**FSDS MAY BATCH 2022(ML Assignment -5)**

**Submitted by: Shubham Tiwari**

Q1: What are the key tasks that machine learning entails? What does data pre-processing imply?

Ans: Machine learning typically entails several key tasks, including data pre-processing, model selection and training, evaluation, and deployment.

Data pre-processing involves cleaning, transforming, and organizing the data to make it suitable for use in a machine learning model. This can include tasks such as missing value imputation, feature scaling, feature selection, and data transformation. The goal of data pre-processing is to prepare the data in a format that can be easily used by the machine learning algorithms and to improve the performance of the model by removing noise and redundant information.

Q2: Describe quantitative and qualitative data in depth. Make a distinction between the two.

Ans: **Quantitative data refers to** data that can be measured and expressed in numbers. It is data that can be analyzed using statistical methods. Examples of quantitative data include measurements of weight, height, time, temperature, and number of items in a set.

**Qualitative data, on the other hand, refers** to data that is descriptive and cannot be measured in numbers. It is used to understand and explain the underlying reasons, opinions, and motivations behind people's actions and decisions. Examples of qualitative data include interview transcripts, open-ended survey responses, and observations of behavior.

**The main difference between the two is that** quantitative data is numerical and can be analyzed using mathematical and statistical techniques, while qualitative data is non-numerical and is analyzed through methods such as content analysis and thematic analysis. Additionally, quantitative data is typically collected through structured methods such as surveys, experiments, and polls, while qualitative data is often collected through unstructured methods such as interviews, observations, and focus groups.

Q3: Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

Ans: **Name, Age, Income, Gender, Purchased**

Arun, 25, 45000, Male, Yes

Dilip, 32, 52000, Female, No

Shreya, 40, 60000, Male, Yes

Shubham, 28, 48000, Female, No

Varun, 36, 55000, Male, Yes

* "Name" is a categorical attribute
* "Age" is a numerical attribute
* "Income" is a numerical attribute
* "Gender" is a categorical attribute
* "Purchased" is a categorical attribute (binary, yes or no)

Q4: What are the various causes of machine learning data issues? What are the ramifications?

Ans: There are several causes of machine learning data issues, including:

1. **Missing data**: This occurs when some observations in the dataset are missing one or more variables. This can lead to bias and reduced accuracy in the model.
2. **Outliers:** These are extreme values that deviate significantly from other observations in the dataset. Outliers can have a large impact on the model, leading to overfitting or poor generalization.
3. **Noise**: This refers to random variations in the data that do not represent the underlying signal. Noise can lead to overfitting and poor generalization of the model.
4. **Non-representative data**: This occurs when the dataset used for training the model does not accurately represent the population it is intended to make predictions about. This can lead to poor model performance and inaccurate predictions.
5. **Data imbalance**: This occurs when one class in a dataset has much more observations than other class, leading to the model being biased towards the class with more observations.

**Ramifications of these issues include**: poor model performance, inaccurate predictions, and bias in the model. These issues can also make it difficult to identify the underlying patterns in the data, leading to incorrect insights and decisions.

Q5: Demonstrate various approaches to categorical data exploration with appropriate examples.

Ans: There are several approaches to exploring categorical data, including:

1. **Frequency tables**: A frequency table shows the number of occurrences of each category in a variable. For example, if we have a variable "gender" with categories "male" and "female", a frequency table would show the number of males and females in the dataset.
2. **Bar plots:** A bar plot is a graphical representation of a frequency table. It shows the number of occurrences of each category as bars. For example, a bar plot of the "gender" variable would show the number of males and females in the dataset as bars.
3. **Pie charts:** A pie chart is a circular graphical representation of a frequency table. It shows the proportion of each category to the whole. For example, a pie chart of the "gender" variable would show the proportion of males and females in the dataset as wedges of a circle.
4. **Cross-tabulation**: Cross-tabulation (or contingency table) is a table that shows the frequency distribution of two or more categorical variables. For example, if we have two variables "gender" and "education" (with categories "male", "female" and "high school", "college", "graduate"), a cross-tabulation table would show the number of males and females in each education level.
5. **Stacked bar chart**: A stacked bar chart is a bar chart where bars are divided into segments to show the proportion of different categories. For example, if we have a variable "gender" and "education" a stacked bar chart would show the proportion of males and females in each education level.
6. **Box plot**: A box plot is a standardized way of displaying the distribution of data based on five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can be used to compare multiple categorical data.

Top of Form

Bottom of Form

Q6: How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Ans: If certain variables have missing values, the learning activity may be affected in several ways. Depending on the amount and distribution of missing values, the model's performance may be negatively impacted. For example, if a large proportion of the data is missing for a particular variable, the model may not be able to effectively learn from that variable. Additionally, if the missing values are not handled properly, the model's predictions may be biased.

There are several methods to handle missing values, including:

1. Removing rows with missing values: This approach can work well if the proportion of missing values is low, but if too many rows are removed, it can lead to a reduction in sample size and a potential loss of valuable information.
2. Imputing missing values: This approach involves replacing missing values with estimated values based on the non-missing values. Common methods of imputation include mean imputation, median imputation, and multiple imputation.
3. Using algorithms that can handle missing values: Some machine learning algorithms, such as random forests, are able to handle missing values in the data without the need for imputation.

Ultimately, the choice of how to handle missing values will depend on the specific dataset and the requirements of the learning task.

Top of Form

Bottom of Form

Q7: Describe the various methods for dealing with missing data values in depth.

Ans: There are several methods to handle missing values, including:

1. **Removing rows with missing values**: This approach can work well if the proportion of missing values is low, but if too many rows are removed, it can lead to a reduction in sample size and a potential loss of valuable information.
2. **Imputing missing values:** This approach involves replacing missing values with estimated values based on the non-missing values. Common methods of imputation include mean imputation, median imputation, and multiple imputation.
3. **Using algorithms that can handle missing values:** Some machine learning algorithms, such as random forests, are able to handle missing values in the data without the need for imputation.

Q8: What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

Ans: Data pre-processing techniques include tasks such as data cleaning, normalization, and feature selection or extraction.**Dimensionality reduction** is a technique used to reduce the number of variables or features in a dataset by projecting the data onto a lower-dimensional space. This can be done through methods such as principal component analysis (PCA) or linear discriminant analysis (LDA).

**Feature selection or extraction** is the process of selecting a subset of relevant features from the original dataset to use in building a model. This can be done through methods such as mutual information or linear regression.

Q9**: i**. What is the IQR? What criteria are used to assess it?

Ans: The interquartile range (IQR) is a measure of the dispersion of a dataset, defined as the difference between the third quartile (Q3) and the first quartile (Q1). It represents the spread of the middle 50% of the data and is a robust measure of variability, as it is not affected by outliers or extreme values.

To calculate the IQR, we first need to find the first and third quartiles of the data. The first quartile (Q1) is the value that separates the lowest 25% of the data from the rest, and the third quartile (Q3) is the value that separates the highest 25% of the data from the rest. The IQR is then calculated by subtracting Q1 from Q3.

IQR is often used to assess the spread of a dataset and to identify outliers. Outliers are considered as observations that are outside of the range of (Q1 - 1.5 \* IQR) and (Q3 + 1.5 \* IQR) . This is known as the Tukey's fences method. In some cases, it is also used to detect a possible skewness on the data.

**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?

Ans: A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that displays the five-number summary of the data. The five-number summary includes the minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

The box in the box plot represents the interquartile range (IQR), which is the range between the first and third quartiles. The line inside the box represents the median of the data. The "whiskers" extend from the box to the minimum and maximum values, with the exception of outliers, which are plotted individually beyond the whiskers.

The lower whisker will surpass the upper whisker in length when the lower whisker extends further from the box than the upper whisker. This can happen when there are more outliers in the lower half of the data than the upper half.

Box plots can be used to identify outliers by looking for data points that fall outside of the whiskers. Outliers are typically defined as data points that are more than 1.5 times the IQR away from the first or third quartile.

Q10: Make brief notes on **any two** of the following:

1. Data collected at regular intervals

Ans: Data collected at regular intervals is known as time series data. Time series data is a set of data points collected at specific intervals of time, such as daily, hourly, or every minute. This type of data is often used in fields such as finance, economics, weather forecasting, and manufacturing. Preprocessing time series data could involve several steps such as handling missing values, resampling to a uniform time frequency, and detrending the data.

Additionally, in time series modeling, feature engineering is an important step in which time-based features such as day of the week, month, and other temporal patterns are extracted and included as input features.

Another important step is the Time series Forecasting, which is the use of a model to predict future values based on previously observed values. This is a common task in time series analysis, and there are many different methods and models that can be used, such as ARIMA, exponential smoothing, and neural networks.

2. The gap between the quartiles

Ans: **The gap between the quartiles, also known as the interquartile range (IQR), is a measure of the dispersion of a dataset**. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1).The first quartile (Q1) is the value that separates the lowest 25% of the data from the highest 75% of the data. The third quartile (Q3) is the value that separates the lowest 75% of the data from the highest 25% of the data. The IQR is a robust measure of the spread of a dataset, as it is not affected by outliers.

The IQR can be used to identify outliers in a dataset. An outlier is a value that falls outside of the range of the IQR. It is defined as a data point which is more than 1.5\*IQR below Q1 or above Q3. This is known as the Tukey's method.

In summary, the IQR is a measure of the spread of a dataset, calculated as the difference between the third quartile and the first quartile. It can be used to identify outliers in a dataset and is a robust measure of spread, not affected by outliers.

3. Use a cross-tab

Ans:xxxxxxxxxxxxxxxxxxxxxxxxxx

Q11: Make a comparison between:

1. Data with nominal and ordinal values.

Ans: **Nominal data are categorical data that do not have** a natural order or ranking. They are used to represent categories or labels that do not have any inherent order or numerical value. Examples of nominal data include gender, eye color, and religious affiliation.

**Ordinal data are categorical data that have a** natural order or ranking. They are used to represent categories or labels that can be ordered or ranked in some way. Examples of ordinal data include educational level (e.g. high school, college, graduate school), income level (e.g. low, medium, high), and customer satisfaction ratings (e.g. very satisfied, satisfied, neutral, unsatisfied, very unsatisfied).

**The main difference between nominal and ordinal data is** the way they are represented and the types of operations that can be performed on them.

**Nominal data can be represented** using nominal scale, where each category is represented by a unique label or name, and the categories do not have any inherent order or numerical value. Nominal data can be used for descriptive statistics such as counting the frequency of each category, and for inferential statistics such as chi-squared tests.

**Ordinal data can be represented** using ordinal scale, where each category is represented by a unique label or name, and the categories have a natural order or ranking. Ordinal data can be used for descriptive statistics such as counting the frequency of each category, and for inferential statistics such as chi-squared tests and non-parametric tests such as Wilcoxon rank-sum test.Top of FormBottom of Form

2. Histogram and box plot.

Ans: Histograms and box plots are two types of graphical representations used to display and summarize the distribution of a set of numerical data.

**A histogram is a** graphical representation of the distribution of a set of numerical data. It is a bar chart that shows the frequency of the data in different bins or intervals. The x-axis represents the intervals or bins, and the y-axis represents the frequency of the data in each interval. Histograms are useful for visualizing the overall shape of a distribution, as well as identifying any outliers or skewness in the data.

**A box plot**, also known as a box-and-whisker plot, is a graphical representation of the distribution of a set of numerical data. It is a standardized way of displaying the distribution of data based on five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). The box in the box-plot represents the interquartile range (IQR) which is the range of the middle 50% of the data, where Q1 is the 25th percentile and Q3 is the 75th percentile. The vertical lines or whiskers represent the minimum and maximum values of the data, excluding outliers. Box plots are useful for comparing the distribution of different sets of data, as well as identifying any outliers or skewness in the data.Bottom of Form

3. The average and median.

Ans: Average (or mean) and median are two measures of central tendency that are used to summarize a set of numerical data. They are used to describe the "center" or "typical" value of a dataset.

**The average, or mean, is calculated by adding up all the values in a dataset and dividing by the total number of values.** It is sensitive to outliers (values that are much larger or smaller than the other values in the dataset) and can be affected by the presence of extreme values.

**The median is the middle value of a dataset when it is arranged in order of magnitude.** It is calculated by arranging all the values in a dataset in order and finding the value that is exactly in the middle of the dataset. The median is not affected by outliers and gives a better representation of the center of the dataset when the data has outliers.

For example, let's say we have a dataset of 5 numbers (1,2,3,4,5) the mean is 3 and the median is 3. If we add an outlier (10) to the dataset, the mean becomes 4 and the median is still 3. In general, when the data is symmetrical and does not have outliers, the mean and median will be similar, but when the data is skewed or has outliers, the median will be a better measure of central tendency.