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?

Answer: Target function is a complex mathematical function generated by supervised ML model to best the data to minimize the error to the least so that our prediction is accurate.

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.

Answer:

Predictive models is used for tasks that involve prediction one particular values from other features in the dataset .

Eg: Logistic regression, linear regression,SVM

Descriptive models are used for tasks that would benefit from the insight gained from data in new and interesting forms.

Eg: K-Means, hierarchical clustering.

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

Answering: After all the EDA, feature engineering and feature selection part, we feed our data to our respective ML model to check the performance. A performance of an ML model is based on how many correct predictions has it made .

Following are some of the performance metrics of measurement parameters:

1. Accuracy
2. Precision
3. Recall
4. F-score
5. Confusion matrix

4.

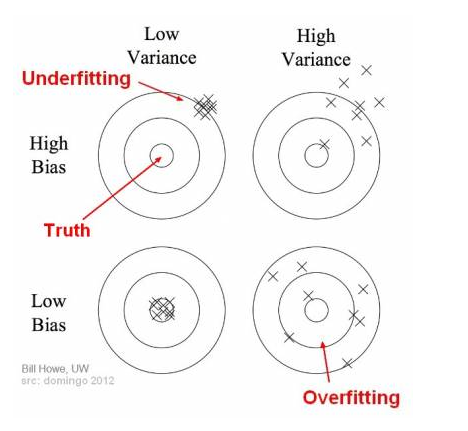
i. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting?

Answer: Underfitting is a condition in machine learning when your model performs really bad for training and testing data as well.

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

Answer: Overfitting is a situation where model performs really good on train data but performs really bad on test/unknown data. Overfitting happens when model learns the details and noise in the data to that much extent that it negatively affects the model performance.

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

Answer: 

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

Answer: Yes, there are several approaches to boost the efficiency of a ML model which are shown below:

1. Adding more data
2. Feature selection
3. Feature engineering
4. Missing values and outlier related treatment
5. Using ensemble models
6. Data distribution and parameter tuning

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

Answer: Following are the measures used for unsupervised learning model’s success and they are as below:

1. Rand index
2. Mutual information based scores
3. Homogeneity, completeness and V-measure
4. Fowlkes-Mallows scores
5. Silhoutte coefficient
6. Calinski Harabasz index
7. Davies-bouldin index
8. Contingency matrix

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.

Answer: Linear regression produces linear hypothesis function but in classification tasks , our data does not show up in a linear distribution but in a grouped distribution. This is because labelled data is a numerical data for regression problems and categorical data for classification problems. Therefore, there will be inconsistencies if we try to do something different.

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

Answer: Predictive modelling for numerical values is very similar to regression problem but it’s outcome is closely related to future forecasting, while is case of categorical variables, it is nothing but taking a new dataset and predicting the future category of new dataset.

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.

10. Make quick notes on:

1. The process of holding out

Answer: Holding out is simply when you split your dataset into a train and test dataset for model training purposes.

2. Cross-validation by tenfold

Answer: The dataset would be split into 10 equal sizes where 1 size will be used as test set randomly and rest will be used as training dataset and we will check the model performance via this method. Also, this same process will be repeated 10 times.

3. Adjusting the parameters

Answer: This is also known as hyperparameter tuning where we give every parameter some set of values and we try to select the best set of values for model parameters by running on each of the parameter’s values one by one.

11. Define the following terms:

1. Purity vs. Silhouette width

2. Boosting vs. Bagging

Answer: Bagging is a machine learning ensemble algorithm designed to reduce the overall variance by combining a set of base models , generally decision trees which helps in avoiding overfitting. It is a model averaging approach. Boosting is ensemble modelling technique that attempts to build a strong classifier from a list of weak classifiers. The process here is that every succeeding models tries to correct on the error that it’s previous model has made. The process is continued till the models are added or training dataset is correctly predicted.

3. The eager learner vs. the lazy learner

Answer: Lazy learners: stores data without learning from it and starts classifying data when model receives test data.

Eager learners: when it receives data, it starts classifying and does not wait for test data to learn accordingly.