**Machine Learning Assignment 6**

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

Machine learning models are a subset of artificial intelligence that enables software to recognize specific patterns and make accurate predictions without needing to be directly programmed to do so.

A machine learning model functions as a mathematical expression of AI that consists of an algorithm capable of going through huge amounts of data to deduce a pattern and predict an outcome based on it.

ML models are divided into two main categories, supervised and unsupervised. Supervised ML models further branch into either regression or a classification pattern. Other than supervised and unsupervised learning, machine learning models can also involve semi-supervised learning or reinforcement learning.

In supervised ML models, the data fed to the algorithm is labeled with input and output being clearly defined and distinguished. The algorithm is also given clear instructions regarding the variable it needs to evaluate for pattern formation and predictions.

Unsupervised machine learning involves models working with data that is not labeled. The main function of the algorithm is to assess the data and pick out patterns based on any perceived similarities in the variables.

Semi-supervised learning is a mix of supervised and unsupervised learning. The algorithm is mostly fed labeled data but it can make out patterns based on this data without any supervision.

Reinforcement learning models are mostly used for operations that incorporate multiple steps. The algorithm is not bound by any strict rules and is free to assess data based on its understanding.

Machine learning models have a wide range of functions, but the scenarios in which they are used normally have some common themes. For example, situations, where a particular decision is repeatedly made, can be automated with the help of an ML model to ensure efficient, accurate, and consistent results because of the algorithm’s ability to recognize the questioning pattern.

Another opportunity to avail the use of an ML model is when data is divided into a distinct of inputs and outputs and needs to be matched together to map out an accurate resulting image.

Machine learning models can also be used in situations where determining a solution or criteria behind a decision is not possible.

Although different [types of machine learning](https://www.seldon.io/four-types-of-machine-learning-algorithms-explained/) will have different approaches to training the model, there are basic steps that are utilised by most models. Algorithms need large amounts of high quality data to be effectively trained. Many of the steps deal with the preparation of this data, so that the model can be as effective as possible. The whole project needs to be properly planned and managed from the beginning, so that a model fits the organisation’s specific requirements. So the initial step deals with contextualising the project within the organisation as a whole.

The six steps to building a machine learning model include:

1. Contextualise machine learning in your organisation
2. Explore the data and choose the type of algorithm
3. Prepare and clean the dataset
4. Split the prepared dataset and perform cross validation
5. Perform machine learning optimisation
6. Deploy the model

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

In the fields of optimization and machine learning, the No Free Lunch Theorem is frequently used, frequently without a clear grasp of what it states or implies.

According to the theorem, all optimization techniques perform equally well when their results are averaged across all possible issues. It suggests that there isn't just one best optimization procedure. There isn't a single machine learning technique that is best for predictive modeling tasks like classification and regression because of the close relationship between optimization, search, and machine learning.

They all concur on one thing: while all algorithms perform equally on average, there is no "optimal" method for any given type of algorithm. When the cost of computation to find a solution is averaged across all the issues in the class, it is the same for all solution methods. As a result, there is no quick cut in any solution.

In general, there are two No Free Lunch (NFL) theorems: one for search and optimization and one for machine learning. It is common to merge these two related theorems into a single general postulate (**the folklore theorem**).

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

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. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 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:
   1. Take the group as a hold out or test data set
   2. Take the remaining groups as a training data set
   3. Fit a model on the training set and evaluate it on the test set
   4. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores
5. **Describe the bootstrap sampling method. What is the aim of it?**

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

Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. This approach to sampling is called sampling with replacement.

The process for building one sample can be summarized as follows:

1. Choose the size of the sample.
2. While the size of the sample is less than the chosen size
   1. Randomly select an observation from the dataset
   2. Add it to the sample

The bootstrap method can be used to estimate a quantity of a population. This is done by repeatedly taking small samples, calculating the statistic, and taking the average of the calculated statistics. We can summarize this procedure as follows:

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.**

**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:

K =

Where is the observed agreement and 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.

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

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in [machine learning](https://courses.corporatefinanceinstitute.com/courses/machine-learning-python-fundamentals).

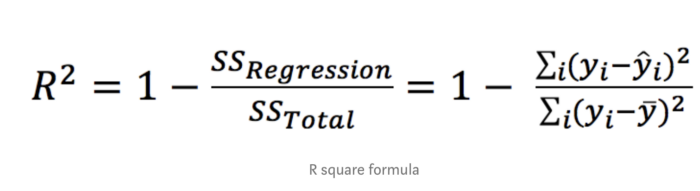
* Ensemble methods aim at improving predictability in models by combining several models to make one very reliable model.
* The most popular ensemble methods are boosting, bagging, and stacking.
* Ensemble methods are ideal for regression and classification, where they reduce bias and variance to boost the accuracy of models.

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

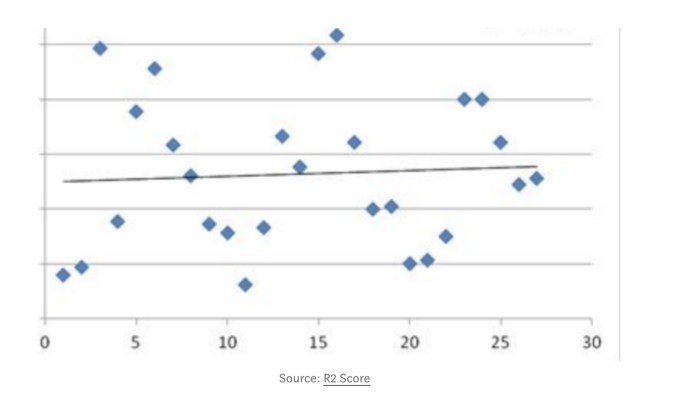
A descriptive model describes a system or other entity and its relationship to its environment. It is generally used to help specify and/or understand what the system is, what it does, and how it does it.

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

In regression problem, accuracy in regression model is slightly harder to illustrate. It is impossible for you to predict the exact value but rather how close your prediction is against the real value. Here , In this blog, I am trying to emphasize that we should evaluate every single metrics to make sure good performance of model rather than only *R square* metric. Because some of time , when we get good R square number like 0.95 then we assume model can predict more accurate . But it doesn’t happen always true. So How to evaluate regression model , let’s start

1. *1. R Square/Adjusted R Square*
2. *2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)*
3. *3. Mean Absolute Error(MAE)*
4. *4. illustrate Residual of model as a normal distribution ( bell shape)*
5. 5. By OLS from statemodels.formula
6. R Square/Adjusted R Square :
7. This is a first measure of regression model especially we, everybody, do during evaluation because it is easy to interpret score between 0 to 1. If we see good score like close to 1, then we assume that model is good fit. Of course , R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data.So I recommend that we have to see all perspective for better evaluation . let’s talk what is actually mean R² . R² is calculated by the sum of squared of prediction error divided by the total sum of square which replace the calculated prediction with mean. R Square value is between 0 to 1 and bigger value indicates a better fit between prediction and actual value.
8. 

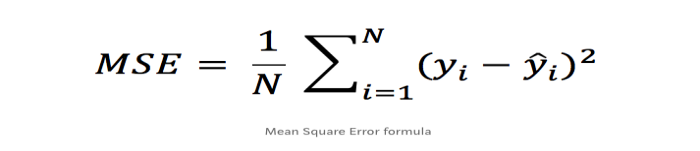
Sometime , R² is very helpful to measure error on model than Mean Square Error(MSE) and Mean Absolute Error(MAE). For instance , We can say R² is perfect measure to give you how is model like on below figure.



Here , I also want to focus R² Adjust measure too. Sometime, We see R² and R² Adjust same score. But When we do fine tuning to model to get better accuracy then R² Adjust help us to better understand. It happen when we add more independent features or penalize more feature due to over fitting . Then we can see different score between on these measures.

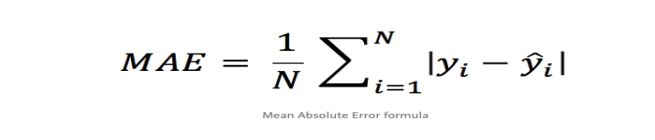
Mean Square Error(MSE)/Root Mean Square Error(RMSE):

while R² is a relative measure of how well the model fit dependent variables, whereas Mean Square Error is an absolute measure of the fit of model. MSE is calculated by sum of square of prediction error. Where prediction error is minus between true values and prediction values, and then it is made by square because we avoid negative error score. It’s result gives us how much deviation from actual number. It’s number might be larger number which may be like uncommon . you might be question how is error score is too big .



Mean Absolute Error(MAE):

This is almost same to Mean Square Error metric but only MAE take absolute error value instead of square of predicted error for avoiding negative score . However, here , we don’t need to calculate Root of MAE score . We can interpret directly the score with real values.



**9. Distinguish :**

**1. Descriptive vs. predictive models**

Below is the detailed explanation of Predictive Analytics and Descriptive Analytics:

* Descriptive Analytics will give you a vision into the past and tells you: what has happened? Whereas the Predictive Analytics will recognize the future and tells you: What might happen in future?
* Descriptive Analytics uses Data Aggregation and [Data Mining techniques](https://www.educba.com/data-mining-techniques/) to give you knowledge about past but Predictive Analytics uses Statistical analysis and Forecast techniques to know the future.
* Descriptive Analytics is used when you need to analyze and explain different aspects of your organization whereas Predictive Analytics is used when you need to know anything about the future and fill the information that you do not know.
* A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report. In a [Predictive model](https://www.educba.com/predictive-modeling/), it identifies patterns found in past and transactional data to find risks and future outcomes.
* Descriptive analytics will help an organization to know where they stand in the market, present facts and figures. Whereas predictive analytics will help an organization to know, how they will stand in the market in future and forecasts the facts and figures about the company.
* Reports generated by Descriptive analysis are accurate but the reports generated by Predictive analysis are not 100% accurate it may or may not happen in future.

1. **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:**

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. 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:

1. High variance and low bias
2. The model is too complex
3. The size of the training data
4. **Bootstrapping vs. cross-validation**

Bootstrapping is any test or metric that relies on random sampling with replacement.It is a method that helps in many situations like validation of a predictive model performance, ensemble methods, estimation of bias and variance of the parameter of a model etc. It works by performing sampling with replacement from the original dataset, and at the same time assuming that the data points that have not been choses are the test dataset. We can repeat this procedure several times and compute the average score as estimation of our model performance. Also, Bootstrapping is related to the ensemble training methods, because we can build a model using each bootstrap datasets and “bag” these models in an ensemble using the majority voting (for classification) or computing the average (for numerical predictions) for all of these models as our final result.

Cross validation is a procedure for validating a model's performance, and it is done by splitting the training data into k parts. We assume that the k-1 parts is the training set and use the other part is our test set. We can repeat that k times differently holding out a different part of the data every time. Finally, we take the average of the k scores as our performance estimation. Cross validation can suffer from bias or variance. Increasing the number of splits, the variance will increase too and the bias will decrease. On the other hand, if we decrease the number of splits, the bias will increase and the variance will decrease.

In summary, Cross validation splits the available dataset to create multiple datasets, and Bootstrapping method uses the original dataset to create multiple datasets after resampling with replacement. Bootstrapping it is not as strong as Cross validation when it is used for model validation. Bootstrapping is more about building ensemble models or just estimating parameters.

**10. Make quick notes on:**

**1. LOOCV.**

Leave-one-out cross-validation, or 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.

The benefit of so many fit and evaluated models is a more robust estimate of model performance as each row of data is given an opportunity to represent the entirety of the test dataset.

Given the computational cost, LOOCV is not appropriate for very large datasets such as more than tens or hundreds of thousands of examples, or for models that are costly to fit, such as neural networks.

**2. F-measurement**

An F-score is the harmonic mean of a system’s precision and recall values. It can be calculated by the following formula: 2 x [(Precision x Recall) / (Precision + Recall)]. Criticism around the use of F-score values to determine the quality of a predictive system is based on the fact that a moderately high F-score can be the result of an imbalance between precision and recall and, therefore, not tell the whole story. On the other hand, systems at a high level of accuracy struggle to improve precision or recall without negatively impacting the other.

Critical (risk) applications that value information retrieval more than accuracy (i.e., producing a large number of false positives but virtually guaranteeing that all the true positives are found) can adopt a different scoring system called F2 measure, where recall is weighed more heavily. The opposite (precision is weighed more heavily) is achieved by using the F0.5 measure.

**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. Silhouette combines two clustering criteria, compactness and separation, which imply that spherical cluster shapes are preferred over others—a property that can be seen as a disadvantage in the presence of complex, nonspherical clusters, which is common in real situations. We suggest a generalization of the silhouette width using the generalized mean

1. **Receiver operating characteristic curve**

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.