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. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

Step 1: Begin with existing data

Machine learning requires us to have existing data—not the data our application will use when we run it, but data to learn from. You need a lot of real data, in fact, the more the better. The more examples you provide, the better the computer should be able to learn. So just collect every scrap of data you have and dump it and voila! Right?

Wrong. In order to train the computer to understand what we want and what we don’t want, you need to prepare, clean and label your data. Get rid of garbage entries, missing pieces of information, anything that’s ambiguous or confusing. Filter your dataset down to only the information you’re interested in right now. Without high quality data, machine learning does not work. So take your time and pay attention to detail.

Step 2: Analyze data to identify patterns

Unlike conventional software development where humans are responsible for interpreting large data sets, with machine learning, you apply a machine learning algorithm to the data. But don’t think you’re off the hook. Choosing the right algorithm, applying it, configuring it and testing it is where the human element comes back in.

There are several platforms to choose from both commercial and open source. Explore solutions from Microsoft, Google, Amazon, IBM or open source frameworks like TensorFlow, Torch and Caffe. They each have their own strengths and downsides, and each will interpret the same dataset a different way. Some are faster to train. Some are more configurable. Some allow for more visibility into the decision process. In order to make the right choice, you need to experiment with a few algorithms and test until you find the one that gives you the results most aligned to what you’re trying to achieve with your data.

When it’s all said and done, and you’ve successfully applied a machine learning algorithm to analyze your data and learn from it, you have a trained model.

Step 3: Make predictions

There is so much you can do with your newly trained model. You could import it into a software application you’re building, deploy it into a web back end or upload and host it into a cloud service. Your trained model is now ready to take in new data and feed you predictions, aka results.

These results can look different depending on what kind of algorithm you go with. If you need to know what something is, go with a classification algorithm, which comes in two types. Binary classification categorizes data between two categories. Multi-class classification sorts data between—you guessed it—multiple categories.

When the result you’re looking for is an actual number, you’ll want to use a regression algorithm. Regression takes a lot of different data with different weights of importance and analyzes it with historical data to objectively provide an end result.

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

The “No Free Lunch” Theorem argues that, without having substantive information about the modeling problem, there is no single model that will always do better than any other model. Because of this, a strong case can be made to try a wide variety of techniques, then determine which model to focus on.

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

k-Fold cross-validation is a technique that minimizes the disadvantages of the hold-out method. k-Fold introduces a new way of splitting the dataset which helps to overcome the “test only once bottleneck”.

The algorithm of the k-Fold technique:

* Pick a number of folds – k. Usually, k is 5 or 10 but you can choose any number which is less than the dataset’s length.
* Split the dataset into k equal (if possible) parts (they are called folds)
* Choose k – 1 folds as the training set. The remaining fold will be the test set
* Train the model on the training set. On each iteration of cross-validation, you must train a new model independently of the model trained on the previous iteration
* Validate on the test set
* Save the result of the validation
* Repeat steps 3 – 6 k times. Each time use the remaining fold as the test set. In the end, you should have validated the model on every fold that you have.
* To get the final score average the results that you got on step 6.

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

The bootstrap method divides the data set with N cases into B samples of identical size with replacement. A separate model of some target variable is built on each of the samples, yielding an n-number of predictions for each record in the data set. The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. It can be used to estimate summary statistics such as the mean or standard deviation

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.

Kappa can range from 0 to 1. A value of 0 means that there is no agreement between the raters (real-world observer vs classification model), and a value of 1 means that there is perfect agreement between the raters. In most cases, anything over 0.7 is considered to be very good agreement. To calculate the Kappa coefficient we will take the probability of agreement minus the probability of disagreement divided by 1 minus the probability of disagreement. This is a positive value which means there is some mutual agreement between the parties.

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

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

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 describes a system or other entity and its relationship to its environment. It is generally used to help specify and/or understand what the system is, what it does, and how it does it. A geometric model or spatial model is a descriptive model that represents geometric and/or spatial relationships. Examples of descriptive data mining include clustering, association rule mining, and anomaly detection

8. Describe how to evaluate a linear regression model.

Mean Squared Error (MSE)

The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.

Mean Absolute Error (MAE)

This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent.

R-squared or Coefficient of Determination

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.

9. Distinguish :

1. Descriptive vs. predictive models

|  |  |
| --- | --- |
| **Descriptive Data Mining** | **Predictive Data Mining** |
| Descriptive mining is generally used to support correlation, cross-tabulation, frequency, etc. | The term 'Predictive' defines to predict something, so predictive data mining is the analysis done to predict the future event or multiple data or trends. |
| It defines the features of the data in a target data set. | It executes the induction over the current and past records so that predictions can appear. |
| It requires data aggregation and data mining. | It requires statistics and data forecasting procedures. |
| The descriptive analysis only responds to the situation. | The predictive analysis includes control over the situation along with responding to it. |
| It can support accurate records. | It makes results does not provide accuracy. |

2. Underfitting vs. overfitting the model

Underfitting: A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data. (It’s just like trying to fit undersized pants!) Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data. In such cases, the rules of the machine learning model are too easy and flexible to be applied to such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.

In a nutshell, Underfitting refers to a model that can neither performs well on the training data nor generalize to new data.

Reasons for Underfitting:

High bias and low variance

The size of the training dataset used is not enough.

The model is too simple.

Training data is not cleaned and also contains noise in it.

Techniques to reduce underfitting:

Increase model complexity

Increase the number of features, performing feature engineering

Remove noise from the data.

Increase the number of epochs or increase the duration of training to get better results.

Overfitting: A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

In a nutshell, Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

Reasons for Overfitting are as follows:

High variance and low bias

The model is too complex

The size of the training data

3. Bootstrapping vs. cross-validation

Bootstrapping is a technique that helps in many situations like validation of a predictive model performance, ensemble methods, estimation of bias and variance of the model.

It works by sampling with replacement from the original data, and take the “not chosen” data points as test cases. We can make this several times and calculate the average score as estimation of our model performance.

In addition, Bootstrapping helps in ensemble methods as we may build a model (like a Decision tree) using each bootstrap data set and “bag” these models in an ensemble (like Random Forest) and take the majority voting for all of these models as our resulting classification.

On the other hand, cross validation is a technique for validating the model performance, and it’s done by split the training data into k parts. We take k-1 parts as our training set and use the “held out” part as our test set. We repeat that k times differently (we hold out different part every time). Finally we take the average of the k scores as our performance estimation.

Cross validation can suffer bias or variance. if we increase the number of splits (k), the variance will increase and bias will decrease. On contrast, if we decrease (k), the bias will increase and variance will decrease. Generally 10-fold CV is used but of course it depends on the size of the training data.

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. It is a computationally expensive procedure to perform, although it results in a reliable and unbiased estimate of model performance. Although simple to use and no configuration to specify, there are times when the procedure should not be used, such as when you have a very large dataset or a computationally expensive model to evaluate.

2. F-measurement

The F-measure of the system is defined as the weighted harmonic mean of its precision and recall, that is, F = {1\over \alpha {1\over P}+(1-\alpha ) {1\over R}}, where the weight α ∈ [0,1]. The balanced F-measure, commonly denoted as F 1 or just F, equally weighs precision and recall, which means α = 1∕2.

3. The width of the silhouette

Silhouette width is a widely used index for assessing the fit of individual objects in the classification, as well as the quality of clusters and the entire classification. The Average Silhouette Width (ASW) of a clustering is. a ( i ) is the average distance of to points in the cluster to which it was assigned, and is the average distance of to the points in the nearest cluster to which it was not assigned.

4. Receiver operating characteristic curve

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. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.