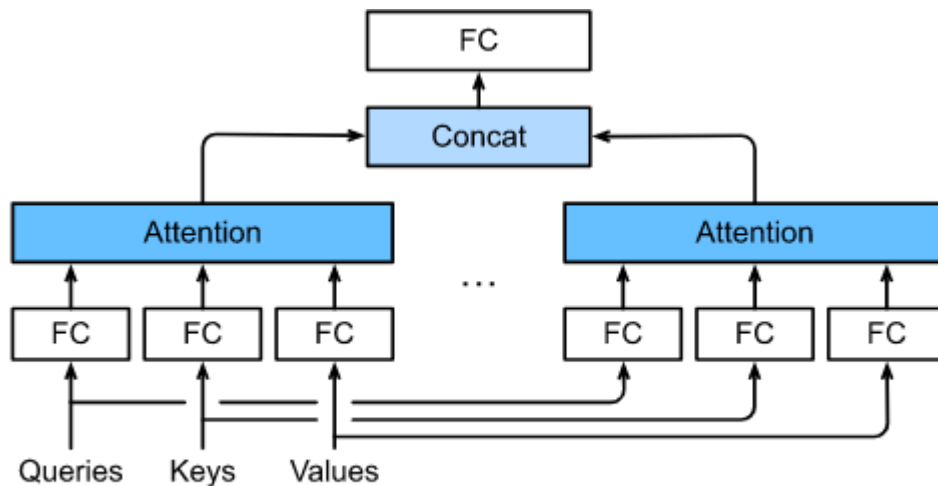
Have you read a research paper Attention you all need? If not, then why you are claiming you know NLP?

→ I have read it. From that research paper, I understood how attention mechanism is affecting encoder-decoder process in neural machine translation system in NLP.

What do you understand by multi headed attention? Explain.

→ Multi headed attention is a mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension.

Basically, multiple attention heads repeats its computations multiple times in parallel. Each of these is called an Attention Head. The Attention module splits its Query, Key, and Value parameters in multiple ways and passes each split independently through a separate Head. All of these calculations are then combined together to produce a final attention output. It gives the transformer greater power to encode multiple relationships for each word.



Tell me something about your project you have done in past?

→ In the past I have done several projects from them one of them is, Number plate Detection. In that project, I have used data from Google Open ImageV6 images with OIDV toolkit & trained using YOLOv4-tiny (for real time detection.) & implemented as a local desktop app with flask API & html.

What was your Dataset size for ML Project?

→ For a ML based project of Flight-Price-Prediction model I have used a dataset with 10684 rows.

What is type of your dataset?

→ For Flight-Price-Prediction model dataset type was a combination of numerical and string data.

What was frequency of your dataset? (E.g. batch, streaming etc)?

→ In this dataset, I have used k-fold Cross Validation with 5 folds & used Xgboost algorithm.

What was source system for your dataset? (E.g. sensor, satellite Kafka, cloud, etc.).

What was kind of derived dataset that you have mentioned in project?

→ As a derived dataset, I have used credit card fraud detection data which is a imbalance dataset problem.

How you have done validation dataset?

→ In Machine learning solutions for robust solution I have used k-fold. If there is n number of folds then, it used (n-1) is for training & the last fold data is for validation.

Have you created any pipeline to validate this dataset or you were using any tool?

→ For Machine learning solutions I have created pipelines using scikit learn.

What do you understand by data lake?

→ Data lakes are next-generation data management solutions that can help businesses data scientists in meet big data challenges and drive new levels of real-time analytics. It is an easy accessible, centralized storage repository for large volumes of structured and unstructured for hosting raw, unprocessed enterprise data.

What do you understand by data warehousing?

→ Data warehouse is a system used for storing and reporting on data. It is basically, data warehousing is an electronic method of organizing, analyzing, and reporting information. In modern business, being able to integrate multiple sources of data is crucial to make better-informed decisions.

A data warehouse essentially combines information from several sources into one comprehensive database.

Can you please name some validations that you have done on top of your data?

→ K-fold, k-means Clustering, Train-test-split etc.

How you have handled streaming dataset?

→ I have handled streaming dataset by taking it as a form of batch & storing it in a database.

How many different types of environments were available in your project?

→ For Number plate detection, I have used multiple environments like- TFOD, OpenCV, Yolo etc. For, different projects, I have used different environments.

What was your delivery mechanism for particular project?

→ For delivery generally, I use Flask API & OpenCV with HTML & CSS implementation as a local/cloud desktop app.

Have you used any OPS pipeline for this current project?

→ NO.

How you were doing model retraining?

→ Using proper pipeline with Py-charm & HTML.

How you have implemented model retraining in your project?

→ I have used sequence of pipeline so that, whenever it can automatically done all the pre-training processes needed and when start training it will replace the old model with the new one & can do prediction.

How frequently you have been doing model retraining and what was the strategy for model retraining?

→ For streaming, I am doing model retraining in every 15 days with proper pipeline.

How you can define Machine Learning?

Machine learning is the concept that a computer program can learn and adapt to new data without human intervention. Machine learning is a field of artificial intelligence (AI)

that keeps a computer's built-in algorithms current regardless of changes in the worldwide economy.

Or

Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions

What do you understand Labelled training dataset?

Labelled training set is a set of training data which has a solution to the problem or task (a.k.a. label). Labelled data is a designation for pieces of data that have been tagged with one or more labels identifying certain properties or characteristics, or classifications or contained objects. Labels make that data specifically useful in certain types of machine learning known as supervised machine learning setups. Labelled dataset is the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative labels to provide context so that a machine learning model can learn from it.

What are 2 most common supervised ML tasks you have performed so far?

Two most common supervised tasks are classification and regression

What kind of Machine learning algorithm would you use to walk robot in various unknown area?

The best Machine Learning algorithm to allow a robot to walk in unknown terrain is Reinforced Learning, where the robot can learn from response of the terrain to optimize itself.

What kind of ML algo you can use to segment your user into multiple groups?

The best algorithm to segment customers into multiple groups is either supervised learning (if the groups have known labels) or unsupervised learning (if there are no group labels).

What type of learning algo realised on similarity measure to make a prediction?

Learning algorithm that relies on a similarity measure to make predictions is instance-based algorithm.

What is an online learning system?

Online learning system is a learning system in which the machine learns as data is given in small streams continuously. In computer science, online machine learning is a method of machine learning in which data becomes available in a sequential order and is used to update the best predictor for future data at each step, as opposed to batch learning techniques which generate the best predictor by learning on the entire training data set at once. Online learning is a common technique used in areas of

machine learning where it is computationally infeasible to train over the entire dataset, requiring the need of out-of-core algorithms. It is also used in situations where it is necessary for the algorithm to dynamically adapt to new patterns in the data, or when the data itself is generated as a function of time, e.g., stock price prediction. Online learning algorithms may be prone to catastrophic interference, a problem that can be addressed by incremental learning approaches.

What is out of core learning?

Out-of-core learning system is a system that can handle data that cannot fit into your computer memory. It uses online learning system to feed data in small bits. Out-of-core learning refers to a set of algorithms working with data that cannot fit into the memory of a single computer, but that can easily fit into some data storage such as a local hard disk or web repository. Your available RAM, the core memory on your single machine, may indeed range from a few gigabytes (sometimes 2 GB, more commonly 4 GB, but we assume that you have 2 GB at maximum) up to 256 GB on large server machines. Large servers are like the ones you can get on cloud computing services such as Amazon Elastic Compute Cloud (EC2), whereas your storage capabilities can easily exceed terabytes of capacity using just an external drive (most likely about 1 TB but it can reach up to 4 TB). As machine learning is based on globally reducing a cost function, many algorithms initially have been thought to work using all the available data and having access to it at each iteration of the optimization process

Can you name couple of ml challenges that you have faced?

Four main challenges in Machine Learning include overfitting the data (using a model too complicated), underfitting the data (using a simple model), lacking in data and nonrepresentative data.

Can you please give 1 example of hyperparameter tuning wrt some classification algorithm?

Machine learning algorithms have hyperparameters that allow you to tailor the behavior of the algorithm to your specific dataset.

Hyperparameters are different from parameters, which are the internal coefficients or weights for a model found by the learning algorithm. Unlike parameters, hyperparameters are specified by the practitioner when configuring the model.

Typically, it is challenging to know what values to use for the hyperparameters of a given algorithm on a given dataset, therefore it is common to use random or grid search strategies for different hyperparameter values.

The more hyperparameters of an algorithm that you need to tune, the slower the tuning process. Therefore, it is desirable to select a minimum subset of model hyperparameters to search or tune. Not all model hyperparameters are equally important. Some hyperparameters have an outsized effect on the behavior, and in turn,

the performance of a machine learning algorithm. As a machine learning practitioner, you must know which hyperparameters to focus on to get a good result quickly.

Logistic Regression

Logistic regression does not really have any critical hyperparameters to tune.

Sometimes, you can see useful differences in performance or convergence with different solvers (solver).

- solver in ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']

Regularization (penalty) can sometimes be helpful.

- penalty in ['none', 'l1', 'l2', 'elasticnet']

Note: not all solvers support all regularization terms.

The C parameter controls the penalty strength, which can also be effective.

- C in [100, 10, 1.0, 0.1, 0.01]

For the full list of hyperparameters, see:

- [sklearn.linear_model.LogisticRegression API](#).

The example below demonstrates grid searching the key hyperparameters for LogisticRegression on a synthetic binary classification dataset.

```
1 # example of grid searching key hyperparameters for logistic regression
2 from sklearn.datasets import make_blobs
3 from sklearn.model_selection import RepeatedStratifiedKFold
4 from sklearn.model_selection import GridSearchCV
5 from sklearn.linear_model import LogisticRegression
6 # define dataset
7 X, y = make_blobs(n_samples=1000, centers=2, n_features=100, cluster_std=20)
8 # define models and parameters
9 model = LogisticRegression()
10 solvers = ['newton-cg', 'lbfgs', 'liblinear']
11 penalty = ['l2']
12 c_values = [100, 10, 1.0, 0.1, 0.01]
13 # define grid search
14 grid = dict(solver=solvers,penalty=penalty,C=c_values)
15 cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
16 grid_search = GridSearchCV(estimator=model, param_grid=grid, n_jobs=-1, cv=cv, sco
17 grid_result = grid_search.fit(X, y)
18 # summarize results
19 print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_))
20 means = grid_result.cv_results_['mean_test_score']
21 stds = grid_result.cv_results_['std_test_score']
22 params = grid_result.cv_results_['params']
23 for mean, stdev, param in zip(means, stds, params):
24     print("%f (%f) with: %r" % (mean, stdev, param))
```

What is out of bag evaluation? What do you understand by hard & soft voting classifier?

In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows:

Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the k th tree.

Put each case left out in the construction of the k th tree down the k th tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case n was oob. The proportion of times that j is not equal to the true class of n averaged over all cases is the oob error estimate. This has proven to be unbiased in many tests.

In classification, a hard voting ensemble involves summing the votes for crisp class labels from other models and predicting the class with the most votes. A soft voting ensemble involves summing the predicted probabilities for class labels and predicting the class label with the largest sum probability.

Let's Suppose I have trained 5 diff model with same training dataset & all of them have achieved 95%precision. Is there any chance that you can combine all these models to get better result? If yes, How? If no, Why?

Yes, we can use K fold

Cross-validation is another method to estimate the skill of a method on unseen data. Like using a train-test split. Cross-validation systematically creates and evaluates multiple models on multiple subsets of the dataset. This, in turn, provides a population of performance measures.

We can calculate the mean of these measures to get an idea of how well the procedure performs on average. We can calculate the standard deviation of these measures to get an idea of how much the skill of the procedure is expected to vary in practice.

This is also helpful for providing a more nuanced comparison of one procedure to another when you are trying to choose which algorithm and data preparation procedures to use. Also, this information is invaluable as you can use the mean and spread to give a confidence interval on the expected performance on a machine learning procedure in practice. Both train-test splits and k-fold cross validation are examples of resampling methods.

What do you understand by Gradient decent? How will you explain Gradient decent to a kid?

Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

Types of gradient Descent:

1. Batch Gradient Descent: This is a type of gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.
2. Stochastic Gradient Descent: This is a type of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.
3. Mini Batch gradient descent: This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where $b < m$ are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

An easy explanation is as follows:

Imagine that you were in the hills, and had to find the lowest valley.

Do this repeatedly:

Start from any point on any hill

Look in all four directions (ahead, behind, left, right) to determine where you might be able to descend (rather than ascend)

Take a step in that direction

Return to step (b) above

If you have reached a point where taking a step in any direction doesn't make a difference, you're at a minimum (you've reached the valley)

Caveat: When you're at a valley, from where you can see some other cavern or valley, you're likely to be at a "local minimum"

Notes:

How big each step you take down the hills are - that represents your learning rate. Too big a step, and you bounce between hills, and too small a step, and you take forever to descend into the valley

Gradient descent algorithms do much the same things, but in an arbitrary number of dimensions. The problem I've described above is in three dimensions of space, with steps taken over time. In mathematical functions, you can define arbitrarily large spaces where gradient descent can take place.

Can you please explain diff between regression & classification?

Classification and Regression are two major prediction problems which are usually dealt with Data mining and machine learning.

Classification is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some

parameters given in input and then the labels are predicted for the data. The derived mapping function could be demonstrated in the form of “IF-THEN” rules. The classification process deal with the problems where the data can be divided into binary or multiple discrete labels.

Let’s take an example, suppose we want to predict the possibility of the wining of match by Team A on the basis of some parameters recorded earlier. Then there would be two labels Yes and No.

Regression is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement depending on the historical data. Because a regression predictive model predicts a quantity, therefore, the skill of the model must be reported as an error in those predictions. Let’s take a similar example in regression also, where we are finding the possibility of rain in some particular regions with the help of some parameters recorded earlier. Then there is a probability associated with the rain.

Explain a clustering algorithm of your choice.

K-Means Clustering

K-Means is by far the most popular clustering algorithm given that it is very easy to understand and apply to a wide range of data science and machine learning problems. Here’s how you can apply the K-Means algorithm to your clustering problem.

The first step is to select a number of clusters randomly, each of which is represented by a variable ‘k’. Next, each cluster is assigned a centroid, i.e., the centre of that particular cluster. It is important to define the centroids as far off from each other as possible to reduce variation. After all the centroids are defined, each data point is assigned to the cluster whose centroid is at the closest distance.

Once all data points are assigned to respective clusters, the centroid is again assigned for each cluster. Once again, all data points are rearranged in specific clusters based on their distance from the newly defined centroids. This process is repeated until the centroids stop moving from their positions.

K-Means algorithm works wonders in grouping new data. Some of the practical applications of this algorithm are in sensor measurements, audio detection, and image segmentation.

How you can explain ML, DL, NLP, Computer vision & reinforcement learning with example in your own term

(AI) is the domain of producing intelligent machines. ML refers to systems that can assimilate from experience (training data) and Deep Learning (DL) states to systems that learn from experience on large data sets. ML can be considered as a subset of AI. Deep Learning (DL) is ML but useful to large data sets. The figure below roughly encapsulates the relation between AI, ML, and DL: In summary, DL is a subset of ML & both were the subsets of AI.

Additional Information: ASR (Automatic Speech Recognition) & NLP (Natural Language Processing) fall under AI and overlay with ML & DL as ML is often utilized for NLP and ASR tasks.

NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors -- such as ears to hear and eyes to see -- computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.

Computer vision:

Computer Vision, often abbreviated as CV, is defined as a field of study that seeks to develop techniques to help computers “see” and understand the content of digital images such as photographs and videos. The problem of computer vision appears simple because it is trivially solved by people, even very young children. Nevertheless, it largely remains an unsolved problem based both on the limited understanding of biological vision and because of the complexity of vision perception in a dynamic and nearly infinitely varying physical world.

Reinforcement

Learning:

The model learns through a trial-and-error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

Machine Learning involves algorithms that learn from patterns of data and then apply it to decision making. Deep Learning, on the other hand, is able to learn through processing data on its own and is quite similar to the human brain where it identifies something, analyse it, and makes a decision.

The key differences are as follow:

The manner in which data is presented to the system.

Machine learning algorithms always require structured data and deep learning networks rely on layers of artificial neural networks.

How you can explain semi-supervised ML in your own way with example?

Semi-supervised machine learning is a combination of supervised and unsupervised machine learning methods.

With more common supervised machine learning methods, you train a machine learning algorithm on a “labeled” dataset in which each record includes the outcome information. This allows the algorithm to deduce patterns and identify relationships

between your target variable and the rest of the dataset based on information it already has. In contrast, unsupervised machine learning algorithms learn from a dataset without the outcome variable. In semi-supervised learning, an algorithm learns from a dataset that includes both labeled and unlabeled data, usually mostly unlabeled.

Examples of semi supervised ML

Speech Analysis: Since labeling of audio files is a very intensive task, Semi-Supervised learning is a very natural approach to solve this problem.

Internet Content Classification: Labeling each webpage is an impractical and unfeasible process and thus uses Semi-Supervised learning algorithms. Even the Google search algorithm uses a variant of Semi-Supervised learning to rank the relevance of a webpage for a given query.

Protein Sequence Classification: Since DNA strands are typically very large in size, the rise of Semi-Supervised learning has been imminent in this field.

What is difference between abstraction & generalization in your own word.

Abstraction is the process of removing details of objects. We step back from concrete objects to consider a number of objects with identical properties. So a concrete object can be looked at as a “superset” of a more abstract object.

A generalization, then, is the formulation of general concepts from specific instances by abstracting common properties. A concrete object can be looked at as a “subset” of a more generalized object.

In other words:

1. For any two concepts A and B, A is an abstraction of B if and only if:
 - Every instance of concept B is also an instance of concept A
2. For any two concepts A and B, A is a generalization of B if and only if:
 - Every instance of concept B is also an instance of concept A
 - There are instances of concept A which are not instances of concept B