**ML Assignment\_4**

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

* Collecting Data: As you know, machines initially learn from the data that you give them. ...
* Preparing the Data: After you have your data, you have to prepare it. ...
* Choosing a Model: ...
* Training the Model: ...
* Evaluating the Model: ...
* Parameter Tuning: ...
* Making Predictions.

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

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**.

Numerical data, or quantitative data, is any form of measurable data such as your height, weight, or the cost of your phone bill. You can determine if a set of data is numerical by attempting to average out the numbers or sort them in ascending or descending order. Exact or whole numbers (ie. 26 students in a class) are considered discrete numbers, while those which fall into a given range (ie. 3.6 percent interest rate) are considered continuous numbers. While learning this type of data, keep in mind that numerical data is not tied to any specific point in time, they are simply raw numbers

Categorical data is sorted by defining characteristics. This can include gender, social class, ethnicity, hometown, the industry you work in, or a variety of other labels. While learning this data type, keep in mind that it is non-numerical, meaning you are unable to add them together, average them out, or sort them in any chronological order. Categorical data is great for grouping individuals or ideas that share similar attributes, helping your machine learning model streamline its data analysis

Time series data consists of data points that are indexed at specific points in time. More often than not, this data is collected at consistent intervals. Learning and utilizing time series data makes it easy to compare data from week to week, month to month, year to year, or according to any other time-based metric you desire. The distinct difference between time series data and numerical data is that time series data has established starting and ending points, while numerical data is simply a collection of numbers that aren’t rooted in particular time periods.

Text data is simply words, sentences, or paragraphs that can provide some level of insight to your machine learning models. Since these words can be difficult for models to interpret on their own, they are most often grouped together or analyzed using various methods such as word frequency, text classification, or sentiment analysis

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

Categorical data refers to a data type that can be stored and identified based on the names or labels given to them. Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form. Also known as qualitative data as it qualifies data before classifying it

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

Feature Selection vs Dimensionality Reduction

While both methods are used for reducing the number of features in a dataset, there is an important difference. **Feature selection is simply selecting and excluding given features without changing them.** Dimensionality reduction transforms features into a lower dimension.25

**4. Make quick notes on any two of the following:**

1. **The histogram**

A histogram is **a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size**. In the most common form of histogram, the independent variable is plotted along the horizontal axis and the dependent variable is plotted along the vertical axis

1. **Use a scatter plot**

A scatter plot is **a set of points plotted on a horizontal and vertical axes**. Scatter plots are important in statistics because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables)

**3.PCA (Personal Computer Aid)**

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed. The underlying data can be measurements describing properties of production samples, chemical compounds or reactions, process time points of a continuous process, batches from a batch process, biological individuals or trials of a DOE-protocol,

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

**It helps determine how best to manipulate data sources to get the answers you need**, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumption.

**The combination of qualitative and quantitative data can also lead to clashes in the philosophical assumptions behind each approach** and therefore recommendations have been made for triangulation to be carried out from a pragmatic, or subtle realist, approach

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

A histogram is a chart that **plots the distribution of a numeric variable's values as a series of bars**. Each bar typically covers a range of numeric values called a bin or class; a bar's height indicates the frequency of data points with a value within the corresponding bin

**7. How do we deal with data outliers?**

**5 ways to deal with outliers in data**

1. Set up a filter in your testing tool. Even though this has a little cost, filtering out outliers is worth it. ...
2. Remove or change outliers during post-test analysis. ...
3. Change the value of outliers. ...
4. Consider the underlying distribution. ...
5. Consider the value of mild outliers

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

There are three main measures of central tendency: **the mode, the median and the mean**. Each of these measures describes a different indication of the typical or central value in the distribution.

Consequently, **when some of the values are more extreme, the effect on the median is smaller**. Of course, with other types of changes, the median can change. When you have a skewed distribution, the median is 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?**

**Scatter plots are a tool you can use to display bivariate data**. You can make coordinate pairs using the data

**If there is a regression line on a scatter plot, you can identify outliers**. An outlier for a scatter plot is the point or points that are farthest from the regression line. There is at least one outlier on a scatter plot in most cases, and there is usually only one outlier

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

To describe the relationship between two categorical variables, we use a special type of table called a cross-tabulation (or "crosstab" for short). In a cross-tabulation, **the categories of one variable determine the rows of the table, and the categories of the other variable determine the columns**.