Machine\_learning\_Assignment\_4

**Q – 1 What are the tasks involved in getting ready to work with machine learning modelling?**

Ans – A machine learning task is the type if prediction or inference being made, based on the problem or question that is being asked, and the available data. For example, the classification task assigns data to categories, and the clustering task groups data according to similarity.

**Q – 2 What are the different form of data in machine learning? Given a specify example for each of them.**

Ans – Data can come in many forms, but in machine learning models rely on four primary types of data like –

* **Numerical data** (Example – Height, weight, age number of movies watched, IQ etc)
* **Categorical data** (Example – Survey could ask the random group of data)
* **Time series** (Example – information sequence that was taken at specify time travel like anything from heart rate to the unite pries of store goods
* **Text data** (Example – “new” and “NEW” variation in the word form, “walk” and “walking”)

**Q – 3 Distinguish**

**Ans – Numeric and categoric attributes: - Numerical** data are values obtained for quantitative variable, and carries a sense of magnitude related to the context of the variable (hence, they are always number or symbols carrying a numerical value). Categorical data are values obtained for a qualitative variable, categorical data numbers do not carry a sense of magnitude.

Numerical data always belongs to either ordinal, ratio, or interval type whereas categorical data belong to nominal type.

Method used to analyse quantitative data are different from the methods used for categorical data, even if the principles are the same at least the application has significant differences,

Numerical data are analysed using statical method in descriptive statistic regression. Time series and many more.

For categorical data usually descriptive method and graphical method are employed. Some non-parametric tests are also used.

**Feature selection**

Feature selection is the process of reducing the number of the input variable when developing a predictive model.

It is a desirable to reduce the number of input variable to both reduce the computational cost of modelling and in some case to improve the performance of the model.

Statistical based feature selection method involving evaluating the relationship between each input variable and the target variable using statistical.

And selection those input variable that have the strongest relationship between with target variable.

**Dimensionality Reduction**

Dimensionality Reduction as the name suggest is the process of transforming the feature into a lower dimension.

It projects the data into lower dimensionality space. That us turn can work quit well or not for your classification algorithm.

Via reduction the Dimensionality of the feature we aim to obtain a set of principle feature.

These features are then used for training the machine learning model in a much more accurate manner.

Q – 4 Make quick notes of any two following.

Ans- **1. Histogram** = Histograms group the data in bins and is the fastest way to get idea about the distribution of each attribute in dataset. The following are some of the characteristics of histograms: -

* It provides us a count of the number of observation in each bin created for visualization.
* From the shape of the bin, we can easily observe the distribution.
* Histogram also help us to see possible outliers.

**2. Use of scatter plot =** All you need are your scatter plot xy variables that can be used to show relationships that exist between them both weather you are trying to improve processes or make a strong argument about your marketing focus, These point are easy and effective way of visualizing where you are now, where you have been and where you want too go. You’ll also learn about scatter plot and correlation throughout this article. Plus a big advantage is that you can learn how to make a scatter plot in machine learning model as well as excel and google sheet.

**Q – 5 Why it is necessary to investigate the data? Is there a discrepancy in hoe qualitive and quantitative data are explored?**

**Ans –** Effective data collection and analysis will allow you to direct scare resources where

They are most needed. If an increase in significant incidents is noted in a particular servise area, this data can be dissected further to determine whether the increase is widespread or isolate to a particular site.

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

**Q – 6 what are the various histogram shape? What exactly are bins?**

**Ans –** The histogram shape is –

* Bell – shaped
* Bimodal
* Skewed right
* Skewed left
* Uniform
* Random

A histogram is a graphical representation of the distribution of data given by the user. Its appearance is similar to bar graph except it is continuous. The towers or bars of a histogram are called bins. The height of each bin show how many value from that data into that range.

**Q – 7 How do we deal with the data outliers?**

**Ans –** Deals with the data outliers –

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

**Q – 8 What is the various central inclination measure? Why does mean vary too much from median on certain data sets?**

**Ans –** There are three main measure of central inclination tendency –

* Mean
* Median
* Mode

Mean is simple and can be applied to any data array sets, whether even or odd. The median is slightly complex and the data needs to be arranged in ascending or descending order before calculation. The mean is generally used for normal distribution.

**Q – 9 Describe how a scatter plot can be used investigate bivariant relationship. It is possible to find outliers by using scatter plot?**

Ans – I suppose you could if one variable of the equation is in a set group variability, but generally a scatter plot does not have any set variables. But if a scatter plot goes straight and down where all of the points have same x coordinate, this would mean that all of the points are in single fine line with no variation, then you can’t really consider that a scatter plot. At that point is just a line. But if there is some variation of the points, them being spread out a little. Them yes this is a scatter plot.

***The bivariant data set shows the linear relationship if the scatter plot shows point bunched randomly around the straight line. The pint might be tight bunched and fall almost exactly on a line, or they be widely scattered, and forming a cloud of point.***

If there is a regression line on scatter plot, you can identify outilers. 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 case, and there is usually only one outlier.

**Q – Describe how cross-tabs can be used to figures out how two variables are related.**

Ans – Cross tabulation is used to quantitatively analyses the relation ship between multiple variable. Cress tabulation also referred to as contingency tables or crosstabs group variable together and enable researchers to understand the correlation between the different variable.