ML question

Q1. What is K in K-Means? How do you arrive at an optimal K? What is plotted in X and Y axis of an Elbow plot? Explain what the Y-axis component means and how it is computed.

Answer: a) K-means clustering searches the clusters with an unlabeled multi-dimensional dataset.

Assumption: 1) Cluster center is the mean of all the points.   
 2) Each point is closest to its own cluster.

It works on expectation maximization approach. In K-means clustering k represents the number of clusters.

b) To find the optimal number of clusters in k means cluster we can use elbow method or silhouette score.

c) In the elbow method plots within cluster sum of square again no. of cluster(k)

d) Y is the sum of all the within cluster sum of square distance. i.e sum of all the distances of each data points to their own clusters.

Q2. How would you assess a Linear Regression model? In other words, how would you know if the model is good or bad? What parameters do you look for?

Answer: To evaluate the performance of linear regression model there are various evaluation metrics such as

1) MSE(mean square error): This is the average of sum square error. Here the value closer to zero indicates a good model.

2) RMSE(Root mean squared error): This is the root of MSE.

3) R2 score: This is also called as coefficient of determination where value close to 1 indicate better model.

4) Adjusted R2 score: In r2 score as we increase the coefficient the score goes on increasing irrespective of the perform. This can misguide the researchers.

To overcome this adjusted r2 score is used which penalizes irrelevant coefficient and only increase the scores when the coefficient adds value to the model.

Q3. How would you assess a Classification model?

Answer: To evaluate the performance of a classification model there are various evaluation metrics such as

1) Accuracy: This is the common method. Accuracy above 0.5 is for class 1 and accuracy below 0.5 is class 0

2) Precision: It is the no. of True positive to the no of predicted positive.

3) Recall: This is the no. of True positive to the total no. of actual positive.

4) F1 score: This is the mean of precision and recall/

5) Confusion matrix: The is the table containing True positive, False positive, true negative, false negative.

6) AUC-ROC: auc(area under curve) is the ability of the classifier to distinguish between the classes and summaries Roc

Roc(receiver operator characteristics) is the probability curve that plots True positive rate vs false positive rate. It also separates signals and noises.

Q4. What is the difference between Association and Clustering? Give examples.

Answer: Association is discovering patterns and relations between the variables.

While clustering is grouping similar data points together

For clustering several models likes k-means, hierarchical clustering are used. For example, Netflix tv shows and movies clustering.

Q5. In a Linear Regression model, what is a 'coefficient estimate' and how do you interpret it? Also, how do you interpret the 'constant' in the equation?

Answer: For equation o linear regression is: "Y = β0 + β1x1 + β2x2 + β3x3 + ε"

Here "Y" is the dependent variable

β0 represents the intercept or the constant

β1,β2,β3,.... are the slope or the coefficient

ε is the irreducible error

If β0 is not present then out model will always pass through origin

β1 or the intercept represent the change in dependent variable or 1 unit increase in independent variable. If β1>0 then x and y both will be positive and if β1<0 then x and y both will be negative.

Q6. What is the difference between Supervised and Unsupervised learning? Give an example for each.

Answer: supervised learning is a type of machine learning where a model is trained on a label data set where the labels are the correct output for each inputs the model learns to make prediction based on the input data and the corresponding labels.

For example:- I want to predict the house price, now house price is dependent on several factors such as location, sqrt feet, etc. Here price is my dependent variable which i want to predict and rest of the of other are independent variable i.e location is independent of sqrt feet and vice versa. As here we will be providing the independent and dependent variable both to the training of the model. house price prediction, Public transportation demand predication, health insurance cross sell, etc. Are the examples of supervised machine learning.

unsupervised learning is a type of machine learning where the model is trained on an unlabeled dataset and is tasked with finding patterns and structures in the data on its own without any prior knowledge of the correct output.

For example:- I am given with a dataset that contains all the tv shows, movies on Netflix. Now I want to segregate them based on genre, age group, ratings, etc. So these are all my independent variable and it has no dependent variable. My model will learn the patterns and the structures inside this data on its own. So this kind of data is and unsupervised machine learning data.