Machine\_learning\_Assignment\_6

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

Ans – A model replicates a decision process to enable automation and understanding. AI/ML models are mathematical algorithms that are **“trained”** using data and human expert input to replicate a decision an expert would make when provided the same information.

This is the best way to train the machine learning model.

1. Being with existing data. Machine learning requires use to have existing data not the data our application will use when we run it. But data to learn from.
2. Analyze to identify pattern
3. Make prediction.

If we are talking about the best algorithm so it’s totally depend upon the problem statement.

**Q – 2 In the sense of machine learning, Explain the “no free lunch” theorem?**

Ans – The “no free lunch” theorem for supervised machine learning is a theorem the essentially implies that no single machine learning algorithm is universally the best performance algorithm for the all problems.

**Q – 3 Describe k-fold cross validation mechanism in details?**

Ans - Let’s say that you have trained a machine learning model. Now, you need to find out how well this model performs. Is it accurate enough to be used? How does it compare to another model? There are several evaluation methods to determine this. One such method is called K-fold cross validation.

Cross validation is an evaluation method used in machine learning to find out how well your machine learning model can predict the outcome of unseen data. It is a method that is easy to comprehend, works well for a limited data sample and also offers an evaluation that is less biased, making it a popular choice.

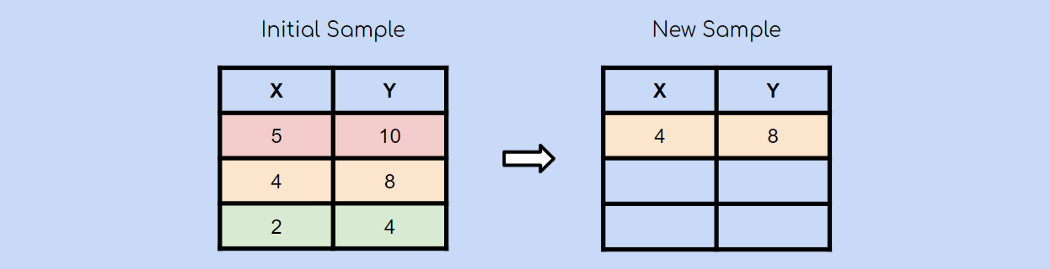
It helps us to avoid **overfitting.** As we know when a model is trained using all of the data in

a single short and give the best performance accuracy. To resist this k fold cross validation helps us to build the model is generalized one.

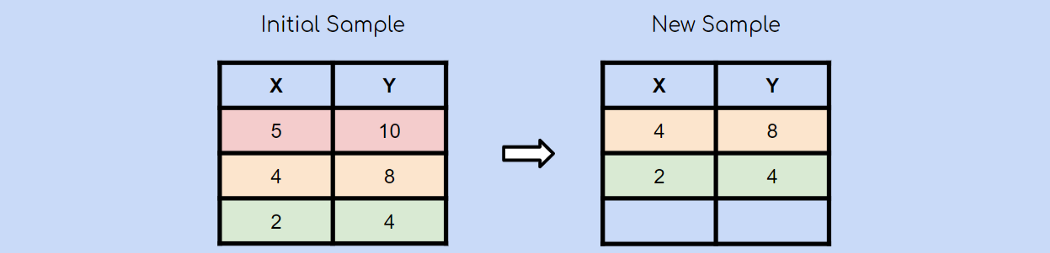
Q – 4 Describe the bootstrap sampling method. What is aim of it?

Ans - Technically speaking, the bootstrap sampling method is a resampling method that uses random sampling with replacement.

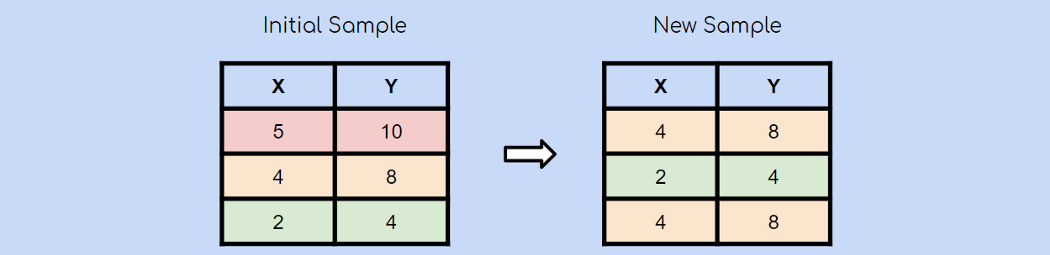
Don’t worry if that sounded confusing, let me explain it with a diagram:



Suppose you have an initial sample with 3 observations. Using the bootstrap sampling method, you’ll create a new sample with 3 observations as well. Each observation has an equal chance of being chosen (1/3). In this case, the second observation was chosen randomly and will be the first observation in our new sample.



After choosing another observation at random, you chose the green observation.



Lastly, the yellow observation is chosen again at random. Remember that bootstrap sampling using random sampling **with replacement**. This means that it is very much possible for an already chosen observation to be chosen again.

And this is the essence of bootstrap sampling!

**Importance of Bootstrap Sampling**

Great, now you understand what bootstrap sampling is, and you know how simple the concept is, but now you’re probably wondering what makes it so useful.

1. It is the building block for many modern machine learning algorithms

2. It can be used to estimate the parameters of a population

**Q – 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 score measures the degree of agreement between the two evaluators, also known as inter-rater 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 of 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 less indicate that the classifier is useless.

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

Ans – Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithm or using different training data sets. The ensemble model then aggregates the prediction of each base model and result in once final prediction for the unseen data.

**Q – 7 What is the descriptive model’s main purpose? Give example of real-world problem that descriptive model where use to solve.**

Ans – A descriptive model describe a system or other entity and its relationship to its environment. It is generally used to used to help specify and or understand what the system is, what it does represent geometric model or spatial model is a descriptive model the represents geometric and/or spatial relationships.

Example of descriptive model is :-

1. Observation.
2. Survey.
3. Case study.

**Q – 8 Describe how to evaluate the linear regression mode.**

Ans – There are three main metrics for model evaluation in regression:

1. R square /Adjust R square.
2. Mean square Error / Root mean square error
3. Mean absolute error.

**Q – 9 Distinguish between Descriptive vs. predictive models.**

Ans -

|  |  |
| --- | --- |
| **Descriptive models** | **Predictive model** |
| 1. Find causality, relationship and pattern between exploratory variable and dependent variable. | Fine cluster of data element with similar characteristic. |
| 1. Focus on specific variable. | Focus on as many variable |
|  |  |

**B. Underfitting vs. overfitting model**

**Underfitting =** A statistical model or a machine learning algorithms is said to have underfitting when it is can not capture underlying trends of the data i.e it is only perform well trained data but perform poorly on tested data (**its try to just trying to fit undersized plant)**

Underfitting destroys the accuracy of our machine learning model. Its occurance simply mean that our model or the algorithm does not fit 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 no linear data. In such a case the rule of the machine learning model are too easy and flexible to be applied to such minimal data and therefor the model will probably make a lot of wrong prediction. Underfitting can be avoided by using more data and also reducing the feature by features by feature selection

**Reasons for underfitting.**

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

**Techniques to reduce the overfitting**

1. Increase model complexity.
2. Increase the number of features, performing feature engineering.
3. Remove noise from the data.
4. Increase the number of epoch or increase the duration of training to get better result.

**Overfitting =** A Statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model get trained with so much data, it start 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 method because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefor they can really build unrealistic model.

A solution to avoid the 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.

**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 the model complexity.
3. Early stopping during the training phase.
4. Ridge regularization and lasso regularization.
5. Use dropout for neural networks to tackle overfitting.

**C. Bootstrapping =** Bootstrapping is the method of sample reuse that is much more general than cross-validation. The idea is to use the observed sample to estimate the population distribution. Then samples can be drawn from the estimated population and the sampling distribution of any type of estimator can itself be estimated.

The training records are sampled with replacement. A common variation is used is the .632 bootstrap which computers the overall accuracy by combining the accuracies of each bootstrap sample with the accuracy computed from a training set that contains all the labels example in the origin data.

**Limitations of the bootstrapping.**

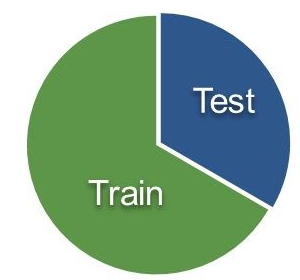
Bootstraps can yield poor results in certain solution.

Through the bootstrap easily accommodate some violation of traditional statistical assumptions.

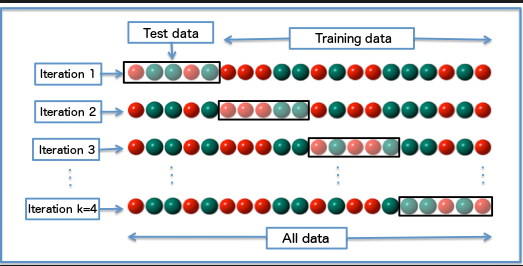
**Cross-Validations =** Cross validation is one of the technique used to test the effectiveness of a machine learning models it is also a re-sampling procedure used to evaluate a model if we have limited data.

**The few techniques of cross validation.**

1. Train\_test\_split approach.



1. **K- fold cross validation.**



**Q – 10 Make quick notes:-**

**A. LOOCV –** Leave one out cross validation is a type of cross validation 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 done and predicting using one observation validation set.

The advantages of LOOCV over random selection is zero randomness. Besides the bias will also be lower as the model is trained on the entire dataset, which consequently will not overestimate the test error rate. But its disadvantage is computational time.

**B. F-measurement =** An F- score is the harmonic mean of a system precision and recall value. It can be calculate by the following formula.

F-score = 2\*[(precision\*Recall)

(Precision + Recall)

The F-Score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural languages processing. If is possible to adjust the f score to give more importance to precision over recall , or vice versa.

**C. The width of the silhouette =** Silhouette width is a widely used index for assessing the fit of individuals objects in the classification as well as the quality of clusters and the entire classification

**D. Recover operating characteristics 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

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

**TPR = TP/TP+FN**

**False Positive Rate** (**FPR**) is defined as follows:

**FPR = FP/FP+TN**

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.