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

**Ans**: A machine learning model is defined as a**mathematical representation of the output of the training process**. Machine learning is the study of different algorithms that can improve automatically through experience & old data and build the model.

A machine learning model is similar to computer software designed to recognize patterns or behaviours based on previous experience or data. The learning algorithm discovers patterns within the training data, and it outputs an ML model which captures these patterns and makes predictions on new data.

**Machine learning models :**

* **Classification Models**
* **Clustering**
* **Regression Models**
* **Dimensionality Reduction**
* **Deep learning**

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

## Ans: *There ain’t no such thing as a free lunch!*

If you are familiar with the world of machine learning in general, you have probably heard about the well-known “No free lunch” theorem. My goal with this article is to present this theorem in the simplest possible way and emphasize why understanding the implications are important when formulating an AIOPS strategy.

The “No Free Lunch” theorem for ML explains that there is no single machine learning algorithm that will perform best for all problems. So there is not just one optimized algorithm for solving all problems or even a given problem in different situations.

There are generally two No Free Lunch theorems, one for **Supervised Machine Learning (Wolpert 1996)**, and the other for **search and optimization (Wolpert and Macready 1997).**

In this article, we are going to focus more on the one written by David Wolpert for supervised machine learning.

The beginnings of this theorem dates back to 1996, where through a paper called *The lack of a Priori Distinctions between Learning Algorithms*, David Wolpert studies the possibilities to get useful theoretical results with a training data set and a learning algorithm with no assumptions about the target variable. Through various mathematical models, Wolpert demonstrates that for every two algorithms A and B, there are as many scenarios that A will perform worse than B as there are scenarios that A will perform better than B. This is true even if one of the provided algorithms is guessed randomly. Wolpert developed a mathematical proof demonstrating that for all possible problem instances drawn from a uniform probability distribution, the average performance of A and B algorithms is the same.

Since machine learning algorithms are built differently for different issues, it is clear that the assumptions generated by these algorithms will not fit all the data sets perfectly. By definition, it also implies that there will be as many data sets that a given algorithm will not be able to model effectively.

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

### Ans: **Definition :**

K-fold cross-validation is a data partitioning technique which splits an entire dataset into k groups. Then, we train and test k different models using different combinations of the groups of data we just partitioned, and use the results from these k models to check the model’s overall performance and generality.

In the context of machine learning, a

[fold](https://qr.ae/pGjnkX)

is a set of rows in a dataset. We will use k-folds to describe a number of groups we decide to partition the data, so in an example of 20 rows, we can split them into 2 folds with 10 rows each, 4 folds with 5 rows each, or 10 folds with 2 rows each.

A simple explanation of how

[k-fold cross validation](https://towardsdatascience.com/k-fold-cross-validation-explained-in-plain-english-659e33c0bc0)

scores a model’s performance is:

* The entire dataset is randomly split into equally-sized, independent

k-folds

, without reusing any of the rows in another fold.

* We use

k-1

folds

for model training, and once that model is complete, we test it using the remaining 1 fold to obtain a score of the model’s performance.

* We repeat this process

k times

, so we have

k number

of models and scores for each.

* Lastly, we take the mean of the

k number

of scores to evaluate the model’s performance.

### Conceptual example

To improve your understanding

twice-fold

😏, consider this analogy about k-fold cross validation with

Twice, a K-pop girl group. Say we are trying to see how well a

Model can dance by inviting different subsets of Twice girls (calledfolds) as training and test samples.

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

**Ans:** The idea behind bootstrap is to use the data of a sample study at hand as a surrogate population, for the purpose of**approximating the sampling distribution of a statistic**; that is, to resample (with replacement) from the sample data at hand and create a large number of phantom samples known as bootstrap samples.

**Boot Strap Method :**

To apply the bootstrap method, the two samples are combined together and two independent bootstrap samples of sizes equal to the original samples are generated from the combined sample.

**Aim of Bootstrap Method :**

Bootstrap’s primary objective is**to create responsive, mobile-first websites**. It ensures all interface elements of a website work optimally on all screen sizes. Bootstrap is available in two variants ‒ precompiled and based on a source code version.

**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.**

**Ans :** The**reliability of clinicians' ratings is an important consideration** in areas such as diagnosis and the interpretation of examination findings. Often, these ratings lie on a nominal or an ordinal scale. For such data, the kappa coefficient is an appropriate measure of reliability.

It basically tells you how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. Cohen’s **kappa** is always less than or equal to 1. **Values** of 0 or less, indicate that the classifier is useless.

**Cohen’s Kappa statistic** is a very useful, but under-utilised, metric. Sometimes in machine learning we are faced with a [multi-class classification](https://en.wikipedia.org/wiki/Multiclass_classification) problem. In those cases, measures such as the accuracy, or precision/recall do not provide the complete picture of the performance of our classifier.

In some other cases we might face a problem with imbalanced classes. E.g. we have two classes, say A and B, and A shows up on 5% of the time. Accuracy can be misleading, so we go for measures such as precision and recall. There are ways to combine the two, such as the[F-measure](https://en.wikipedia.org/wiki/F1_score), but the F-measure does not have a very good intuitive explanation, other than it being the harmonic mean of precision and recall.

[Cohen’s kappa](https://en.wikipedia.org/wiki/Cohen%27s_kappa) statistic is a very good measure that can handle very well both multi-class and imbalanced class problems.

**Cohen’s kappa** is defined as:

cohen's kappa statistic

where po is the observed agreement, and pe is the expected agreement. It basically tells you how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class.

When the number of codes is less than five, and especially when K = 2, lower values of Kappa are acceptable, but prevalence variability also needs to be considered. For only two codes, the highest kappa value is .80 from observers with accuracy .95, and the lowest is kappa value is .02 from observers with accuracy .80.

# Interpretation of Kappa Values

## Evaluate the agreement level with condition

The kappa statistic is frequently used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured. Measurement of the extent to which data collectors (raters) assign the same score to the same variable is called interrater reliability. In 1960, Jacob Cohen critiqued the use of percent agreement due to its inability to account for chance agreement. He introduced the Cohen’s kappa, developed to account for the possibility that raters actually guess on at least some variables due to uncertainty. The scale of Kappa value interpretation is the as following:

*Kappa value interpretation Landis & Koch (1977):  
<0 No agreement  
0 — .20 Slight  
.21 — .40 Fair  
.41 — .60 Moderate  
.61 — .80 Substantial  
.81–1.0 Perfect*

However past researches indicated that multiple factors have influences on Kappa value: observer accuracy, # of code in the set, the prevalence of specific codes, observer bias, observer independence (Bakeman & Quera, 2011). As a result, interpretations of Kappa, including definitions of what constitutes a good kappa should take circumstances into account.

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

**Ans:** Ensemble method in Machine Learning is defined as the**multimodal system in which different classifier and techniques are strategically combined into a predictive model** (grouped as Sequential Model, Parallel Model, Homogeneous and Heterogeneous methods etc.) Ensemble method also helps to reduce the variance in the predicted data, minimize the biasness in the predictive model and to classify and predict the statistics from the complex problems with better accuracy.

#### 1. Sequential Methods

In this kind of Ensemble method, there are sequentially generated base learners in which data dependency resides. Every other data in the base learner is having some dependency on previous data. So, the previous mislabeled data are tuned based on its weight to get the performance of the overall system improved.

**Example**: Boosting

#### 2. Parallel Method

In this kind of Ensemble method,  the base learner is generated in parallel order in which data dependency is not there. Every data in the base learner is generated independently.

**Example**: Stacking

#### 3. Homogeneous Ensemble

Such an ensemble method is a combination of the same types of classifiers. But the dataset is different for each classifier. This will make the combined model work more precisely after the aggregation of results from each model. This type of ensemble method works with a large number of datasets. In the homogeneous method, the feature selection method is the same for different training data. It is computationally expensive.

**Example:** Popular methods like bagging and boosting comes into the homogeneous ensemble.

#### 4. Heterogeneous Ensemble

Such an ensemble method is the combination of different types of classifiers or [machine learning models](https://www.educba.com/machine-learning-models/) in which each classifier built upon the same data. Such a method works for small datasets. In heterogeneous, the feature selection method is different for the same training data. The overall result of this ensemble method is carried out by averaging all the results of each combined model.

**Example**: Stacking

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

**Ans:** A descriptive model is usually an equation chosen to**fit experimental or observational data**. For example, Kepler’s law concerning the period of a planet’s motion was obtained by fitting to observational data recorded by the astronomer Tycho Brahe.

## [Goals, Power, and Sample Size](https://www.sciencedirect.com/science/article/pii/B9780124058880000131)

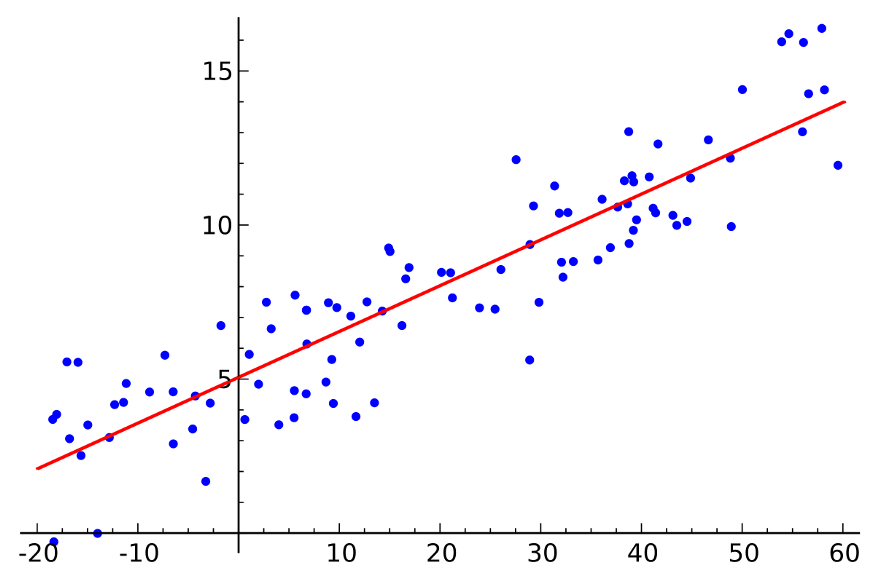
## Power analysis requires verisimilitude of simulated data

Power analysis is only useful when the simulated data imitate actual data. We generate simulated data from a descriptive model that has uncertainty in its parameter values, but we assume that the model is a reasonably good description of the actual data. If the model is instead a poor description of the actual data, then the simulated data do not imitate actual data, and inferences from the simulated data are not very meaningful. It is advisable, therefore, to check that the simulated data accurately reflect the actual data.

When simulated data differ from actual data, strange results can arise in power analysis. Consider an analysis of replication probability in which the simulated data are quite different to the actual data. The novel simulated data are combined with the original data to conduct the replication analysis. The combined data are a mixture of two different trends (i.e., the actual trend and the different simulated trend), and therefore the estimates of the parameters become more uncertain than for the original data alone. It is only when the simulated sample size becomes large, relative to the original sample size, that the simulated trend overwhelms the actual trend, and the replication uncertainty becomes smaller again. If you find in your analyses of replication power that parameter uncertainty initially gets larger as the simulated sample size increases, then you may have a situation in which the model does not faithfully mimic the actual data.

**8. Describe how to evaluate a linear regression model.**

# **Ans:** Evaluation of Linear Regression Models



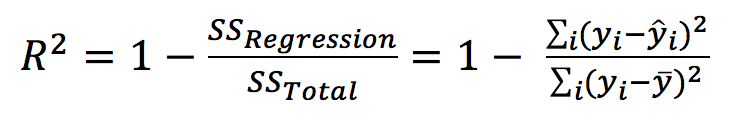
Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted is called the dependent variable and the factors that are used to predict the value of the dependent variable are called independent variables.

Evaluating a machine learning model is as important as building it. We are creating models to perform on new and unseen data. Hence, we need to evaluate if our model is performing correctly. Evaluating a Linear Regression model is not easy because there are a lot of evaluation metrics. When to use which metric depends on the data and problem of the project.

In this post, I will go over some evaluation metrics for Regression models.

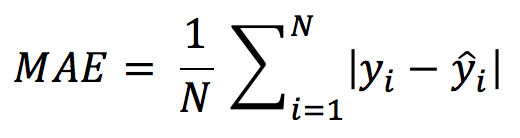
**R Squared(R²)**

R-squared is a goodness of fit measure for linear regression models. This indicates the percentage of the variance in the dependent that the independent variables explain collectively. R-squared measures the strength of the relationship between the model and the dependent variable. R Squared value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value. Here is the formula for R-squared and the calculation of R² with sci-kit Learn is the following:



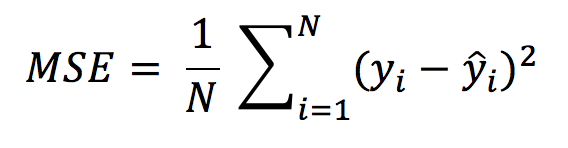
**Mean Absolute Error(MAE)**

Mean Absolute Error is a measure of errors between observations and predictions. It is the average magnitude of the errors in a set of predictions, without considering their directions. It is the absolute value of error between actual and predicted value. Following is the formula and way to calculate with sci-kit learn.



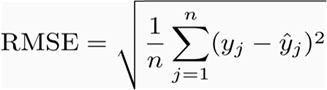
**Mean Squared Error(MSE)**

Mean Squared Error is the sum of the square of prediction error. Mean Squared Error is similar to Mean Absolute Error. Mean Absolute Error takes the absolute value of error but Mean Squared Error takes the square of error. MSE penalize big prediction error by square while MAE treats all the errors the same.



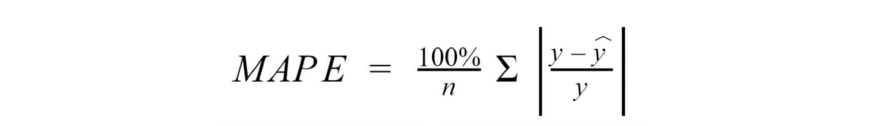
**Root Mean Squared Error(RMSE)**

Root Mean Squared Error is the square root of the mean squared error. RMSE is always non-negative and a value of 0 would indicate a perfect fit to the data. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. Following is the formula of RMSE and how to calculate RMSE in python.



**Mean Absolute Percentage Error(MAPE)**

Mean Absolute Percentage Error measures the accuracy as a percentage and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.



**9. Distinguish :**

**1. Descriptive vs. predictive models**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Comparison | Descriptive Data | |  |  | | --- | --- | |  |  |   Predictive Data |
| **1** | **Basic** | It determines, what happened in the past by analyzing stored data. | It determines, what happened in the past by analyzing stored data. |
| **2** | **Preciseness** | It provides accurate data. | It produces results does not ensure accuracy.3 |
| **3** | Practical analysis methods | Standard reporting, query/drill down and ad-hoc reporting. | Predictive modelling, forecasting, simulation and alerts. |
| **4** | **Require** | It requires data aggregation and data mining | It requires statistics and forecasting methods |
| **5** | **Type of approach** | Reactive approach | Proactive approach |
| **6** | **Describe** | Describes the characteristics of the data in a target data set. | Carry out the induction over the current and past data so that predictions can be made. |

**2. Underfitting vs. overfitting the model**

Overfitting and underfitting are commonplace issues that you are sure to encounter during your machine learning or deep learning training. It’s important to understand what these terms mean in order to spot them when they arise. Building a good model takes time and effort which includes dealing with issues like these and performing balancing acts as you optimize your project. This also involves lots of study and practice to improve your skillset. Ready to dive deeper into both theory and practice and learn how to build well-trained models? Try our Deep Learning Course with TensorFlow 2 for free.

**3. Bootstrapping vs. cross-validatio**

### Boostraping

Bootstrapping is a resampling technique with replacement; that is, we can choose on every sample a subset of elements that might be repeated.

#### Boostraping with Scikit-Learn

In Scikit-Learn, we can get a random sample of a dataset D with the resample method. Let's select from a 4-row matrix 3 samples of two rows each.

### Cross-Validation

Cross-validations is a very similar technique to Bootstrapping, with the difference that it selects its samples without replacement; that is, there are no repeated elements in every subset. Selecting k-samples with Cross-Validation is called K-Fold CrossValidation. Usually, on each cross-validation exercise, we define a portion of the selected sample to be the train and the other the test set (for example 70/30 or 80/20). The type of cross-validation that selects a test set with one single example is called LOOCV (leave one out cross-validation).

**10. Make quick notes on:**

**1. LOOCV (Leave One Out Cross-Validation)**

**LOOCV(Leave One Out Cross-Validation)** is a type of [cross-validation](https://www.geeksforgeeks.org/cross-validation-in-r-programming/) approach in which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set. In LOOCV, fitting of the model is done and predicting using one observation validation set. Furthermore, repeating this for N times for each observation as the validation set. Model is fitted and the model is used to predict a value for observation. This is a special case of **K-fold cross-validation** in which the number of folds is the same as the number of observations(K = N). This method helps to reduce **Bias**and **Randomness.**The method aims at reducing the Mean-Squared error rate and prevent over fitting. It is very much easy to perform LOOCV in [R programming](https://www.geeksforgeeks.org/introduction-to-r-programming-language/).

**2. F-measurement**

In [statistical](https://en.wikipedia.org/wiki/Statistics) analysis of [binary classification](https://en.wikipedia.org/wiki/Binary_classification" \o "Binary classification), the **F-score** or **F-measure** is a measure of a test's [accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision#In_binary_classification). It is calculated from the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) and [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. Precision is also known as [positive predictive value](https://en.wikipedia.org/wiki/Positive_predictive_value), and recall is also known as [sensitivity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity) in diagnostic binary classification.

The **F1** score is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) of the precision and recall. The more generic {\displaystyle F\_{\beta }} score applies additional weights, valuing one of precision or recall more than the other.

The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if both precision and recall are zero.

**3. The width of the silhouette**

Image data are normally unstructured and high dimensional due to the photography technology advancement such that an image can be taken at a wide range of resolution levels. To overcome such problem, data miners may consider selecting only a minimal set of features that are really important for classifying their images. Feature selection is a popular method for reducing dimensions in data. However, most feature selection algorithms return results in form of score for each feature. It is still difficult for data miners to choose features based on such scoring scheme because they may not know which score range is the best for their data classification at hand. Therefore, in this research, we aim to assist data miners and novice data analysts on solving dimensionality problem by finding for them the best optimal set of features, instead of just reporting the scores of all features and leaving the selection step to be the burden of miners. We select optimal set of features by firstly apply clustering technique to group similar features based on their scores. We thus propose the silhouette width criterion for selecting the optimal number of clusters during the cluster analysis step. After that we perform association mining to analyze relationships that may exist among different subsets of features toward the target attribute. Our method finally reports user the best subset of features to be potentially used further for data classification. We demonstrate performance of our proposed method on the satellite forest image data in Japan. © 2018 International Association of Computer Science and Information Technology.

**4. Receiver operating characteristic curve(ROC)**

A Receiver Operating Characteristics (ROC) Curve is used to describe the trade-off between correct classifications and wrong classifications.

The ROC curve displays a plot of the True Positive (TP) against the False Positive (FP).

The performance of a classifier is represented as a point in the curve.

The total performance of a classifier is summarized over all possible threshold in the curve. The overall performance is given by area under the curve (AUC).

A high-performing model will have an ROC that will pass close to the upper left side of the curve and provide a large area under