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

Answer: A model represents what was learned by a machine learning algorithm. The model is the “thing” that is saved after running a machine learning algorithm on training data and represents the rules, numbers, and any other algorithm-specific data structures required to make predictions.

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model.

This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model.

Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved.

1. Define adequately our problem (objective, desired outputs…).
2. Gather data.
3. Choose a measure of success.
4. Set an evaluation protocol and the different protocols available.
5. Prepare the data (dealing with missing values, with categorial values…).
6. Correctly split the data.

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

Answer: The No Free Lunch Theorem, often abbreviated as NFL or NFLT, is a theoretical finding that suggests all optimization algorithms perform equally well when their performance is averaged over all possible objective functions. The theorem applies to optimization generally and to search problems, as optimization can be described or framed as a search problem. The implication is that the performance of your favourite algorithm is identical to a completely naive algorithm, such as random search.

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

Answer:

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

The general procedure is as follows:

* Shuffle the dataset randomly.
* Split the dataset into k groups.
* For each unique group:
  + Take the group as a hold out or test data set
  + Take the remaining groups as a training data set
  + Fit a model on the training set and evaluate it on the test set
  + Retain the evaluation score and discard the model
* Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

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

Answer: The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples.

Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. This approach to sampling is called sampling with replacement.

Importantly, any data preparation prior to fitting the model or tuning of the hyperparameter of the model must occur within the for-loop on the data sample. This is to avoid data leakage where knowledge of the test dataset is used to improve the model. This, in turn, can result in an optimistic estimate of the model skill.

A useful feature of the bootstrap method is that the resulting sample of estimations often forms a Gaussian distribution. In additional to summarizing this distribution with a central tendency, measures of variance can be given, such as standard deviation and standard error. Further, a confidence interval can be calculated and used to bound the presented estimate. This is useful when presenting the estimated skill of a machine learning model.

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.

Answer: The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. In addition, it takes into account random chance (agreement with a random classifier), which generally means it is less misleading than simply using accuracy as a metric (an Observed Accuracy of 80% is a lot less impressive with an Expected Accuracy of 75% versus an Expected Accuracy of 50%). Computation of Observed Accuracy and Expected Accuracy is integral to comprehension of the kappa statistic and is most easily illustrated through use of a confusion matrix.

Kappa is an important measure on classifier performance, especially on imbalanced data set.

For example, in credit card fraud detection, the marginal distribution of the response variable is high skewed, that using accuracy as a measure will not be useful. In other words, for given fraud detection example, 99.9% of the transactions will be non-fraud transactions. We can have a trivial classifier that always says non-fraud to every transaction, and we will still have 99.9% of the accuracy.

On the other hand, Kappa will "fix" this problem by consider the marginal distribution of the response variable. Using Kappa, the afore mentioned trivial classifier will have a very small Kappa.

In plain English, it measures how much better the classifier is, compared to guessing with the target distribution.

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

Answer: Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. That is why ensemble methods placed first in many prestigious machine learning competitions, such as the Netflix Competition, KDD 2009, and Kaggle.

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking).

Ensemble methods can be divided into two groups:

* Sequential ensemble methods where the base learners are generated sequentially (e.g. AdaBoost). The basic motivation of sequential methods is to exploit the dependence between the base learners. The overall performance can be boosted by weighing previously mislabeled examples with higher weight.
* Parallel ensemble methods where the base learners are generated in parallel (e.g. Random Forest). The basic motivation of parallel methods is to exploit independence between the base learners since the error can be reduced dramatically by averaging.

Most ensemble methods use a single base learning algorithm to produce homogeneous base learners, i.e. learners of the same type, leading to homogeneous ensembles.

7. What is a descriptive model's main purpose? Give examples of real-world problems that descriptive models were used to solve.

Answer: A descriptive model is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. As opposed to predictive models that predict a target of interest, in a descriptive model, no single feature is more important than any other. In fact, because there is no target to learn, the process of training a descriptive model is called unsupervised learning. Although it can be more difficult to think of applications for descriptive models, what good is a learner that isn't learning anything in particular - they are used quite regularly for data mining.

For example, the descriptive modeling task called pattern discovery is used to identify useful associations within data. Pattern discovery is often used for market basket analysis on retailers' transactional purchase data. Here, the goal is to identify items that are frequently purchased together, such that the learned information can be used to refine marketing tactics. For instance, if a retailer learns that swimming trunks are commonly purchased at the same time as sunglasses, the retailer might reposition the items more closely in the store or run a promotion to "up-sell" customers on associated items.

8. Describe how to evaluate a linear regression model.

Answer:

A good regression model is one where the difference between the actual or observed values and predicted values for the selected model is small and unbiased for train, validation, and test data sets.

To measure the performance of your regression model, some statistical metrics are used. Here we will discuss four of the most popular metrics. They are-

* Mean Absolute Error (MAE)
* Root Mean Square Error (RMSE)
* Coefficient of determination or R2
* Adjusted R2

Mean Absolute Error (MAE):

This is the simplest of all the metrics. It is measured by taking the average of the absolute difference between actual values and the predictions.

Root Mean Square Error (RMSE)

The Root Mean Square Error is measured by taking the square root of the average of the squared difference between the prediction and the actual value. It represents the sample standard deviation of the differences between predicted values and observed values (also called residuals).

Coefficient of Determination or R^2

It measures how well the actual outcomes are replicated by the regression line. It helps you to understand how well the independent variable adjusted with the variance in your model. That means how good is your model for a dataset.

Adjusted R-squared

There is a drawback of R^2 that it improves every time when we add new variables in the model.

Think about it, whenever you add a new variable there can be two circumstances, either the new variable improves your model or not. When the new variable improves your model then it is ok. But what if it does not improve your model? Then the problem occurs. The value of R^2 keeps on increasing with the addition of more independent variables even though they may not have a significant impact on the prediction.

9. Distinguish :

1. Descriptive vs. predictive models

2. Underfitting vs. overfitting the model

3. Bootstrapping vs. cross-validation

Answer:

1. A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report. In a Predictive model, it identifies patterns found in past and transactional data to find risks and future outcomes.
2. Overfitting is when the model's error on the training set (i.e. during training) is very low but then, the model's error on the test set (i.e. unseen samples) is large! Underfitting is when the model's error on both the training and test sets (i.e. during training and testing) is very high.
3. Cross validation splits the available dataset to create multiple datasets, and Bootstrapping method uses the original dataset to create multiple datasets after resampling with replacement. Bootstrapping it is not as strong as Cross validation when it is used for model validation. Bootstrapping is more about building ensemble models or just estimating parameters.

10. Make quick notes on:

1. LOOCV.

2. F-measurement

3. The width of the silhouette

4. Receiver operating characteristic curve

Answer:

1. 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-Measure provides a way to combine both precision and recall into a single measure that captures both properties. Alone, neither precision or recall tells the whole story. We can have excellent precision with terrible recall, or alternately, terrible precision with excellent recall. F-measure provides a way to express both concerns with a single score. Once precision and recall have been calculated for a binary or multiclass classification problem, the two scores can be combined into the calculation of the F-Measure.

The traditional F measure is calculated as follows:

F-Measure = (2 \* Precision \* Recall) / (Precision + Recall)

1. Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.
2. Image result for receiver operating characteristic curve machine learning. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.