1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Answer: Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. Data cleaning.

Data integration.

Data reduction.

Data transformation. Data Preprocessing includes the steps we need to follow to transform or encode data so that it may be easily parsed by the machine.

The main agenda for a model to be accurate and precise in predictions is that the algorithm should be able to easily interpret the data's features. The majority of the real-world datasets for machine learning are highly susceptible to be missing, inconsistent, and noisy due to their heterogeneous origin.

Applying data mining algorithms on this noisy data would not give quality results as they would fail to identify patterns effectively. Data Processing is, therefore, important to improve the overall data quality.

Duplicate or missing values may give an incorrect view of the overall statistics of data. Outliers and inconsistent data points often tend to disturb the model’s overall learning, leading to false predictions.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

Answer: Quantitative data is numbers-based, countable, or measurable. Qualitative data is interpretation-based, descriptive, and relating to language. Quantitative data tells us how many, how much, or how often in calculations. Qualitative data can help us to understand why, how, or what happened behind certain behaviors When it comes to conducting data research, you’ll need different collection, hypotheses and analysis methods, so it’s important to understand the key differences between quantitative and qualitative data:

Quantitative data is numbers-based, countable, or measurable. Qualitative data is interpretation-based, descriptive, and relating to language.

Quantitative data tells us how many, how much, or how often in calculations. Qualitative data can help us to understand why, how, or what happened behind certain behaviors.

Quantitative data is fixed and universal. Qualitative data is subjective and unique.

Quantitative research methods are measuring and counting. Qualitative research methods are interviewing and observing.

Quantitative data is analyzed using statistical analysis. Qualitative data is analyzed by grouping the data into categories and themes. Each type of data set has its own pros and cons.

Advantages of quantitative data It's relatively quick and easy to collect and it's easier to draw conclusions from. When you collect quantitative data, the type of results will tell you which statistical tests are appropriate to use. As a result, interpreting your data and presenting those findings is straightforward and less open to error and subjectivity. Another advantage is that you can replicate it. Replicating a study is possible because your data collection is measurable and tangible for further applications.

Disadvantages of quantitative data Quantitative data doesn't always tell you the full story (no matter what the perspective). With choppy information, it can be inconclusive. Quantitative research can be limited, which can lead to overlooking broader themes and relationships. By focusing solely on numbers, there is a risk of missing larger focus information that can be beneficial.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

Answer: Data can come in many forms, but machine learning models rely on four primary data types. These include numerical data, categorical data, time series data, and text data.

Preparing Your Dataset for Machine Learning: 10 Basic Techniques That Make Your Data Better Articulate the problem early. Establish data collection mechanisms. ... Check your data quality. Format data to make it consistent. Reduce data. Complete data cleaning. Create new features out of existing ones.

How to Prepare Data For Machine Learning by Jason Brownlee on December 25, 2013 in Data Preparation Tweet Tweet Share Last Updated on August 16, 2020

Machine learning algorithms learn from data. It is critical that you feed them the right data for the problem you want to solve.

Even if you have good data, you need to make sure that it is in a useful scale, format and even that meaningful features are included.

In this post you will learn how to prepare data for a machine learning algorithm. This is a big topic and you will cover the essentials.

Kick-start your project with my new book Data Preparation for Machine Learning, including step-by-step tutorials and the Python source code files for all examples.

Let's get started.

lots of data Lots of Data Photo attributed to cibomahto, some rights reserved

Data Preparation Process The more disciplined you are in your handling of data, the more consistent and better results you are like likely to achieve. The process for getting data ready for a machine learning algorithm can be summarized in three steps:

Step 1: Select Data Step 2: Preprocess Data Step 3: Transform Data You can follow this process in a linear manner, but it is very likely to be iterative with many loops.

Want to Get Started With Data Preparation? Take my free 7-day email crash course now (with sample code).

Click to sign-up and also get a free PDF Ebook version of the course.

Download Your FREE Mini-Course

Step 1: Select Data This step is concerned with selecting the subset of all available data that you will be working with. There is always a strong desire for including all data that is available, that the maxim “more is better” will hold. This may or may not be true.

You need to consider what data you actually need to address the question or problem you are working on. Make some assumptions about the data you require and be careful to record those assumptions so that you can test them later if needed.

Below are some questions to help you think through this process:

What is the extent of the data you have available? For example through time, database tables, connected systems. Ensure you have a clear picture of everything that you can use. What data is not available that you wish you had available? For example data that is not recorded or cannot be recorded. You may be able to derive or simulate this data.

How to Prepare Data For Machine Learning by Jason Brownlee on December 25, 2013 in Data Preparation Tweet Tweet Share Last Updated on August 16, 2020

Machine learning algorithms learn from data. It is critical that you feed them the right data for the problem you want to solve.

Even if you have good data, you need to make sure that it is in a useful scale, format and even that meaningful features are included.

In this post you will learn how to prepare data for a machine learning algorithm. This is a big topic and you will cover the essentials.

Kick-start your project with my new book Data Preparation for Machine Learning, including step-by-step tutorials and the Python source code files for all examples.

Let’s get started.

lots of data Lots of Data Photo attributed to cibomahto, some rights reserved

Data Preparation Process The more disciplined you are in your handling of data, the more consistent and better results you are like likely to achieve. The process for getting data ready for a machine learning algorithm can be summarized in three steps:

Step 1: Select Data Step 2: Preprocess Data Step 3: Transform Data You can follow this process in a linear manner, but it is very likely to be iterative with many loops.

Want to Get Started With Data Preparation? Take my free 7-day email crash course now (with sample code).

Click to sign-up and also get a free PDF Ebook version of the course.

Download Your FREE Mini-Course

Step 1: Select Data This step is concerned with selecting the subset of all available data that you will be working with. There is always a strong desire for including all data that is available, that the maxim “more is better” will hold. This may or may not be true.

You need to consider what data you actually need to address the question or problem you are working on. Make some assumptions about the data you require and be careful to record those assumptions so that you can test them later if needed.

Below are some questions to help you think through this process:

What is the extent of the data you have available? For example through time, database tables, connected systems. Ensure you have a clear picture of everything that you can use. What data is not available that you wish you had available? For example data that is not recorded or cannot be recorded. You may be able to derive or simulate this data. What data don’t you need to address the problem? Excluding data is almost always easier than including data. Note down which data you excluded and why. It is only in small problems, like competition or toy datasets where the data has already been selected for you.

Step 2: Preprocess Data After you have selected the data, you need to consider how you are going to use the data. This preprocessing step is about getting the selected data into a form that you can work.

Three common data preprocessing steps are formatting, cleaning and sampling:

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file. Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely. Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

How to Prepare Data For Machine Learning by Jason Brownlee on December 25, 2013 in Data Preparation Tweet Tweet Share Last Updated on August 16, 2020

Machine learning algorithms learn from data. It is critical that you feed them the right data for the problem you want to solve.

Even if you have good data, you need to make sure that it is in a useful scale, format and even that meaningful features are included.

In this post you will learn how to prepare data for a machine learning algorithm. This is a big topic and you will cover the essentials.

Kick-start your project with my new book Data Preparation for Machine Learning, including step-by-step tutorials and the Python source code files for all examples.

Let's get started.

lots of data Lots of Data Photo attributed to cibomahto, some rights reserved

Data Preparation Process The more disciplined you are in your handling of data, the more consistent and better results you are like likely to achieve. The process for getting data ready for a machine learning algorithm can be summarized in three steps:

Step 1: Select Data Step 2: Preprocess Data Step 3: Transform Data You can follow this process in a linear manner, but it is very likely to be iterative with many loops.

Want to Get Started With Data Preparation? Take my free 7-day email crash course now (with sample code).

Click to sign-up and also get a free PDF Ebook version of the course.

Download Your FREE Mini-Course

Step 1: Select Data This step is concerned with selecting the subset of all available data that you will be working with. There is always a strong desire for including all data that is available, that the maxim "more is better" will hold. This may or may not be true.

You need to consider what data you actually need to address the question or problem you are working on. Make some assumptions about the data you require and be careful to record those assumptions so that you can test them later if needed.

Below are some questions to help you think through this process:

What is the extent of the data you have available? For example through time, database tables, connected systems. Ensure you have a clear picture of everything that you can use. What data is not available that you wish you had available? For example data that is not recorded or cannot be recorded. You may be able to derive or simulate this data. What data don't you need to address the problem? Excluding data is almost always easier than including data. Note down which data you excluded and why. It is only in small problems, like competition or toy datasets where the data has already been selected for you.

Step 2: Preprocess Data After you have selected the data, you need to consider how you are going to use the data. This preprocessing step is about getting the selected data into a form that you can work.

Three common data preprocessing steps are formatting, cleaning and sampling:

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file. Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely. Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset. It is very likely that the machine learning tools you use on the data will influence the preprocessing you will be required to perform. You will likely revisit this step.

So much data So much data Photo attributed to Marc_Smith, some rights reserved

Step 3: Transform Data The final step is to transform the process data. The specific algorithm you are working with and the knowledge of the problem domain will influence this step and you will very likely have to revisit different transformations of your preprocessed data as you work on your problem.

Three common data transformations are scaling, attribute decompositions and attribute aggregations. This step is also referred to as feature engineering.

Scaling: The preprocessed data may contain attributes with a mixtures of scales for various quantities such as dollars, kilograms and sales volume. Many machine learning methods like data attributes to have the same scale such as between 0 and 1 for the smallest and largest value for a given feature. Consider any feature scaling you may need to perform. Decomposition: There may be features that represent a complex concept that may be more useful to a machine learning method when split into the constituent parts. An example is a date that may have day and time components that in turn could be split out further. Perhaps only the hour of day is relevant to the problem being solved. consider what feature decompositions you can perform. Aggregation: There may be features that can be aggregated into a single feature that would be more meaningful to the problem you are trying to solve. For example, there may be a data instances for each time a customer logged into a system that could be aggregated into a count for the number of logins allowing the additional instances to be discarded. Consider what type of feature aggregations could perform.

4. What are the various causes of machine learning data issues? What are the ramifications?

Answer: The number one problem facing Machine Learning is the lack of good data. While enhancing algorithms often consumes most of the time of developers in AI, data quality is essential for the algorithms to function as intended.The process is transforming, and hence there are high chances of error which makes the learning complex. It includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, and a lot more. List Of Common Practical Issues In Machine Learning

1) Lack Of Quality Data One of the main issues in Machine Learning is the absence of good data. While upgrading, algorithms tend to make developers exhaust most of their time on artificial intelligence. Data quality is fundamental for the algorithms to work as proposed. Incomplete data, unclean data, and noisy data are the quintessential foes of ideal ML. Different reasons for low data quality are-

Data can be noisy which will result in inaccurate predictions. This often leads to less accuracy in classification and low-quality results. It is noted as one of the most common errors faced in terms of data. Incorrect or incomplete information can also lead to faulty programming through Machine Learning. Having less information will lead the program to analyze based on the minimal data present. Hence, decreasing the accuracy of the results. For better future actions, the generalizing of input and output of past data is crucial. But a common issue that occurs is, that the output data can become difficult to generalize. 2) TALENT DEFICIT Albeit numerous individuals are pulled into the ML business, however, there are still not many experts who can take complete control of this innovation. It is quite rare to find a trained professional who is capable of comprehending the problems in Machine Learning and being able to reach out to a reliable software solution for the same.

3) IMPLEMENTATION Organizations regularly have examination engines working with them when they decide to move up to ML. The usage of fresher ML strategies with existing procedures is a complicated errand. Keeping up legitimate documentation and interpretation needs to go a long way to facilitating maximum usage. There are issues in Machine Learning when it comes to implementation-

Slow deployment – Although the models of Machine Learning are time efficient the creation process of the same is quite the opposite. As it is still a young innovation the implementation time is slow. Data Security – Saving confidential data on ML servers is a risk as the model will not be able to differentiate between sensitive and insensitive data. Lack of data is another key issue faced during the implementation of the model. Without adequate data, it is not possible to give out valuable intel.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Answer: Categorical Variable/Data (or Nominal variable): Such variables take on a fixed and limited number of possible values. For example – grades, gender, blood group type, etc. Also, in the case of categorical variables, the logical order is not the same as categorical data e.g. “one”, “two”, “three”. But the sorting of these variables uses logical order. For example, gender is a categorical variable and has categories – male and female and there is no intrinsic ordering to the categories. A purely categorical variable is one that simply allows you to assign categories, but you cannot clearly order the variables. Terms related to Variability Metrics : Exploring categorical variables is generally simpler than working with numeric variables because we have fewer options, or at least life is simpler if we only require basic summaries. We’ll work with the year and type variables in storms to illustrate the key ideas.

Which kind of categorical variable is type? There are four storm categories in type. We can use the unique function to print these for us:

unique(storms$type)

[1] "Tropical Depression" "Tropical Storm" "Hurricane"

[4] "Extratropical"

The first question we should ask is, is type an ordinal or nominal variable? It's hard to know how to classify type without knowing something about tropical storms. Some googling indicates that type can reasonably be considered an ordinal variable: a tropical depression is the least severe storm and a hurricane is the most severe class; in between are extra tropical and tropical storm categories.

What about the year variable? Years are obviously ordered from early to late and we might be interested in how some aspect of our data changes over time. In this case we might consider treating year either as a numeric variable, or perhaps as an ordinal categorical variable. Alternatively, if the question is simply, 'do the data vary from one year to the next' without any concern for trends, it's perfectly reasonable to treat year as a nominal categorical variable.

This illustrates an important idea: the classification of a variable will often depend on the objectives of an analysis. The classification of a variable matters because it influences how we choose to summarise it, how we interpret its relationship with other variables, and whether a specific statistical model is appropriate for our data or not. Fortunately, our choice of classification is less important when we are just trying to summarise the variable numerically or graphically. For now, let's assume that it's fine to treat year as a categorical variable.

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Answer: Missing data is an everyday problem that a data professional need to deal with. Though there are many articles, blogs, videos already available, I found it is difficult to find a piece of concise consolidated information in a single place. That's why I am putting my effort here, hoping it will be useful to any data practitioner or enthusiast.Missing data are defined as not available values, and that would be meaningful if observed. Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc. Most datasets in the real world contain missing data. Before you can use data with missing data fields, you need to transform those fields to be used for analysis and modelling. Like many other aspects of data science, this too may actually be more art than science. Understanding the data and the domain from which it comes is very important.

Having missing values in your data is not necessarily a setback. Still, it is an opportunity to perform the right feature engineering to guide the model to interpret the missing information the right way. There are machine learning algorithms and packages that can automatically detect and deal with missing data. However, it's still recommended to transform the missing data manually through analysis and coding strategy. First, we need to understand what are the types of missing data. Missingness is broadly categorized into 3 categories: Missing data (or missing values) is defined

as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data [1]. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research [2,3].

However, until recently, most researchers have drawn conclusions based on the assumption of a complete data set. The general topic of missing data has attracted little attention in the field of anesthesiology.

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions.

7. Describe the various methods for dealing with missing data values in depth.

```
Answer: Missing data (or missing values) is defined as the data value that is
not stored for a variable in the observation of interest. The problem of
missing data is relatively common in almost all research and can have a
significant effect on the conclusions that can be drawn from the data [1].
Accordingly, some studies have focused on handling the missing data, problems
caused by missing data, and the methods to avoid or minimize such in medical
research [2,3].

However, until recently, most researchers have drawn conclusions based on the
assumption of a complete data set. The general topic of missing data has
attracted little attention in the field of anesthesiology.

Missing data present various problems. First, the absence of data reduces
statistical power, which refers to the probability that the test will reject
the null hypothesis when it is false. Second, the lost data can cause bias in
the estimation of parameters. Third, it can reduce the representativeness of
the samples. Fourth, it may complicate the analysis of the study. Each of these
distortions may threaten the validity of the trials and can lead to invalid
conclusions.

Go to:
Types of Missing Data
Rubin first described and divided the types of missing data according to the
assumptions based on the reasons for the missing data [4]. In general, there
are three types of missing data according to the mechanisms of missingness.

Missing completely at random
Missing completely at random (MCAR) is defined as when the probability that the
data are missing is not related to either the specific value which is supposed
to be obtained or the set of observed responses. MCAR is an ideal but
unreasonable assumption for many studies performed in the field of
anesthesiology. However, if data are missing by design, because of an equipment
failure or because the samples are lost in transit or technically
unsatisfactory, such data are regarded as being MCAR.
```

The statistical advantage of data that are MCAR is that the analysis remains unbiased. Power may be lost in the design, but the estimated parameters are not biased by the absence of the data.

Missing at random
Missing at random (MAR) is a more realistic assumption for the studies performed in the anesthetic field. Data are regarded to be MAR when the probability that the responses are missing depends on the set of observed responses, but is not related to the specific missing values which is expected to be obtained.

As we tend to consider randomness as not producing bias, we may think that MAR does not present a problem. However, MAR does not mean that the missing data can be ignored. If a dropout variable is MAR, we may expect that the probability of a dropout of the variable in each case is conditionally independent of the variable, which is obtained currently and expected to be obtained in the future, given the history of the obtained variable prior to that case.

Missing not at random
If the characters of the data do not meet those of MCAR or MAR, then they fall into the category of missing not at random (MNAR).

The cases of MNAR data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data. The model may then be incorporated into a more complex one for estimating the missing values.

Go to:
Techniques for Handling the Missing Data
The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully [5,6]. The following are suggested to minimize the amount of missing data in the clinical research [7].

First, the study design should limit the collection of data to those who are participating in the study. This can be achieved by minimizing the number of follow-up visits, collecting only the essential information at each visit, and developing the userfriendly case-report forms.

Second, before the beginning of the clinical research, a detailed documentation of the study should be developed in the form of the manual of operations, which includes the methods to screen the participants, protocol to train the investigators and participants, methods to communicate between the investigators or between the investigators and participants, implementation of the treatment, and procedure to collect, enter, and edit data.

Third, before the start of the participant enrollment, a training should be conducted to instruct all personnel related to the study on all aspects of the study, such as the participant enrollment, collection and entry of data, and implementation of the treatment or intervention [8].

Fourth, if a small pilot study is performed before the start of the main trial, it may help to identify the unexpected problems which are likely to occur during the study, thus reducing the amount of missing data.

Fifth, the study management team should set a priori targets for the unacceptable level of missing data. With these targets in mind, the data collection at each site should be monitored and reported in as close to real-time as possible during the course of the study.

Sixth, study investigators should identify and aggressively, though not coercively, engage the participants who are at the greatest risk of being lost during follow-up.

Finally, if a patient decides to withdraw from the follow-up, the reasons for the withdrawal should be recorded for the subsequent analysis in the interpretation of the results.

It is not uncommon to have a considerable amount of missing data in a study. One technique of handling the missing data is to use the data analysis methods which are robust to the problems caused by the missing data. An analysis method is considered robust to the missing data when there is confidence that mild to moderate violations of the assumptions will produce little to no bias or distortion in the conclusions drawn on the population. However, it is not always possible to use such techniques. Therefore, a number of alternative ways of handling the missing data has been developed.

Listwise or case deletion
By far the most common approach to the missing data is to simply omit those cases with the missing data and analyze the remaining data. This approach is known as the complete case (or available case) analysis or listwise deletion.

Listwise deletion is the most frequently used method in handling missing data, and thus has become the default option for analysis in most statistical software packages. Some researchers insist that it may introduce bias in the estimation of the parameters. However, if the assumption of MCAR is satisfied, a listwise deletion is known to produce unbiased estimates and conservative results. When the data do not fulfill the assumption of MCAR, listwise deletion may cause bias in the estimates of the parameters [9].

If there is a large enough sample, where power is not an issue, and the assumption of MCAR is satisfied, the listwise deletion may be a reasonable strategy. However, when there is not a large sample, or the assumption of MCAR is not satisfied, the listwise deletion is not the optimal strategy.

Pairwise deletion
Pairwise deletion eliminates information only when the particular data-point needed to test a particular assumption is missing. If there is missing data elsewhere in the data set, the existing values are used in the statistical testing. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion, which may delete the case with any missing data. This approach presents the following problems: 1) the parameters of the model will stand on different sets of data with different statistics, such as the sample size and standard errors; and 2) it can produce an intercorrelation matrix that is not positive definite, which is likely to prevent further analysis [10].

Pairwise deletion is known to be less biased for the MCAR or MAR data, and the appropriate mechanisms are included as covariates. However, if there are many missing observations, the analysis will be deficient.

Mean substitution

In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This allows the researchers to utilize the collected data in an incomplete dataset. The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution. However, with missing values that are not strictly random, especially in the presence of a great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias. Furthermore, this approach adds no new information but only increases the sample size and leads to an underestimate of the errors [11]. Thus, mean substitution is not generally accepted.

Regression imputation
Imputation is the process of replacing the missing data with estimated values. Instead of deleting any case that has any missing value, this approach preserves all cases by replacing the missing data with a probable value estimated by other available information. After all missing values have been replaced by this approach, the data set is analyzed using the standard techniques for a complete data.

In regression imputation, the existing variables are used to make a prediction, and then the predicted value is substituted as if an actual obtained value. This approach has a number of advantages, because the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced.

Last observation carried forward
In the field of anesthesiology research, many studies are performed with the longitudinal or time-series approach, in which the subjects are repeatedly measured over a series of time-points. One of the most widely used imputation methods in such a case is the last observation carried forward (LOCF). This method replaces every missing value with the last observed value from the same subject. Whenever a value is missing, it is replaced with the last observed value [12].

This method is advantageous as it is easy to understand and communicate between the statisticians and clinicians or between a sponsor and the researcher.

Although simple, this method strongly assumes that the value of the outcome remains unchanged by the missing data, which seems unlikely in many settings (especially in the anesthetic trials). It produces a biased estimate of the treatment effect and underestimates the variability of the estimated result. Accordingly, the National Academy of Sciences has recommended against the uncritical use of the simple imputation, including LOCF and the baseline observation carried forward, stating that:

Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified [13].

Maximum likelihood

There are a number of strategies using the maximum likelihood method to handle the missing data. In these, the assumption that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand. After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.

When there are missing but relatively complete data, the statistics explaining the relationships among the variables may be computed using the maximum likelihood method. That is, the missing data may be estimated by using the conditional distribution of the other variables.

Expectation-Maximization
Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods [14]. This approach begins with the expectation step, during which the parameters (e.g., variances, covariances, and means) are estimated, perhaps using the listwise deletion. Those estimates are then used to create a regression equation to predict the missing data. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to "fill in" the missing data. The expectation and maximization steps are repeated until the system stabilizes, when the covariance matrix for the subsequent iteration is virtually the same as that for the preceding iteration.

An important characteristic of the expectation-maximization imputation is that when the new data set with no missing values is generated, a random disturbance term for each imputed value is incorporated in order to reflect the uncertainty associated with the imputation. However, the expectation-maximization imputation has some disadvantages. This approach can take a long time to converge, especially when there is a large fraction of missing data, and it is too complex to be acceptable by some exceptional statisticians. This approach can lead to the biased parameter estimates and can underestimate the standard error.

For the expectation-maximization imputation method, a predicted value based on the variables that are available for each case is substituted for the missing data. Because a single imputation omits the possible differences among the multiple imputations, a single imputation will tend to underestimate the standard errors and thus overestimate the level of precision. Thus, a single imputation gives the researcher more apparent power than the data in reality.

Multiple imputation
Multiple imputation is another useful strategy for handling the missing data. In a multiple imputation, instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values.

This approach begin with a prediction of the missing data using the existing data from other variables [15]. The missing values are then replaced with the predicted values, and a full data set called the imputed data set is created. This process iterates the repeatability and makes multiple imputed data sets (hence the term "multiple imputation"). Each multiple imputed data set produced is then analyzed using the standard statistical analysis procedures for complete data, and gives multiple analysis results. Subsequently, by combining these analysis results, a single overall analysis result is produced.

The benefit of the multiple imputation is that in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in a valid statistical inference. Restoring the natural variability of the missing data can be achieved by replacing the missing data with the imputed values which are predicted using the variables correlated with the missing data. Incorporating uncertainty is made by producing different versions of the missing data and observing the variability between the imputed data sets.

Multiple imputation has been shown to produce valid statistical inference that reflects the uncertainty associated with the estimation of the missing data. Furthermore, multiple imputation turns out to be robust to the violation of the normality assumptions and produces appropriate results even in the presence of a small sample size or a high number of missing data.

With the development of novel statistical software, although the statistical principles of multiple imputation may be difficult to understand, the approach may be utilized easily.

Sensitivity analysis
Sensitivity analysis is defined as the study which defines how the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs.

When analyzing the missing data, additional assumptions on the reasons for the missing data are made, and these assumptions are often applicable to the primary analysis. However, the assumptions cannot be definitively validated for the correctness. Therefore, the National Research Council has proposed that the sensitivity analysis be conducted to evaluate the robustness of the results to the deviations from the MAR assumption [13].

Go to:
Recommendations
Missing data reduces the power of a trial. Some amount of missing data is expected, and the target sample size is increased to allow for it. However, such cannot eliminate the potential bias. More attention should be paid to the missing data in the design and performance of the studies and in the analysis of the resulting data.

The best solution to the missing data is to maximize the data collection when the study protocol is designed and the data collected. Application of the sophisticated statistical analysis techniques should only be performed after the maximal efforts have been employed to reduce missing data in the design and prevention techniques.

A statistically valid analysis which has appropriate mechanisms and assumptions for the missing data should be conducted. Single imputation and LOCF are not optimal approaches for the final analysis, as they can cause bias and lead to invalid conclusions. All variables which present the potential mechanisms to explain the missing data must be included, even when these variables are not included in the analysis [16]. Researchers should seek to understand the reasons for the missing data. Distinguishing what should and should not be imputed is usually not possible using a single code for every type of the missing value [17]. It is difficult to know whether the multiple imputation or full maximum likelihood estimation is best, but both are superior to the traditional approaches. Both techniques are best used with large samples. In general, multiple imputation is a good approach when analyzing data sets with missing data.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

Answer: Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Steps Involved in Data Preprocessing: 1. Data Cleaning: The data can have many irrelevant and missing parts. Data preprocesing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results. Data quality assessment. Data cleaning. Data transformation. Data reduction. Fill the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data: Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways : Binning Method: This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

Regression: Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

2. Data Transformation: This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

Normalization: It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

Attribute Selection: In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

Discretization: This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

Concept Hierarchy Generation: Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

9.

i. What is the IQR? What criteria are used to assess it?

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

Answer: The IQR of a set of values is calculated as the difference between the upper and lower quartiles, Q3 and Q1. Each quartile is a median calculated as follows. The second quartile Q2 is the same as the ordinary median. IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts. Q1 represents the 25th percentile of the data. We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. The interquartile range (IQR) measures the spread of the middle half of your data. It is the range for the middle 50% of your sample. Use the IQR to assess the variability where most of your values lie. Larger values indicate that the central portion of your data spread out further. Conversely, smaller values show that the middle values cluster more tightly.

In this post, learn what the interquartile range means and the many ways to use it! I’ll show you how to find the interquartile range, use it to measure variability, graph it in boxplots to assess distribution properties, use it to identify outliers, and test whether your data are normally distributed.

The interquartile range is one of several measures of variability. To learn about the others and how the IQR compares, read my post, Measures of Variability.

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

Answer: The lowest point of the lower whisker is called the lower limit. Lower limit value equals Q1 – 1.5 * (Q3-Q1). The highest point of the upper whisker is the called the upper limit. The upper limit value equals Q3 + 1.5 * (Q3-Q1). A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.The left edge of the box represents the lower quartile; it shows the value at which the first 25 % of the data falls up to. The right edge of the box shows the upper quartile; it shows that 25 % of the data lies to the right of the upper quartile value.In descriptive statistics, a box plot or boxplot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.

10. Make brief notes on any two of the following:

1. Data collected at regular intervals
2. The gap between the quartiles
3. Use a cross-tab

Answer: 1.Interval data, also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another. Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal.Interval data is a type of data which is measured along a scale, in which each point is placed at an equal distance (interval) from one another. Interval data is one of the two types of discrete data. An example of interval data is the data collected on a thermometer—its gradation or markings are equidistant. The simplest levels of measurement are nominal and ordinal data. These are both types of categorical data that take useful but imprecise measures of a variable. They are easier to work with but offer less accurate insights. Building on these are interval data and ratio data, which are both types of numerical data. While these are more complex, they can offer much richer insights.

Nominal data is the simplest (and most imprecise) data type. It uses labels to identify values, without quantifying how those values relate to one another e.g. employment status, blood type, eye color, or nationality.

Ordinal data also labels data but introduces the concept of ranking. A dataset of different qualification types is an example of ordinal data because it contains an explicit, increasing hierarchy, e.g. High School Diploma, Bachelor's, Master's, Ph.D., etc.

Interval data categorizes and ranks data, and introduces precise and continuous intervals, e.g. temperature measurements in Fahrenheit and Celsius, or the pH scale. Interval data always lack what's known as a 'true zero.' In short, this means that interval data can contain negative values and that a measurement of 'zero' can represent a quantifiable measure of something.

Ratio data categorizes and ranks data, and uses continuous intervals (like interval data). However, it also has a true zero, which interval data does not. Essentially, this means that when a variable is equal to zero, there is none of this variable. An example of ratio data would be temperature measured on the Kelvin scale, for which there is no measurement below absolute zero (which represents a total absence of heat).

2. The Gap between the quantile Quartiles are three values that split sorted data into four parts, each with an equal number of observations. Quartiles are a type of quantile.Quartiles are a set of descriptive statistics. They summarize the central tendency and variability of a dataset or distribution.

Quartiles are a type of percentile. A percentile is a value with a certain percentage of the data falling below it. In general terms, k% of the data falls below the kth percentile.

The first quartile (Q1, or the lowest quartile) is the 25th percentile, meaning that 25% of the data falls below the first quartile. The second quartile (Q2, or the median) is the 50th percentile, meaning that 50% of the data falls below the second quartile. The third quartile (Q3, or the upper quartile) is the 75th percentile, meaning that 75% of the data falls below the third quartile. By splitting the data at the 25th, 50th, and 75th percentiles, the quartiles divide the data into four equal parts.

In a sample or dataset, the quartiles divide the data into four groups with equal numbers of observations. In a probability distribution, the quartiles divide the distribution's range into four intervals with equal probability.

First quartile: Also known as Q1, or the lower quartile. This is the number halfway between the lowest number and the middle number. Second quartile: Also known as Q2, or the median. This is the middle number halfway between the lowest number and the highest number. Third quartile: Also known as Q3, or the upper quartile. This is the number halfway between the middle number and the highest number.

In [ ]: