Machine Learning Assignment-1

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

**R-squared** (also known as the coefficient of determination) measures the proportion of variation in the dependent variable that is explained by the independent variables in the model. In other words, it indicates how well the model fits the data, with values ranging from 0 to 1. Higher R-squared values indicate a better fit, as they mean that a larger proportion of the variation in the dependent variable is explained by the independent variables in the model.

On the other hand, **RSS** measures the total sum of squared differences between the actual values of the dependent variable and the predicted values by the model. It represents the amount of unexplained variation in the data, and lower RSS values indicate a better fit, as they mean that the model is able to explain more of the variation in the data.

Therefore, both measures are useful in evaluating the goodness of fit of a model, but they serve different purposes. R-squared is a useful measure to assess the overall fit of the model and to compare different models, while RSS is useful to identify the degree of the error in the model's predictions.

In general, a good model should have both a high R-squared value and a low RSS value, indicating that it explains a large proportion of the variation in the dependent variable and has a low degree of error in its predictions.

1. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

**TSS**- The Total SS (TSS or SST) tells you how much variation there is in the [dependent variable](https://www.statisticshowto.com/dependent-variable-definition/).  
Total SS = Σ(Yi – mean of Y)2.

**ESS**- The Explained SS tells you how much of the variation in the dependent variable your model explained.  
Explained SS = Σ(Y-Hat – mean of Y)2.

**RSS-** The residual sum of squares tells you how much of the dependent variable’s variation your model **did not explain**. It is the sum of the squared differences between the actual Y and the predicted Y:  
Residual Sum of Squares = Σ e2

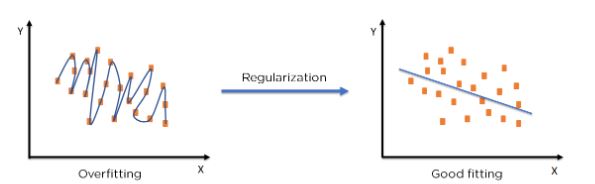
Equation relating these three metrics are as follows

TSS=ESS+ RSS

1. What is the need of regularization in machine learning?

**Regularization** refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.

While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.



1. What is Gini–impurity index?

**Gini Impurity index** is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

It is the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled according to the class distribution in the dataset.

1. Are unregularized decision-trees prone to overfitting? If yes, why?

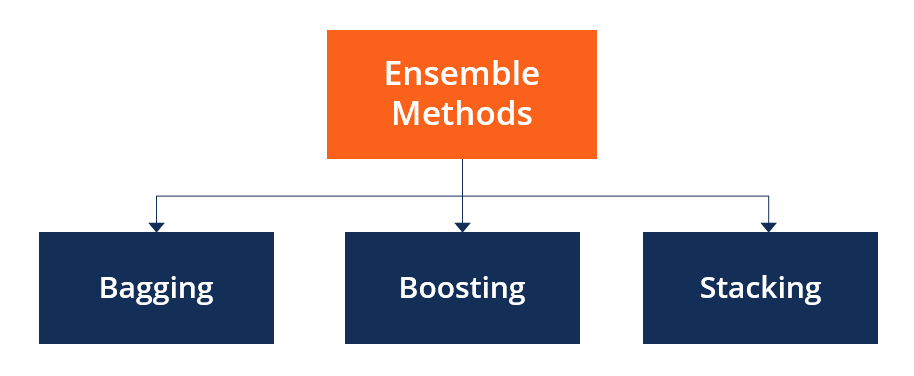
Yes, Unregularized decision- trees prone to overfitting.

**Regularization** is a technique to prevent the model from overfitting by adding extra information to it. It mainly regularizes or reduces the coefficient of features toward zero.

Unregularized decision tree is not able to predict the output when deals with unseen data by introducing noise in the output, hence leading a situation called overfitting.

1. What is an ensemble technique in machine learning?

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://corporatefinanceinstitute.com/course/learn-python/).



Ensemble methods fall into two broad categories, i.e., sequential ensemble techniques and parallel ensemble techniques.

**Sequential ensemble techniques** generate base learners in a sequence, e.g., Adaptive Boosting (AdaBoost). The sequential generation of base learners promotes the dependence between the base learners. The performance of the model is then improved by assigning higher weights to previously misrepresented learners.

**Parallel ensemble techniques**, base learners are generated in a parallel format, e.g., random forest. Parallel methods utilize the parallel generation of base learners to encourage independence between the base learners. The independence of base learners significantly reduces the error due to the application of averages.

1. What is the difference between Bagging and Boosting techniques?

Difference between bagging and Boosting

|  |  |
| --- | --- |
| Bagging | Boosting |
| It is a homogenous weak learner’s model that learns from each other independently in parallel and combines them for determining the model’s average | It is also a homogenous weak learner’s model but here, learners learn sequentially and adaptively to improve model predictions of a learning algorithm |
| It decreases variance, not bias, and solves over-fitting issues in a model | It decreases bias, not variance. |
| Here, each model receives an equal weight | Here, models are weighed based on their performance. |

1. What is out-of-bag error in random forests?

**Out-of-bag( Oob)** errors are an estimate of the performance of a random forest classifiers or regressor of unseen data. The Oob error can be obtain using the oo\_ score\_ attribute of the random forest classifier or regressor.

The oob error is computed using the samples that were not included in the training of the individual trees.

The oob error can be useful for evaluating the performance of the random forest on unseen data . It is not always the reliable estimate of the generalization error model of the, but it can provide a useful indication of how well the model is performing.

1. What is K-fold cross-validation?

**Cross-validation** is a resampling technique used to validate machine learning models against a limited sample of data. In this article we will talk about K-fold Cross-validation and its advantages and disadvantages.Its working is shown with python program.

In **K-fold cross-validation**, the data set is divided into a number of K-folds and used to assess the model’s ability as new data become available. K represents the number of groups into which the data sample is divided. For example, if you find the k value to be 5, you can call it 5-fold cross-validation. Each fold is used as a test set at some point in the process.

1. Randomly shuffle the dataset.
2. Divide the dataset into k folds
3. For each unique group:
   * Use one fold as test data
   * Use remaining groups as training dataset
   * Fit model on training set and evaluate on test set
   * Keep Score

4. Get accuracy score by applying mean to all the accuracies received for all folds

A diagram of a cross-validation process

Description automatically generated

A diagram of training and test

Description automatically generated

1. What is hyper parameter tuning in machine learning and why it is done?

**Hyperparameters** directly control model structure, function, and performance. Hyperparameter tuning allows data scientists to tweak model performance for optimal results. It is used to select the optimal values for a machine learning model’s hyperparameters. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success.

It is used to select the optimal values for a machine learning model’s hyperparameters.

1. What issues can occur if we have a large learning rate in Gradient Descent?

The choice of learning rate can significantly impact the performance of gradient descent. If the learning rate is too high, the algorithm may overshoot the minimum, and if it is too low, the algorithm may take too long to converge.

1. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

**Logistic regression** has traditionally been used to come up with a hyperplane that separates the feature space into classes. But if we suspect that the decision boundary is nonlinear we may get better results by attempting some nonlinear functional forms for the logit function.

1. Differentiate between Adaboost and Gradient Boosting.

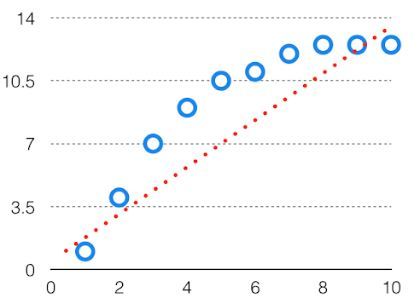
**Gradient boosting** is a type of boosting algorithm. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. Gradient Boosting can be used for both Classification and Regression. It is a greedy algorithm and can overfit a training dataset easily. It can benefit from regularization methods that penalize various parts of the algorithm and generally improve the performance of the algorithm by reducing overfitting.

The difference between Gradient Boost and Ada Boost lies in what it does with the underfitted values of its predecessor.

There is not reweighting of the samples take place in GBM. Contrary to AdaBoost, which tweaks the instance weights at every interaction, this method tries to fit the new predictor to the residual errors made by the previous predictor.

1. What is bias-variance trade off in machine learning?

**Bias is** the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.



**Variance** is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before. As a result, such models perform very well on training data but has high error rates on test data.

A graph showing a curve

Description automatically generated

In **bias variance tradeoff,** if the algorithm is too simple (hypothesis with linear equation) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree equation) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as a Trade-off or Bias Variance Trade-off. This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time. For the graph, the perfect tradeoff will be like this.

A graph showing the value of a number of points

Description automatically generated with medium confidence

1. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

**Linear Kernel**

It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for [**text-classification problems**](https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/) as most of these kinds of classification problems can be linearly separated.

Linear kernel functions are **faster** than other functions.

Linear Kernel Formula

**F(x, xj) = sum( x.xj)**

Here, **x, xj** represents the data you’re trying to classify.

**Radial Basis Function (RBF)**

It is one of the most preferred and used kernel functions in svm. It is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data

Gaussian Radial Basis Formula

**F(x, xj) = exp(-gamma \* ||x - xj||^2)**

The value of gamma varies from **0 to 1**. You have to manually provide the value of gamma in the code. The most preferred value for **gamma is 0.1**.

**Polynomial Kernel**

It is a more generalized representation of the linear kernel. It**is not** as preferred as other kernel functions as it is **less efficient** and accurate.

#### Polynomial Kernel Formula

**F(x, xj) = (x.xj+1)^d**

Here ‘.’ shows the **dot product** of both the values, and **d** denotes the degree.

F(x, xj) representing the **decision boundary** to separate the given classes.

**MACHINE LEARNING ASSIGNMENT-2**

1. Using a goodness of fit,we can assess whether a set of obtained frequencies differ from a set of frequencies.

a) Mean

b) Actual  
c) Predicted

d) Expected

1. Chisquare is used to analyse

a) Score

b) Rank  
c) Frequencies

d) All of these

1. What is the mean of a Chi Square distribution with 6 degrees of freedom?

a) 4

b) 12

c) 6

d) 8

1. Which of these distributions is used for a goodness of fit testing?

a)  Normal distribution

b)  Chi-squared distribution

c)  Gamma distribution

d)  Poission distribution

1. Which of the following distributions is Continuous

a) Binomial Distribution  
b) Hypergeometric Distribution

c) F-Distribution  
d) Poisson Distribution

1. A statement made about a population for testing purpose is called?

a) Statistic

b) Hypothesis  
c) Level of Significance

d) TestStatistic

1. If the assumed hypothesis is tested for rejection considering it to be true is called?

a) Null Hypothesis  
b) Statistical Hypothesis  
c) SimpleHypothesis

d) Composite Hypothesis

1. If the Critical region is evenly distributed then the test is referred as?

a) Two-tailed

b) One tailed

c) Three-tailed

d) Zero tailed

1. Alternative Hypothesis is also called as?

a) Composite hypothesis  
b) Research Hypothesis  
c) SimpleHypothesis

d)Null hypothesis

1. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by

a)np

b) n