**ML Assignment\_5**

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

Machine Learning is **gaining some useful information from the data**. Usually, Machine Learning is of two types Supervised Learning and Unsupervised Learning. Classification and Regression are examples of Machine Learning. The task of classification is to predict what class an instance of data should fall into

Data preprocessing is **the process of transforming raw data into an understandable format**. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

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

**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 measurement involves running statistical analysis on data that has numerical values. Qualitative measurement looks for patterns in non-numerical data.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Types | Default Task | Attribute Types | # Instances | # Attributes | Year |
| Multivariate | Classification | Categorical, Integer, Real | 4177 | 8 | 1995 |
| Multivariate | Classification | Categorical, Integer | 48842 | 14 | 1996 |
| Multivariate | Classification | Categorical, Integer, Real | 798 | 38 |  |

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

**5 Common Machine Learning Problems & How to Solve Them**

* 1) Understanding Which Processes Need Automation.
* 2) Lack of Quality Data.
* 3) Inadequate Infrastructure.
* 4) Implementation.
* 5) Lack of Skilled Resources
* The ramifications are **the broader effects that fan out into the world from one situation, or decision, that kicks it all off**. Ramification also refers to something branching out, like limbs on a tree — which is what bad decisions tend to do. Ramification is like consequence, but usually unintended and bad.

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

**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 :**

* **Mode :**Most frequently occurring value in the given data **Example-**

Data = ["Car", "Bat", "Bat", "Car", "Bat", "Bat", "Bat", "Bike"]

Mode = "Bat"

* **Expected Value :**When working in machine learning, categories have to be associated with a numeric value, so as to give understanding to the machine. This gives an average value based on a category’s probability of occurrence i.e. Expected Value. It is calculated by –

-> Multiply each outcome by its probability of occurring.

-> Sum these values

* So, it is the sum of values times their probability of occurrence often used to sum up factor variable levels.
* **Bar Charts :**Frequency of each category plotted as bars

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

**Many machine learning algorithms fail if the dataset contains missing values**. However, algorithms like K-nearest and Naive Bayes support data with missing values. You may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly

**Handling `missing` data**

1. Use the 'mean' from each column. Filling the NaN values with the mean along each column. [ ...
2. Use the 'most frequent' value from each column. Now let's consider a new DataFrame, the one with categorical features. ...
3. Use 'interpolation' in each column. ...
4. Use other methods like K-Nearest

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

One way of handling missing values is the **deletion of the rows or columns having null values**. If any columns have more than half of the values as null then you can drop the entire column. In the same way, rows can also be dropped if having one or more columns values as null

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

**Important Data Preprocessing Techniques**

* Data Cleaning.
* Dimensionality Reduction.
* Feature Engineering.
* Sampling Data.
* Data Transformation.
* Imbalanced Data.

Dimensionality reduction technique can be defined as, "**It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information**." These techniques are widely used in machine learning for obtaining a better fit predictive model while solving the classification

**9.**

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

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

**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?**

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. In a box plot, we draw a box from the first quartile to the third quartile.

If the left whisker is longer than the right whisker, **the distribution is negatively skewed**. The length of the whiskers also gives you information about how spread out the data is. A box-and-whisker plot is often used when the number of data values is large

In the boxplot below, the length of the box is IQR, and the minimum and maximum values are represented by the whiskers. The whiskers are generally extended into 1.5\*IQR distance on either side of the box. Therefore, **all data points outside these 1.5\*IQR values are flagged as outliers**.

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

1. **Data collected at regular intervals**

**Interval data is measured along a numerical scale that has equal distances between adjacent values**. These distances are called “intervals.” There is no true zero on an interval scale, which is what distinguishes it from a ratio scale

**2. The gap between the quartiles**

1. **Use a cross-tab**

Cross-tabulation is **a mainframe statistical model that follows similar lines**. It helps you make informed decisions regarding your research by identifying patterns, trends, and the correlation between your study parameters. When conducting a study, the raw data can usually be daunting.

**11. Make a comparison between:**

1. **Data with nominal and ordinal values**

**Nominal data is classified without a natural order or rank, whereas ordinal data has a predetermined or natural order**. On the other hand, numerical or quantitative data will always be a number that can be measured

1. **Histogram and box plot**

Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data. The 'whiskers' of a box plot show the least and greatest values in the data set

1. **The average and median**

The average is calculated by adding up all of the individual values and dividing this total by the number of observations. The median is calculated by taking the “middle” value, the value for which half of the observations are larger and half are smaller