**FSDS MAY BATCH 2022(ML Assignment -4)**

**Submitted by: Shubham Tiwari**

Q1: What are the key tasks involved in getting ready to work with machine learning modeling?

Ans: Some key tasks involved in getting ready to work with machine learning modeling include:

1. Defining the problem and determining the desired outcome.
2. Collecting and preparing the data for modelling.
3. Exploring and analyzing the data to gain insights and identify patterns.
4. Selecting and applying appropriate models and algorithms.
5. Tuning and evaluating the models to optimize performance.
6. Deploying the final model for use in a production environment.

Q2: What are the different forms of data used in machine learning? Give a specific example for each of them.

Ans: There are several different forms of data used in machine learning, including:

1. Structured data: This refers to data that is organized in a tabular format, such as a spreadsheet or a database table. An example of structured data would be a dataset of customer information, including their name, address, and purchase history.
2. Unstructured data: This refers to data that does not have a pre-defined format, such as text, images, or audio. An example of unstructured data would be a collection of customer reviews for a product or service.
3. Time-series data: This refers to data that is collected over time, such as stock prices or temperature measurements. An example of time-series data would be daily closing prices of a stock over a period of several years.
4. Semi-structured data: This refers to data that has some structure, but not as much as structured data. An example of semi-structured data would be a dataset of customer reviews that also includes information like date of purchase and product id.
5. Spatial data: This refers to data that has a geographic or spatial component, such as location data or GIS data. An example of spatial data would be a dataset of GPS coordinates for a fleet of delivery vehicles.
6. Streaming data: This refers to data that is generated in real-time and received in a continuous flow, such as social media feeds, sensor data, or log files. An example of streaming data would be a dataset of sensor data from a fleet of vehicles on a highway.

Top of Form

Bottom of Form

Q3: Distinguish:

1. Numeric vs. categorical attributes

Ans: Attributes are the individual features or variables that are used to describe the data. Numeric attributes are attributes that have numerical values, such as age, income, or temperature. Categorical attributes are attributes that have categorical values, such as color, gender, or brand.

Numeric attributes can be further divided into two types:

* Continuous: Numeric attributes that can take any value within a certain range, for example, temperature, weight, height.
* Discrete: Numeric attributes that can take only specific values, for example, the number of children.

**The main difference** between numeric and categorical attributes is the type of data they represent and the types of operations that can be performed on them.

**Numeric attributes** can be used for mathematical operations such as addition, subtraction, and averaging. They can also be used to calculate measures of central tendency such as mean and median, and measures of variation such as standard deviation.

**Categorical attributes** are used to represent non-numerical data, such as categories, labels, or names. Categorical data can be used to calculate frequencies and proportions, and to compare the relative frequencies of different categories.

Categorical data can also be converted into numerical data by using techniques such as one-hot encoding, in which each category is represented by a binary column.

In summary, numeric attributes are numerical features of the data, while categorical attributes are non-numerical features of the data. They are used differently in machine learning and statistics, depending on the type of data they represent, and the operations that can be performed on them.

Top of Form

Bottom of Form

2. Feature selection vs. dimensionality reduction.

Ans: Feature selection and dimensionality reduction are two techniques used to improve the performance of a model by reducing the number of features or variables in the data.

**Feature selection** is the process of selecting a subset of the most relevant features from the original dataset, with the goal of improving the accuracy and interpretability of the model. This is done by identifying the features that have the most impact on the target variable and removing or reducing the impact of irrelevant or redundant features. There are various feature selection techniques such as filter, wrapper and embedded methods.

**Dimensionality reduction** is the process of reducing the number of features in the data by mapping the data to a lower-dimensional space. The goal of dimensionality reduction is to preserve the most important information in the data while reducing the size and complexity of the dataset. There are various dimensionality reduction techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and t-SNE.

**Both feature selection and dimensionality reduction can help to improve the performance of a model by reducing overfitting and increasing the interpretability of the model.** However, they differ in terms of the specific goals they aim to achieve and the methods they use to achieve them.

Feature selection aims to select a subset of the most relevant features from the original dataset, while dimensionality reduction aims to reduce the number of features by mapping the data to a lower-dimensional space. Dimensionality reduction techniques can be used in conjunction with feature selection methods to further improve the performance of the model.Bottom of Form

Q4: Make quick notes on **any two** of the following:

1. The histogram

Ans: A histogram is a graphical representation of the distribution of a dataset, showing the frequency or number of observations that fall within certain ranges of values (also known as bins). The shape of a histogram can provide information about the underlying distribution of the data, and there are several common shapes of histograms, including:

1. **Normal distribution**: A histogram with a normal distribution has a bell-shaped curve, with the majority of the data points concentrated in the middle and tapering off at the ends. This indicates that the data is symmetric and the mean, median, and mode are approximately equal.
2. **Skewed Distribution**: A histogram with a skewed distribution has a tail that extends in one direction, indicating that the data is not symmetric. A histogram with a long tail to the right (positive skew) has a mean that is larger than the median, while a histogram with a long tail to the left (negative skew) has a mean that is smaller than the median.
3. **Multi-modal Distribution**: A histogram with multi-modal distribution has more than one peak, indicating that there are multiple clusters of data points with different values.
4. **Bimodal Distribution**: A histogram with bimodal distribution has two peaks, indicating that there are two clusters of data points with different values.

2. Use a scatter plot

Ans: A scatter plot is a graphical representation of a bivariate relationship between two variables. It is a useful tool for investigating the relationship between two variables by plotting individual data points on a two-dimensional graph, with one variable on the x-axis and the other on the y-axis.

When a scatter plot is used to investigate bivariate relationships, it can reveal patterns and trends in the data, such as:

1. **Positive correlation**: A positive correlation occurs when the values of the two variables increase or decrease together. In a scatter plot, this appears as a upward or downward slope, with the data points forming a line or a cluster.
2. **Negative correlation:** A negative correlation occurs when the values of one variable increase as the values of the other variable decrease. In a scatter plot, this appears as a downward or upward slope, with the data points forming a line or a cluster.
3. **No correlation**: A scatter plot can also reveal no correlation between two variables, where the data points are scattered randomly without any visible pattern or trend.
4. **Outliers:** A scatter plot can also be used to identify outliers, which are data points that fall outside of the general pattern or trend of the data. Outliers can be identified as data points that are far away from the other data points in the plot.

3.PCA (Personal Computer Aid)

Ans: xxxxxxxxxxxxxxx

Q5: Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

Ans: Investigating data is necessary for several reasons:

1. **Data cleaning**: Data investigation is the first step in data cleaning process where you can identify and correct errors, inconsistencies, and missing values in the data. This can help to improve the quality of the data and ensure that it is accurate and reliable for analysis.
2. **Gaining insights**: Data investigation can help to reveal patterns, trends, and relationships within the data that can provide valuable insights into the problem or question being studied. This can help to inform decisions and guide the selection of appropriate models and algorithms.
3. **Identifying outliers**: Data investigation can help to identify outliers, which are data points that are significantly different from the rest of the data. Outliers can have a significant impact on the analysis and results, so it is important to understand their nature and handle them appropriately.
4. **Selecting features:** Data investigation can help to select the most important features for the model and discard the features that are not informative.

The process of investigating data is similar for both qualitative and quantitative data, but the methods used can be different. For quantitative data, statistical methods and visualization techniques are often used to explore the data and identify patterns. For qualitative data, techniques such as content analysis and thematic analysis are often used to explore and interpret the data.

In summary, data investigation is an essential step in the machine learning process, as it helps to understand the data, identify problems, and prepare the data for modeling and analysis. The process of data investigation is similar for both qualitative and quantitative data, but the methods used may differ depending on the type of data.Top of Form

Bottom of Form

Q6: How do we deal with data outliers?

Ans: Dealing with outliers in data can be a challenging task and there are several ways to approach it, depending on the specific situation and the goals of the analysis. Some common methods include:

1. Identifying and removing outliers: This approach involves identifying the outlier data points and removing them from the dataset. This is typically done by using statistical methods to identify data points that fall outside of a certain range or by using visualization techniques to identify data points that are significantly different from the rest of the data.
2. Imputing or replacing outliers: This approach involves replacing the outlier data points with a more reasonable value, such as the mean or median of the dataset. This method can be useful when the outlier data points are a result of data entry errors or other errors in the data collection process.
3. Transforming the data: This approach involves applying a mathematical transformation to the data, such as taking the logarithm or square root, in order to reduce the impact of outliers on the analysis.
4. Anomaly Detection: This approach involves finding the pattern in data and then finding the data point that does not follow the pattern. It can be done by using various techniques like clustering, density-based methods, statistical methods etc.
5. Keeping the outliers: Depending on the nature of the problem, outliers may contain valuable information and therefore, it might be important to keep them in the dataset.Top of Form

Bottom of Form

Q7: What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

Ans: Central inclination measures are methods used to summarize and describe the central tendency or average value of a dataset. Some common central inclination measures include:

1. Mean: The mean is the sum of all the values in a dataset divided by the number of values. It is a commonly used measure of central tendency, but it can be affected by outliers and skewed data.
2. Median: The median is the middle value in a dataset when the data is arranged in order. It is less affected by outliers and skewed data than the mean.
3. Mode: The mode is the most frequently occurring value in a dataset. It is commonly used with categorical data and can be affected by the number of categories.

Mean can vary too much from median in certain data sets when the data set is skewed or contains outliers. Skewness refers to the lack of symmetry in the data distribution, positive skewness means the tail on the right side of the probability density function is longer or fatter and the mean is larger than the median. On the other hand, negative skewness means the tail on the left side of the probability density function is longer or fatter and the mean is smaller than the median. Outliers are extreme values that are significantly different from the rest of the data. These values can have a large impact on the mean, but not as much of an impact on the median.

In conclusion, depending on the nature of the data, different central inclination measures may be more appropriate. Mean is a commonly used measure of central tendency, but it can be affected by outliers and skewed data. Median and mode are less affected by outliers and skewed data and can provide a more accurate representation of the central tendency in such cases.Top of Form

Bottom of Form

Q8: Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

Ans: A scatter plot is a graphical representation of a bivariate relationship between two variables. It is a useful tool for investigating the relationship between two variables by plotting individual data points on a two-dimensional graph, with one variable on the x-axis and the other on the y-axis.

When a scatter plot is used to investigate bivariate relationships, it can reveal patterns and trends in the data, such as:

1. **Positive correlation**: A positive correlation occurs when the values of the two variables increase or decrease together. In a scatter plot, this appears as a upward or downward slope, with the data points forming a line or a cluster.
2. **Negative correlation:** A negative correlation occurs when the values of one variable increase as the values of the other variable decrease. In a scatter plot, this appears as a downward or upward slope, with the data points forming a line or a cluster.
3. **No correlation**: A scatter plot can also reveal no correlation between two variables, where the data points are scattered randomly without any visible pattern or trend.
4. **Outliers:** A scatter plot can also be used to identify outliers, which are data points that fall outside of the general pattern or trend of the data. Outliers can be identified as data points that are far away from the other data points in the plot.

In summary, a scatter plot is a useful tool for investigating bivariate relationships between two variables by revealing patterns, trends, and outliers in the data. It can help to identify the correlation and the strength of the correlation between two variables and also helps to identify outliers in the data which can be further investigated.Top of Form

Bottom of Form

Q9: What are the various histogram shapes? What exactly are ‘bins?

Ans: A histogram is a graphical representation of the distribution of a dataset, showing the frequency or number of observations that fall within certain ranges of values (also known as bins). The shape of a histogram can provide information about the underlying distribution of the data, and there are several common shapes of histograms, including:

1. **Normal distribution**: A histogram with a normal distribution has a bell-shaped curve, with the majority of the data points concentrated in the middle and tapering off at the ends. This indicates that the data is symmetric and the mean, median, and mode are approximately equal.
2. **Skewed Distribution**: A histogram with a skewed distribution has a tail that extends in one direction, indicating that the data is not symmetric. A histogram with a long tail to the right (positive skew) has a mean that is larger than the median, while a histogram with a long tail to the left (negative skew) has a mean that is smaller than the median.
3. **Multi-modal Distribution**: A histogram with multi-modal distribution has more than one peak, indicating that there are multiple clusters of data points with different values.
4. **Bimodal Distribution**: A histogram with bimodal distribution has two peaks, indicating that there are two clusters of data points with different values.

**A bin** is a range of values that the data is divided into in a histogram. The height of each bar corresponds to the number of observations that fall within each bin. The width of the bin represents the range of values that are included in the bin, and the choice of bin width can affect the shape of the histogram and the conclusions that can be drawn from it. A larger bin width will result in fewer bars and a smoother histogram, while a smaller bin width will result in more bars and a more detailed histogram.

Top of Form

Bottom of Form

Q10: Describe how cross-tabs can be used to figure out how two variables are related.

Ans: Cross-tabulation, also known as a contingency table or crosstab, is a method of summarizing and analyzing the relationship between two categorical variables. It involves creating a table that shows the frequency or count of observations for each combination of categories for the two variables.

For example, if the two variables are gender and purchase decision, a cross-tabulation table would show the number of male and female individuals who made a purchase or not.

Cross-tabs can be used to figure out how two variables are related in several ways:

1. **Identifying patterns:** Cross-tabs can reveal patterns and trends in the data, such as which categories of one variable are more likely to be associated with certain categories of the other variable.
2. **Measuring association**: Cross-tabs can be used to measure the strength and direction of association between the two variables. This can be done using measures such as chi-squared test and Cramer's V.
3. **Identifying outliers**: Cross-tabs can be used to identify outliers, which are observations that fall outside of the general pattern or trend of the data.
4. **Identifying independence**: Cross-tabs can be used to identify independence between two variables, which means that the probability of one variable does not depend on the other variable.

In summary, cross-tabulation is a powerful tool for exploring the relationship between two categorical variables by summarizing the frequency of observations for each combination of categories. It can reveal patterns, trends, and outliers in the data, measure the strength and direction of association.

Top of Form

Bottom of Form