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

Data collection

Data cleaning and preprocessing

Feature selection

Model training

Hyper parameter tuning

Deployment

1. What are the different forms of data used in machine learning. Give examples

Numerical data, categorical data, time series

1. Distinguish-

numerical vs categorical attributes

numerical attributes(quantitative data) means continuous data like height ,weight or any numerical quantity whereas categorical data (qualitative data)is a category means hair color, quality of food

to process categorical data in ml encoding is performed where certain numerical value is assigned to different categories of the data.categorical data is of two types-

ordinal -the categories have ranks like good better best

nominal- data which has name labels like name of a person, hair color

feature selection vs dimensionality reduction

feature selection- means only selection those features which define new pattern and can be used by our model to study new relationships between data for proper training and dropping those that serve no purpose

for example – for two highly correlated data we drop one of them

feature selection is done so as to reduce time complexity

dimensionality reduction- means arranging the axis of the data in such a way that the original meaning of the data is retained and no feature is dropped. It is not an elimination method but a reduction method.eg PCA(principle component analysis)

4.define

Histogram

A graph which displays frequency of occurrence of a certain data

Use of scatter plot –

Scatter plot is a plot where data is represented by a dot.

Scatter plot is used to visualize and observe relationship between two different dataset

PCA

means arranging the axis of the data in such a way that the original meaning of the data is retained and no feature is dropped. It is not an elimination method but a reduction method

steps for PCA-

1. Always check if you have standard normal distribution or not, if not then use standardization (z score method) to make mean=0, variance=1
2. Draw a best fit line which describe relation between features which is called as principle component axis
3. PCs are drawn until pcs are able to define 90% of given relationships in data which are measured by EVR(explained variance ratio) which is given by

EVR(pc1)=distance of PC1 pts/pc1+pc2\_...pcn

5 why is it necessary to investigate data?is there a discrepancy of how qualitative and quantitative data is explored.

6.what are various histogram shapes. What are bins

A histogram displays numerical data by grouping data into "bins" of equal width.

Shapes are-bell curve, right skewed, left skewed, uniform, random, standard

7.how to deal with outliers

There are various methods to remove outliers from the data

1. Z score-

Z-score is just the number of standard deviations away from the mean that a certain data point is.

For example if data is 1,2,3,4,5 mean=3

Then zscore of each point is

|  |  |
| --- | --- |
| 1 | -2 |
| 2 | -1 |
| 3 | 0 |
| 4 | 1 |
| 5 | 2 |
|  |  |
|  |  |

Now we can put a threshold value for z score, any value having z score greater than threshold value (means that is many standard deviations away from our mean) will be considered as outlier and can be removed.

Generally the threshold value is considered as 3.

1. Hypothesis testing
2. IQR method

5-point summary

For any given dataset we sort it and try to find these five points

Minimum

First quartile (25%) q1

Second quartile (50)q2

Third quartile (75)q3

maximum

now we find IQR=q3-q1

then we determine lower fence=q1-1.5(IQR)

higher fence=q3+1.5(IQR)

The data which does not lie in this range of lower and higher fence will be considered as an outlier.

1. DBSCAN-

In DBSCAN we make clusters of data based on epsilon that we define and minimum points required to form a cluster, for the data points which are not able to form any cluster are considered as noise or outliers.

1. Standard deviation
2. By visualizing the data

By drawing a box plot we can visualize the position of outlier.

8.what are various central inclination measures? Why do mean vary too much from median for certain dataset.

9.describe how scatter plot can be used to investigate bivariate relationships.is it possible to find outliers using scatter plot?

A scatter plot shows our data in form of points it can be used to determine relationship between two data points by simply analysing shape of the dataset for eg if graph with x and y is straight line with increasing slope means x and y share linear relationship

Yes we can determine outlier as well by visualizing given data , if data is far away from other data points that means that data is an outlier.

10.describe how cross tags can be used to figure out how two variables are related.?