1. In the sense of machine learning, what is a model? What is the best way to train a model?

A machine learning model is a file that has been trained to recognize certain types of patterns. We train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

We should clean data and pre-process them, analysis data which help to find the best features and models for our data.

2. In the sense of machine learning, explain the ‘No Free Lunch’ theorem.

It is a theoretical finding that suggests all optimization algorithms perform equally well when their performance is averaged over all possible objective functions.

3. Describe the K-fold cross-validation mechanism in detail.

In K fold we divide data into K parts and each time we select one of the parts as validation and other data for training the model and finally select the model and fold which shows the best performance. It is a method of optimization that verify impact of data on performance of model.

4. Describe the bootstrap sampling method. What is the aim of it?

Bootstrap is a technique of selecting randomly some of the value and start training based on those selected value and expand model, it use in decision tree and in start model select randomly some of the data as root and create the tree and calculate loss function and based on that select the best root for data. it help to reduce the cost of training and reaching to the result faster with less computation process

5. What is the significance of calculating the Kappa value for a classification model? Demonstrate how to measure the Kappa value of a classification model using a sample collection of results.

It tells how much better classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. kappa is always less than or equal to 1. And 0 or less, indicate that the classifier is not good at all.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 15 | 4 |
| 1 | 7 | 20 |

Accuracy = (15 + 20) / (15+4+7+20) = 35/46 = 0.78

Expected Acc = (((22 \* 19) / 46) + ((27\*24) /46)) / 46 = 0.50

Kappa = (accuracy – expected acc) / (1- expected acc) = (0.78 – 0.5) / (1-0.5) = 0.56

6. Describe the model ensemble method. In machine learning, what part does it play?

When we combine some models to improve the final accuracy it calls ensemble, there are various ensemble techniques such as bagging, boosting and stacking. Bagging is training some model parallelly and use max voting or averaging for getting final result such as random forest. Boosting means series, a week model to increase the final performance such as Ada boost and XGboost

7. What is a descriptive model’s main purpose? Give examples of real-world problems that

descriptive models were used to solve.

Descriptive Analytics will give you a vision of the past and tells you: what has happened? Whereas Predictive Analytics will recognize the future and tells you: What might happen in the future and a descriptive model will exploit the past data that are stored in databases and provide an accurate report. A Predictive model, identifies patterns found in past and transactional data to find risks and future outcomes. Most of the unsupervised model are descriptive such as K-means and as an example customer segmentation is one of them

8. Describe how to evaluate a linear regression model.

There different methods such as Mean Square Error, Mean absolute Error, Root of mean square error and R square and adjusted R square

9. Distinguish:

1. Descriptive vs. predictive models

Descriptive is unsupervised and data does not have target value while predictive have target value and model train based on them. descriptive analytics, which tell what has already happened while predictive analytics shows what could happen

2. Underfitting vs. overfitting the model

Underfitting means model is not enough accurate, and makes some wrong prediction it means the model cannot train properly. It happens when the error is high in both train and test data. Overfitting means the model makes not accurate predictions and it is so adjust with train data which is not flexible toward unseen data. in overfitting the error is less on train data and high in test data.

3. Bootstrapping vs. cross-validation

In bootstrapping we select data randomly ad it is possible some of the data never select and then the model miss those data during training which may impact on performance of the model but is cross-validation we divide data into specific partition and will run the model by selecting one partition as test data and other for train and repeat this process for all partition and finally select the best model, then performance will be more sure and all data will select

10. Make quick notes on:

1. LOOCV.

The Leave-One-Out Cross-Validation, or LOOCV, procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model

2. F-measurement

It is F-score and is the mean of precision and recall values. It can be calculated by the:

2 x [(Precision x Recall) / (Precision + Recall)] formula

3. The width of the silhouette

It is a method for calculating the performance of unsupervised model. It has a ranges between −1 and 1, where a higher silhouette coefficient refers to a model with more coherent clusters and silhouette coefficients close to +1 means the sample is far away from the neighbouring clusters.