1. What is the definition of a target function? In the sense of a real-life example, express the target function. How is a target function's fitness assessed?

A: Target function describes the relation between input and output. We can think of target function as an ideal model.

Let's say sales of a product appears to be three times the marketing cost after you trained a linear regression model using sales and market cost. In this case (sales = marketing cost* 3) is the target function.

Target function's fitness can be assessed using hypothesis testing.

2. What are predictive models, and how do they work? What are descriptive types, and how do you use them? Examples of both types of models should be provided. Distinguish between these two forms of models.

A: Predictive models take historical data as input to predict future behaviour. For example, we can predict whether a borrower will default or not based on historical data.

Descriptive models are used to understand the relations between variables or attributes. For example, market analysis is done to understand the products that can be sold together.

In predictive models, there is a target that must be predicted while it's not the case for descriptive models.

3. Describe the method of assessing a classification model's efficiency in detail. Describe the various measurement parameters.

A: There are many measures that can be used to assess the efficiency of classification models.

Accuracy: Overall fraction of correctly predicted values

Confusion matrix: Shows the false positive rate, true positive rate, false negative rate and true negative rate.

Precision: Ration of true positives among the predicted positives

Recall: Ration of correctly predicted positive among all the actual positives

F1 score: Harmonic mean of precision and recall

Area under ROC curve: ROC curve is a plot between true positive and false positive rates for different thresholds of probability for +ve class. Area under this curve indicates how well a model can distinguish positive and negative classes

4.

i. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting?

A: When a model fails to capture the trends to a significant level, it is called underfitting. It happens when the model is too simple.

ii. What does it mean to overfit? When is it going to happen?

A: When a model fits too close to the data so that it loses its generalizability. It happens when the model is too complex, or when the data is not enough.

iii. In the sense of model fitting, explain the bias-variance trade-off.

A: It's not possible to completely eliminate both bias and variance. If the model is underfitting both bias and variance are high. If the model is overfitting, there will be little bias but the model will pick up errors and noises within data and increase variance.

5. Is it possible to boost the efficiency of a learning model? If so, please clarify how.

A: Yes it is possible to increase the efficiency of a model. Methods such as feature selection, cross validation can be used to do this.

6. How would you rate an unsupervised learning model's success? What are the most common success indicators for an unsupervised learning model?

A: Clustering is the most common form of unsupervised learning. For a clustering model, evaluation means measuring the similarity or dissimilarity. Silhouette coefficient and Dunn's index are two of metrics used for evaluating clustering models.

7. Is it possible to use a classification model for numerical data or a regression model for categorical data with a classification model? Explain your answer.

A: Depending on the use case, numerical data can be converted to categorical data but the reverse is not possible. Hence classification models can be applied on number data but regression models can't be used on categorical data.

8. Describe the predictive modeling method for numerical values. What distinguishes it from categorical predictive modeling?

A: Regression models can be used for predictive modeling for numeric values. On the other hand, it is difficult for the machine to understand the meaning of categorical values. However we can convert categorical values to numeric by encoding them or by creating dummy variables.

9. The following data were collected when using a classification model to predict the malignancy of a group of patients' tumors:

i. Accurate estimates – 15 cancerous, 75 benign

ii. Wrong predictions – 3 cancerous, 7 benign

Determine the model's error rate, Kappa value, sensitivity, precision, and F-measure.

Error rate = (3 + 7)/(15 + 75 + 3 + 7) = 10/100 = 0.10

p_o = 90/100 = 0.9

p_e = (15/100 * 3/100) + (75/100 * 7/100) = 0.057

Kappa = (0.9 - 0.057)/(1-0.057) = 0.943

Sensitivity(recall) = 15/(15 + 7) = 0.68

Precision = 15/(15+3) = 0.83

F-measure = 2(1/0.68 + 1/0.83) = 0.75

10. Make quick notes on:

1. The process of holding out

A: Splitting data into training, test and validation data to check the performance of data

2. Cross-validation by tenfold

A: Training data is split into 10 folds . In each of 10 iterations, one of these 10 folds is used as a validation set and the rest together as a training set.

3. Adjusting the parameters

A: Parameters can be used to increase the performance of a model. For example, threshold can be changed in classification models.

11. Define the following terms:

1. Purity vs. Silhouette width

A: Purity relates to overall number of points that are correctly labelled while silhouette width deals with a single point in terms of closeness to its current cluster.

2. Boosting vs. Bagging

A: Bagging involves training the model with random samples of data and can be done simultaneously while boosting involves creation of a new model based on the errors on the previous model(needs to be done sequentially)

3. The eager learner vs. the lazy learner

A: Lazy learners have more time to predict and less time to learn compared to eager learners.