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

In machine learning, a model is a file or an object that has been trained in data using algorithms to detect or “learn” certain patterns. This can then be used to predict future values for unseen input data.

The best way to train a model varies according to the algorithm being used and the task that needs to be performed. All the algorithms however, benefit from a large amount of good quality data to make better predictions. The more data is used for training, the better predictions can be expected of the model.

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

The “No Free Lunch” (NFL) theorem asserts that when the performance of all optimization algorithms is averaged across all conceivable problems, they all perform equally. This suggests that there is no single “best” or optimal algorithm present.

In the context of machine learning, all ML algorithms are in some form optimization problems, hence this expands to ML algorithms as well asserting that there is no single best ML algorithm for a particular kind of predictive modelling tasks like classification and regression.

This theory essentially means that for a particular ML task, there is no single algorithm that will perform best across all the possible tasks for that class of algorithms, rather it depends upon the specific problem and a trail and comparison approach has to be taken for finding the optimal algorithm for that particular problem.

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

K-fold Cross Validation (CV) is a resampling procedure used to evaluate the ML models on a limited dataset. Cross validation is used to evaluate the performance of the ML model in a generalized scenario to ensure it performs optimally on unseen data with minimal bias.

The k-fold CV uses the following general procedure:

1. Randomize the dataset
2. Split the dataset into k groups
3. Hold one of the groups as a test set
4. Use the remaining groups as training set
5. Fit the model on the training set and evaluate on the test set
6. Retain the evaluation score and discard the model
7. Create a different group with different train test split and repeat the process
8. Summarize the performance of the different group models

The scikit-learn library had the Kfold() method that can be used for implementing the method for model evaluation. It can generate the train test splits that can be used to create the final model.

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

Bootstrap sampling (or bootstrapping) is a concept in statistics where a sample data is repeatedly drawn from the data source to estimate the parameter for the whole data. For example, for calculation of average height of 1000 students instead if doing the tedious process of collecting the details of each of the students, we can take a sample size of 5 random students and take a mean of their heights. We repeat this process 20 times and take the average of the result of each of the samples. This would allow us to take a mean of 100 students representative of the result for the whole population.

In machine learning context, bootstrap sampling – also referred to as bootstrap aggregation or bagging technique – is used to create a cluster of models using small samples of the data and combining them. This helps mitigate the problem of overfitting and improves the overall stability of the 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.

The Cohen’s kappa value for a classification model can be used to evaluate its performance. It does this by comparing the true and model classifications on the test set. This is then used to determine how many of the model’s predictions match the real values which determines the kappa statistic.

There is no standardized way to interpret the values but an assumption of performance can be made using the range a value falls in. if the value is 0 or less then the model predicted and true values have no agreement, for 0 - 0.2 there is a slight agreement, 0.2 - 0.4 is fair, 0.4 – 0.6 is moderate, 0.6 – 0.8 is high and 0.8 – 1(max) is a perfect score.

For demonstration, consider the following case, for a test set of 1000 items with 500 positive and 500 negative samples, the model correctly gets 400 positive and 300 negative for a total of 700 correct classifications.

We can create a truth table comparing the predicted and true values from the test data.

|  |  |  |
| --- | --- | --- |
| Real/Model | Positive | Negative |
| Positive | 400 | 100 |
| Negative | 200 | 300 |

The observed agreement rate is 700/1000 = 0.70

expected\_agreement = (400+100) \* (400+400) / (400+100+400+400) + (400+400) \* (100+400) / (400+100+400+400) = 0.50

kappa = (0.70 - 0.50) / (1 - 0.50) = 0.4

Here, from the value of kappa, we can summarize that the model is moderately good at the given classification task.

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

Ensemble methods in machine learning essentially are used to combine multiple models to produce a better model. This can help in overcoming many of the issues encountered by single algorithms like underfitting, overfitting, suboptimal performance etc.

There are many ensemble techniques but three are widely used:

Bagging

Stacking

Boosting

Bagging – short for “Bootstrap Aggregation” – works by training a model on several data samples and combining the predicted output produced by the models by using a simple statistical methods like aggregation. This can help build a better model than using a single model to train on the whole data as this creates a more generalized and stable model

Stacking is another ensemble technique. Here, multiple models are trained on the dataset and another model is used to combine the outputs. The ensemble members (level 1 models) are used to predict an outcome and another model (level 0 model) is used to combine their outcomes to make the final prediction. This method is beneficial in creating several targeted models that are each better at certain predictions (overfitted) and the outcome of all the predictions can be fed to the level 0 model to give the final output

Boosting has come into popularity after the introduction of AdaBoost and subsequent algorithms that proved it to be a viable ensemble technique. This method uses the multiple models in a stepped form wherein the outcome of one model is “boosted” by the subsequent model. Here, the models are fit and sequentially added to the ensemble such that each subsequent model attempts to correct the prediction of the previous model to arrive at an optimal solution. The objective here is to develop a “strong learner” by combining the predictions of several purpose built “weak learners”

Combining different algorithms to produce the optimal solution can provide us with more flexibility in choice of the models and overcome the shortcomings of the individual models. Problems like bias can be used to an advantage by creating several biased models and combining their outcomes to produce a more robust generalized model, output of the individual models can be optimized and data can be used in a much more efficient manner for creating an optimized generalized model.

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

A descriptive model is used to gain insights on the data to generate a better understanding of it. Here there is not target variable or difference in the importance of the various features. The process of training a descriptive model is called unsupervised learning.

Descriptive models are widely used in the field of data mining where pattern discovery on unlabelled data is required. This is often used in for market basket analysis for retail transactions. We can identify the items that buyers typically buy together and this information can be used for store organization or promotional offer bundles to upsell a customer to buy more items. A good real world example of this is the “frequently bought together” recommendations on amazon and similar e-commerce sites which can encourage a customer to purchase more by suggesting items of interest.

9. Distinguish:

1. Descriptive vs. predictive models

Descriptive models are used on unlabelled data to gain insights using pattern recognition which can help understand it better. This is used for unsupervised learning tasks where there are no target variables to predict.

Predictive models on the other hand are used for predicting future values based on historical understanding of the data. These are more widely used where a target variable is to be predicted using various features associated with it

2. Underfitting vs. overfitting the model

Underfitting of a model is when the model is unable to properly identify the trends in the data which causes it to not be as effective in predicting output based on unseen data. This can be caused by suboptimal quality of data, wrong choice of algorithm, too low model complexity etc.

Overfitting is quite the opposite but produces a similar end result of poor model performance on unseen data. This happens when a model is able to predict the training data values with high accuracy but this behaviour is not reflected on unseen data. This can happen due to the model being too complex, not enough diversity in the used training data or just not enough training data.

3. Bootstrapping vs. cross-validation

Bootstrapping is an ensemble technique which is used to improve model performance where different models are used for prediction and an aggregate of the predictions is used to generate the final output.

Cross validation is quite different as it is used to produce a more generalized model by training the model on different distributions of the data by splitting the training data into several parts. One part is kept for validation and others are used for training the model, this process is then repeated several times keeping a different part for testing each time. This helps in reducing the problem of overfitting allowing the model to perform better on unseen data.

10. Make quick notes on:

1. LOOCV.

LOOCV stands for Leave One Out Cross Validation. This is a cross validation technique that involves holding out a single observation from a dataset as validation data and training a model on rest of the data. This is an extreme version of the K-Fold CV where each observation gets a chance to be a part of the training dataset. This method is suitable for only small datasets for improving the accuracy as it is computationally expensive.

2. F-measurement

In binary classification, F-measure or F-score is the measure of a test’s accuracy. In case of machine learning, it gives the accuracy of the model. The F-measure is calculated as the harmonic mean of the precision and recall values of the model. Highest possible value for F-score is 1 indicating perfect precision and recall and lowest is 0 in case of 0 precision or recall value.

3. The width of the silhouette

Silhouette is an interpretation and validation method for consistency within the clusters of data. It provides a graphical representation of how well an object has been classified into the right cluster. The value of a silhouette width ranges from +1 to -1 and where higher value indicates better matching of an object to objects of its own cluster and poorer with the others. If many points have a low value, it may indicate that the distribution may have too many or too few clusters

4. Receiver operating characteristic curve

ROC curve gives a measure of performance of a classification algorithm at various thresholds. The ROC is a probability curve and represents the capability of the algorithm to correctly identify the object cluster. It is plotted as TPR vs FPR graph where TPR is the true positive rate and FPR is the false positive rate which is calculated as 1 – [Specificity].

The area under the ROC curve, called AUC gives the measure of separability. An AUC near 1 represents a good measure of separability and 0 represents a poor measure.

For a multiclass problem, an ROC-AUC curve for each of the classes can be plotted against the remaining classes. This means that for a problem with three possible classes, three separate ROC graphs can be plotted.