Machine Learning #04

▼ 1. What are the key tasks involved in getting ready to work with machine learning modeling?

Before working with machine learning modeling, there are several key tasks that should be completed to ensure success:

1. ***Data collection:*** Collecting and preparing data is a crucial task in machine learning modeling. ***It involves identifying relevant data sources, gathering and cleaning data, and ensuring the data is in a suitable format for modeling.***
2. ***Data exploration and visualization:*** ***Exploring and visualizing data helps to identify patterns, relationships, and potential issues in the data.*** This is an important step in understanding the data and selecting appropriate modeling techniques.
3. ***Feature selection and engineering:*** Selecting and engineering features ***involves identifying the relevant variables or features that will be used in the modeling process.*** This includes identifying important predictors and transforming variables to improve the performance of the model.
4. ***Model selection:*** ***Selecting an appropriate modeling technique is essential to achieving good results.*** This involves selecting from a range of models, such as decision trees, neural networks, or regression models.
5. ***Model training and evaluation:*** Training and evaluating the model involves dividing the data into training and testing sets, training the model on the training set, and evaluating the performance of the model on the testing set.
6. ***Model optimization:*** Optimization involves ***fine-tuning the model to achieve better performance.*** This includes adjusting model parameters, regularization, and ensemble techniques.
7. ***Deployment and monitoring:*** Once the model has been trained and optimized, ***it can be deployed in a production environment.*** Ongoing monitoring and maintenance is necessary to ensure the model continues to perform well over time.

▼ 2. What are the different forms of data used in machine learning? Give a specific example for each of them.

There are generally three types of data used in machine learning:

1. ***Numerical Data:*** ***This type of data consists of numbers and can be either continuous or discrete.*** Examples of numerical data include temperature, height, weight, and income.
2. ***Categorical Data:*** ***This type of data consists of categories or labels that are not numerical in nature.*** Examples of categorical data include gender, color, occupation, and zip code.
3. ***Text Data:*** ***This type of data consists of text in the form of words, sentences, or paragraphs.*** Examples of text data include emails, social media posts, and product reviews.

Here are some specific examples for each type of data:

1. ***Numerical Data:*** In a healthcare setting, numerical data could include a patient's blood pressure, cholesterol level, and body mass index (BMI). These data points could be used to predict the

likelihood of the patient developing a specific disease.

2. ***Categorical Data:*** In an e-commerce setting, categorical data could include a customer's age group, gender, and location. These data points could be used to predict the likelihood of the customer purchasing a specific product or service.
3. ***Text Data:*** In a social media setting, text data could include posts, comments, and messages. These data points could be used to predict the sentiment of the user towards a specific topic or product.

▼ 3. Distinguish:
1. Numeric vs. categorical attributes
2. Feature selection vs. dimensionality reduction

1. ***Numeric vs. categorical attributes:***

   Numeric attributes are those that have a numeric value or can be measured on a numerical scale, such as age, temperature, or height. ***Categorical attributes, on the other hand, are non-numeric and can be divided into discrete categories or groups, such as color, gender, or occupation.***

   The main difference between the two is that numeric attributes can be used in mathematical operations and statistical analysis, while categorical attributes cannot. Numeric attributes can also be scaled, making them suitable for distance-based algorithms such as K-means clustering, while categorical attributes require special encoding techniques.

2. ***Feature selection vs. dimensionality reduction:***

   Feature selection and dimensionality reduction are techniques used in machine learning to reduce the number of features or variables in a dataset.

   Feature selection involves selecting a subset of the original features based on their importance or relevance to the target variable. This technique is used to eliminate irrelevant or redundant features, which can reduce overfitting and improve model performance.

   Dimensionality reduction, on the other hand, involves transforming the original feature space into a lower-dimensional space while preserving the important information. This technique is used to address the curse of dimensionality, which can make it difficult to analyze and visualize high-dimensional datasets. Dimensionality reduction techniques include principal component analysis (PCA), singular value decomposition (SVD), and t-distributed stochastic neighbor embedding (t-SNE).

▼ 4. Make quick notes on any two of the following:
1. The histogram
2. Use a scatter plot
3.PCA (Personal Computer Aid)

1. ***The histogram:*** ***A histogram is a graphical representation of the distribution of a dataset.*** It is a way to ***represent the frequency of observations in a given range or bin***. Histograms are commonly used to visualize the distribution of numeric data, such as age, income, or test scores.
2. ***Scatter plot:*** ***A scatter plot is a type of chart that displays the relationship between two continuous variables.*** The data is represented by a collection of points, where each point represents the values of the two variables. Scatter plots are commonly used to investigate the correlation between two variables.

3. ***PCA (Principal Component Analysis):*** ***PCA is a statistical technique used to reduce the number of dimensions in a dataset while retaining the most important information.*** It does this by identifying the principal components, which are the directions in which the data varies the most. PCA is commonly used in data preprocessing to reduce the complexity of a dataset before performing machine learning algorithms.

▼ 5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

It is necessary to investigate data to gain insights, identify patterns, relationships, and trends that can aid decision-making, problem-solving, and hypothesis testing. Exploring data is a critical step in the machine learning process, where data scientists try to extract meaningful insights from the available data.

Exploring quantitative and qualitative data are not inherently different, but the methods used to analyze them may differ. For example, qualitative data may require more interpretive analysis than quantitative data, which may be more amenable to statistical analysis. However, in both cases, the goal is to identify patterns and trends that can help answer research questions or support decision-making. Therefore, while the techniques used may differ, the objective is the same.

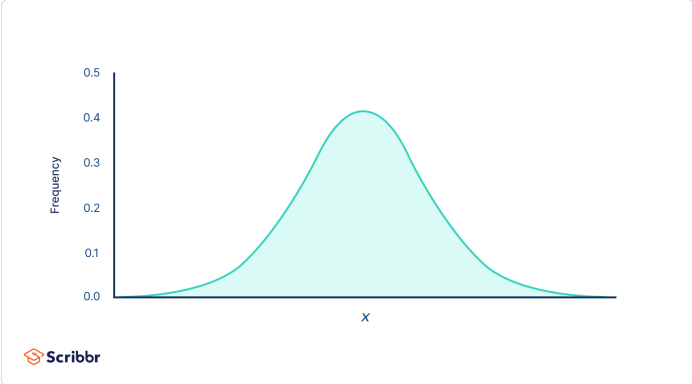
▼ 6. What are the various histogram shapes? What exactly are ‘bins'?

A histogram is a graphical representation of the distribution of a dataset. It is created by dividing the entire range of values in the dataset into a series of intervals or "bins" and then counting how many values fall into each bin.

The various shapes of histograms are:

1. ***Normal Distribution:***

   Also known as Gaussian distribution, it is a bell-shaped curve. The values are concentrated around the mean, with few outliers.



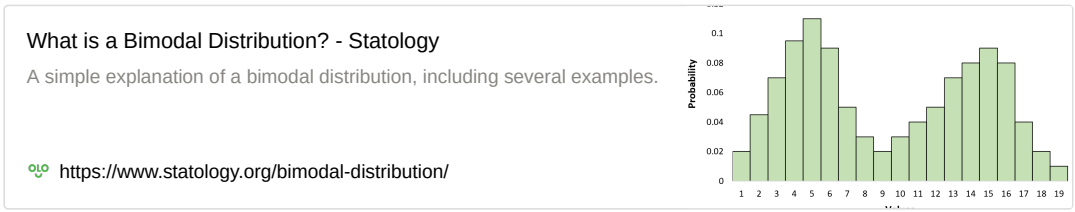
2. ***Skewed Distribution:***

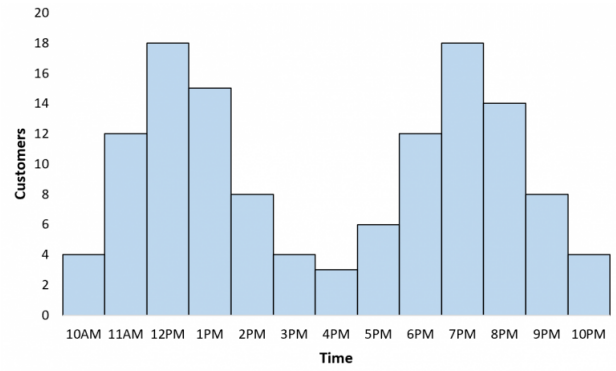
   A distribution is skewed when the data is not symmetric. It can either be positively skewed or negatively skewed.

3. ***Bimodal Distribution:***

   A distribution with two modes is called a bimodal distribution. It occurs when the data consists of two separate groups, each with its own peak.

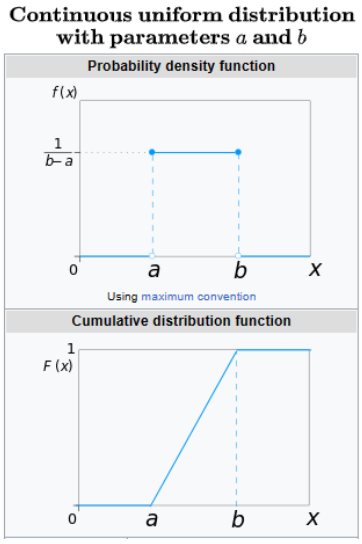
   ▼ **Reference Link**





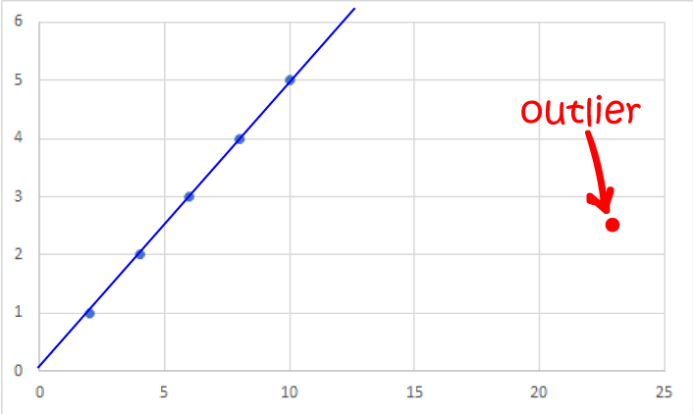
4. ***Uniform Distribution:***

   A uniform distribution occurs when the data is evenly distributed across all possible values.



Bins are the intervals or ranges of values that the data is divided into for the creation of the histogram. The width of each bin is determined by the range of values and the number of bins specified. The number of bins has a significant impact on the resulting histogram's shape and the level of detail shown in the data.

▼ 7. How do we deal with data outliers?



Outliers are data points that are significantly different from other observations in a dataset.
Dealing with outliers is an important step in the data pre-processing stage of machine learning. Here are a few common methods to deal with data outliers:

1. ***Removal:*** ***Outliers can be removed from the dataset***. However, this method can result in the loss of valuable information and must be used with caution.
2. ***Imputation:*** ***Outliers can be replaced with a value that is more representative of the dataset.*** This can be done by using statistical measures such as the mean, median or mode.
3. ***Binning:*** ***Outliers can be placed in a separate bin***, thereby ***reducing their impact*** on the analysis.
4. ***Transformation:*** Data transformation techniques such as logarithmic transformation, square root transformation, or box-cox ***transformation can be applied to the dataset to reduce the effect of outliers.***

It is important to note that the approach to dealing with outliers may vary depending on the type of data and the analysis being performed.

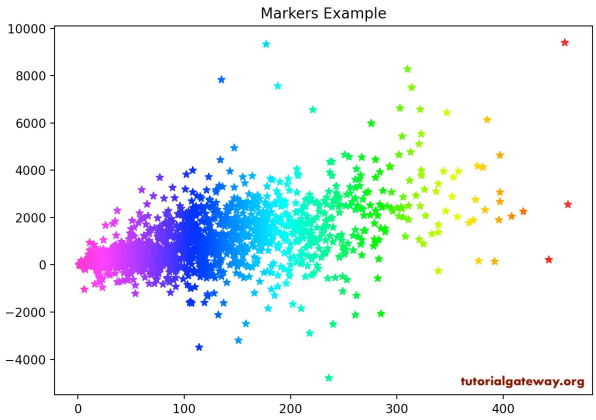
Additionally, there is no significant difference in how qualitative and quantitative data are explored for outliers. In both cases, the data points that deviate significantly from the expected range should be identified and dealt with appropriately.

▼ 8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

The various central inclination measures are mean, median, and mode. Mean is the average value of a dataset, calculated by adding up all values and dividing by the number of observations. Median is the middle value in a dataset when the values are arranged in order, and ***mode is the most common value in the dataset.***

Mean can vary too much from median in certain datasets due to the presence of outliers. Outliers are data points that are significantly different from the rest of the dataset. When outliers are present, the mean can be heavily influenced by their extreme values, while the median remains unaffected. Therefore, in datasets with outliers, the median can be a better measure of central tendency than the mean.

▼ 9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?



A scatter plot is a type of graph used to investigate the relationship between two variables in a bivariate dataset. The two variables are represented on the x-axis and y-axis, with each data point plotted as a dot. The pattern of the dots on the graph can reveal the nature of the relationship between the variables.

For example, ***if the dots cluster closely together and form a straight line, it suggests a strong positive or negative linear relationship between the variables***. ***If the dots form a curved pattern, it suggests a non-linear relationship. If the dots are randomly scattered, it suggests that there is no relationship between the variables.***

Scatter plots can also be used to identify outliers in the data. Outliers are data points that are significantly different from the rest of the data. ***In a scatter plot, outliers may appear as individual data points that are far away from the main cluster of points.*** By identifying outliers, we can investigate whether they are genuine data points or if they are errors in the data that need to be corrected.

▼ 10. Describe how cross-tabs can be used to figure out how two variables are related.

Cross-tabulation, ***also known as contingency table analysis***, ***is a statistical tool that allows for the examination of the relationship between two categorical variables.*** ***It is used to summarize and compare the distribution of two or more variables.*** The basic idea is to tabulate the data in a matrix, with the rows and columns representing the two variables being examined, and the cells containing the frequency or count of each combination of the two variables.

For example, suppose we want to investigate the relationship between gender and voting preference in a particular election. We can use cross-tabulation to see how many men and women voted for each candidate. The resulting table will have two rows (male and female) and several columns (one for each candidate), and the cells will contain the frequency of each combination of gender and voting preference.

Once the table has been constructed, it can be analyzed in various ways to determine the strength and nature of the relationship between the two variables. For example, we can calculate the marginal frequencies (the totals for each row and column), the conditional frequencies (the proportion of each row or column that belongs to a particular category), and the chi-square statistic (a measure of the independence or association between the two variables).

Cross-tabs can also be used to identify patterns and trends in the data, such as the presence of clusters or outliers. ***For example, if there are large differences in the voting patterns of men and women, this might suggest that gender is a significant factor in determining voting preference. If there are a small number of cases that fall into a particular cell of the table, this might indicate that there are outliers or anomalies in the data.***

Overall, cross-tabulation is a useful tool for exploring the relationship between two categorical variables and for identifying patterns and trends in the data.

▼ Reference Link

Contingency Table: Definition, Examples & Interpreting

A contingency table displays frequencies for two categorical variables. Use two-way tables to see relationships between the variables.

α https://statisticsbyjim.com/basics/contingency-table/