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

Ans: Machine learning models are computer programs that are used to recognize patterns in data or make predictions. Machine learning models are created from [machine learning algorithms](https://www.coursera.org/articles/machine-learning-algorithms), which are trained using either labeled, unlabeled, or mixed data. Different machine learning algorithms are suited to different goals, such as classification or prediction modeling.As data is introduced to a specific algorithm, it is modified to better manage a specific task and becomes a machine learning model.

* Supervised learning: Supervised learning occurs when an algorithm is trained using “labeled data”, or data that is tagged with a label so that an algorithm can successfully learn from it. Training an algorithm with labeled data helps the eventual machine learning model know how to classify data in the manner that the researcher desires.
* Unsupervised learning: Unsupervised learning uses unlabeled data to train an algorithm. In this process, the algorithm finds patterns in the data itself and creates its own data clusters. Unsupervised learning is helpful for researchers who are looking to find patterns in data that are currently unknown to them.
* Semi-supervised learning: Semi-supervised learning uses a mix of labeled and unlabeled data to train an algorithm. In this process, the algorithm is first trained with a small amount of labeled data before being trained with a much larger amount of unlabeled data.

The best way to train the model:

1. First determine problem statement, access data set and clean the data to be presented to the model. try to find out correlations among independent variable. Determine which algorithms and what parameters will use.
2. Split dataset into a training set and a testing set, Use K fold cross-validation. For e.g  [10-fold cross-validation](https://www.dominodatalab.com/blog/guide-to-building-models-with-cross-validation), the data is split into ten sets, allowing to train and test the data ten times.
   1. Split the data into ten equal parts or folds.
   2. Designate onefold as the hold-out fold.
   3. Train the model on the other nine folds.
   4. Test the model on the hold-out fold.
3. Select algorithm: Selecting candidate algorithms will often depend on:

* Size of the training data.
* Accuracy and interpretability of the required output.
* Speed of training time required, which is inversely proportional to accuracy.
* Linearity of the training data.
* Number of features in the data set.

1. Tune the Hyperparameters: Evaluate optimal values for input parameters ( **Hyperparameters)** which would generate the best model output.Comapre the model performance after hyperparameter tuning.

and select the best model based on model evaluation score.

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

Ans: The No Free Lunch Theorem is often used in optimization and machine learning, with little comprehension of what it means or implies. The theory asserts that when the performance of all optimization methods is averaged across all conceivable problems, they all perform equally well. It indicates that no one optimum optimization algorithm exists. Because of the strong link between optimization, search, and machine learning, there is no one optimum machine learning method for predictive modelling tasks like classification and regression.They all agree on one point: there is no “best” algorithm for specific kinds of algorithms, since they all perform similarly on average. Mathematically, the computing cost of finding a solution is the same for any solution technique when averaged across all problems in the class. As a result, no solution provides a shortcut.

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

Ans: 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. For k=10 , it is 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:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. 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
4. Summarize the skill of the model using the sample of model evaluation scores

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

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

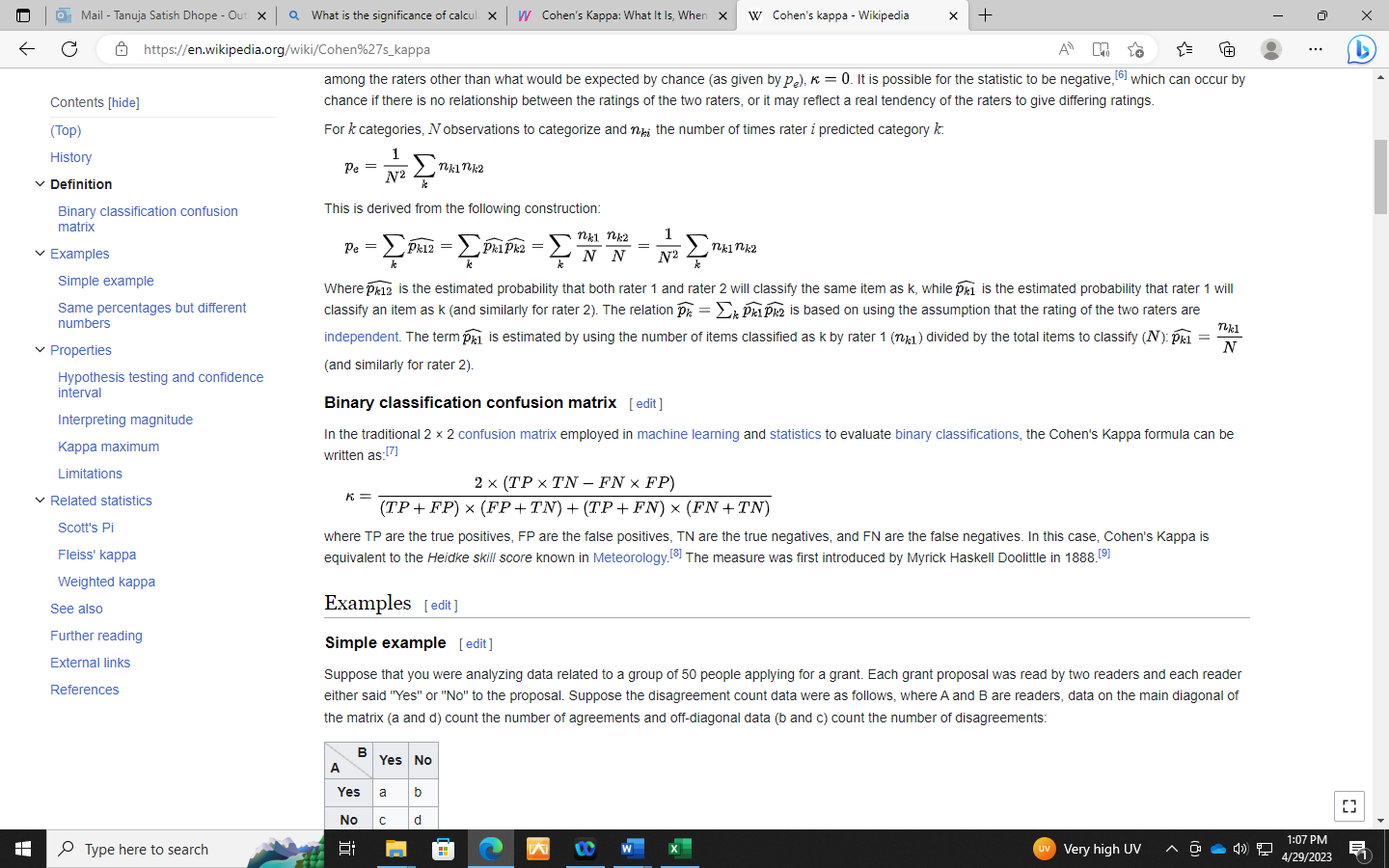
1. Choose a number of bootstrap samples to perform
2. Choose a sample size
3. For each bootstrap sample
   1. Draw a sample with replacement with the chosen size
   2. Calculate the statistic on the sample
4. Calculate the mean of the calculated sample statistics.

The procedure can also be used to estimate the 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.

Ans: The Kappa Statistic or Cohen’s\* Kappa is a statistical measure of inter-rater [reliability](https://www.theanalysisfactor.com/what-is-reliability/) for categorical variables in ML. It is usually described is **the amount of agreement correct by the agreement expected by chance**. However, it technically isn't corrected by chance but instead reports if the agreement is greater than by chance. It is most generally applied to predictive models built from unbalanced data (i.e. class distributions not equivalent). It is generally thought to be a more robust measure than simple percent agreement calculation, as *κ* takes into account the possibility of the agreement occurring by chance.

In the traditional 2 × 2 [confusion matrix](https://en.wikipedia.org/wiki/Confusion_matrix) employed in [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [statistics](https://en.wikipedia.org/wiki/Statistics) to evaluate [binary classifications](https://en.wikipedia.org/wiki/Binary_classification), the Cohen's Kappa formula can be written as:[[7]](https://en.wikipedia.org/wiki/Cohen%27s_kappa#cite_note-7)

�=2×(��×��−��×��)(��+��)×(��+��)+(��+��)×(��+��) 

where TP are the true positives, FP are the false positives, TN are the true negatives, and FN are the false negatives. In this case, Cohen's Kappa is equivalent to the *Heidke skill score* known in [Meteorology](https://en.wikipedia.org/wiki/Meteorology).

For instance, in the following two cases there is equal agreement between A and B (60 out of 100 in both cases) in terms of agreement in each class, so we would expect the relative values of Cohen's Kappa to reflect this. However, calculating Cohen's Kappa for each:

|  |  |  |
| --- | --- | --- |
| B  A | Yes | No |
| Yes | 45 | 15 |
| No | 25 | 15 |

�=0.60−0.541−0.54=0.1304

|  |  |  |
| --- | --- | --- |
| B  A | Yes | No |
| Yes | 25 | 35 |
| No | 5 | 35 |

�=0.60−0.461−0.46=0.2593

It shows greater similarity between A and B in the second case, compared to the first. This is because while the percentage agreement is the same, the percentage agreement that would occur 'by chance' is significantly higher in the first case (0.54 compared to 0.46).

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.

It mainly focus to reduce variance and bias.

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

Ans: Discriminative models focus on modeling the decision boundary between classes in a classification problem. The goal is to learn a function that maps inputs to binary outputs, indicating the class label of the input. Maximum likelihood estimation is often used to estimate the parameters of the discriminative model, such as the coefficients of a logistic regression model or the weights of a neural network. Discriminative models (just as in the literal meaning) separate classes instead of modeling the conditional probability and don’t make any assumptions about the data points. But these models are not capable of generating new data points. Therefore, the ultimate objective of discriminative models is to separate one class from another.

Example 1:

Descriptive statistics about a college involve the average math test score for incoming students.

Example 2:

a survey to 40 respondents about their favorite car color.

8. Describe how to evaluate a linear regression model.

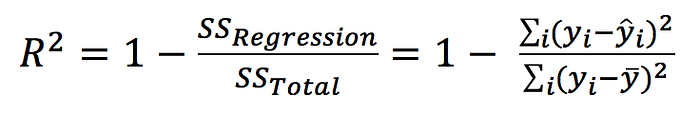
Ans: Evaluation metrics:

## Root Mean Squared Error(RMSE): 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.

A picture containing text, clock

Description automatically generated

# Coefficient of Determination (R2):R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that’s explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model’s inputs.



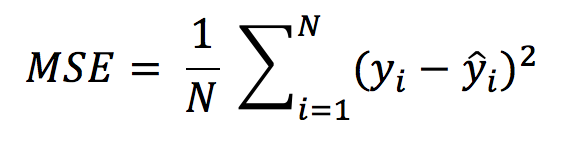
Mean absolute error (MAE) : is a measure of errors between observations and predictions. It is the average magnitude of the errors in a set of predictions, without considering their directions. It is the absolute value of error between actual and predicted value. MAE is calculated as:

Table

Description automatically generated with medium confidence

## 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.



9. Distinguish :

1. Descriptive vs. predictive models

The goal of descriptive models is to find an accurate understanding of the problem by asking questions from historical data, helps us understand the deep patterns from the data to uncover all those special features that were overlooked at the initial stage.

[Predictive modelling](https://www.educba.com/course/predictive-modling-and-implementation-using-ms-excel/) is a mathematical technique which uses statistics for prediction. It aims to work upon the provided information to reach an end conclusion after an event has been triggered. Models are known to make use of classifiers and detection theory to guess the probability of an outcome given a set of input data, it uses Group method of data handling approaches and models like [Naïve Bayes](https://www.educba.com/naive-bayes-algorithm/),K-nearest neighbor algorithm, Majority classifier,support vector machines,Boosted trees

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. Underfitting destroys the accuracy of our machine learning model.

 Reasons for Underfitting:

1. High bias and low variance
2. The size of the training dataset used is not enough.
3. The model is too simple.
4. Training data is not cleaned and also contains noise in it.

Techniques to reduce underfitting:

1. Increase model complexity
2. Increase the number of features, performing feature engineering
3. Remove noise from the data.
4. 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.

Reasons for Overfitting are as follows:

1. High variance and low bias
2. The model is too complex
3. The size of the training data

Techniques to reduce overfitting:

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization
5. Use dropout for neural networks to tackle overfitting.

3. Bootstrapping vs. cross-validation

1. Bootstrapping selects samples with replacements that can be as big as the dataset.
2. Cross-validation samples are smaller than the dataset.
3. Bootstrapping contains repeated elements in every subset. Bootstrapping relies on random sampling.
4. Cross-validation does not rely on random sampling, just splitting the dataset into k unique subsets.
5. Cross-validation is usually used to test an ML model's generalization capabilities.
6. Bootstrapping is used more for statistical tests, ensemble machine learning, and parameter estimation.

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. LOOCV, is a configuration of k-fold cross-validation where *k* is set to the number of examples in the dataset. LOOCV is an extreme version of k-fold cross-validation that has the maximum computational cost. It requires one model to be created and evaluated for each example in the training dataset.

### 2.F-measurement: The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which [classify](https://deepai.org/machine-learning-glossary-and-terms/classifier) examples into ‘positive’ or ‘negative’. The F-score is a way of combining the [precision and recall](https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall) of the model, and it is defined as the [harmonic mean](https://deepai.org/machine-learning-glossary-and-terms/harmonic-mean) of the model’s precision and recall.

### Fβ-score Formula:

The adjusted F-score allows us to weight precision or recall more highly if it is more important for our use case.

3. The width of the silhouette:  The Average Silhouette Width (ASW) is a popular cluster validation index to estimate the number of clusters. The ASW is an intuitive and simple measurement of cluster quality that does not rely on statistical model assumptions. Given that it is widely used and trusted for comparing the qualities of clusterings produced by various clustering methods over different numbers of clusters, it seems natural to investigate optimal ASW quality clustering not only over k but also for fixed k in order to integrate the problem for fixed k and the problem of finding the best k.

1. 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

1. True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

TPR=TP/TP+FN

1. False Positive Rate (FPR) is defined as follows:

FPR=FN/TP+FN

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

