ASSIGNMENT- 2
MACHINE LEARNING (SET 1)

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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?**

   Answer: R-squared is generally a better measure of the goodness of fit for a regression model than the residual sum of squares (RSS). R-squared, denoted as R^2, is a statistical measure that represents the portion of the variance for the Labels(Dependent Variables) that's explained by the Features(Independent Variables) in the model. R-squared is dimensionless and ranges from (0-1), where a value closer to 1 indicates a better fit and near zero indicates a poor fit. R^2 is calculated using the formula: [R^2= 1 – RSS/TSS]

2. **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.**

   Answer: a.) TSS(Total sum of squares)- It measures how much variation there is in the observed data.

   b.) ESS(Explained sum of squares)- It is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.

   c.) RSS(Residuals sum of squares)- It is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term. It also known as SSE(sum of square error).

   TSS (total sum of squares) is equal to ESS (explained sum of squares) plus RSS (residual sum of squares).

   TSS = $\sum(Yi - \bar{Y})^2$, where Yi is the actual value of the response variable for observation i, and $\bar{Y}$ is the mean of the response variable.

ESS = ∑(Ŷi - Ȳ)^2, where Ŷi is the predicted value of the response variable for observation i.

RSS = ∑(Yi - Ŷi)^2, which is the sum of squared differences between the actual and predicted values of the response variable.

TSS = ∑(Yi - Ȳ)² = ∑[(Yi - Ŷi) + (Ŷi - Ȳ)]² = ∑(Yi - Ŷi)² + ∑(Ŷi - Ȳ)² + 2∑(Yi - Ŷi)(Ŷi - Ȳ)

Next, we can use the definition of RSS to simplify the first term on the right-hand side:

∑(Yi - Ŷi)² = RSS

Similarly, we can use the definition of ESS to simplify the second term:

∑(Ŷi - Ȳ)² = ESS

Now, we need to simplify the third term using some algebraic manipulations. We can start by expanding the product:

2∑(Yi - Ŷi)(Ŷi - Ȳ) = 2∑(YiŶi - YiȲ - ŶiȲ + Ŷi²)

Then, we can use the fact that the sum of the residuals (Yi - Ŷi) is zero, which can be shown as follows:

∑(Yi - Ŷi) = ∑Yi - ∑Ŷi = nȲ - nȲ = 0

Using this fact, we can simplify the third term:

$2\sum(Yi - \hat{Y}i)(\hat{Y}i - \bar{Y}) = 2\sum(Yi\hat{Y}i - \hat{Y}i\bar{Y}) = 2(\sum Yi\hat{Y}i - \sum\hat{Y}i\bar{Y}) = 2(\sum\hat{Y}iYi - n\bar{Y}^2) = 2(ESS - n(\bar{Y} - \hat{Y})^2)$

where $\hat{Y}$ is the sample mean of the predicted values, which is equal to $\bar{Y}$.

Substituting these results back into the equation for TSS, we get:

$TSS = RSS + ESS + 2(ESS - n(\bar{Y} - \hat{Y})^2) = RSS + 2ESS - 2n(\bar{Y} - \hat{Y})^2 = RSS + 2ESS - 2n(\bar{Y} - \bar{Y})^2 = RSS + 2ESS - 0 = RSS + ESS$
Therefore, we have shown that TSS is equal to ESS plus RSS.

3. **What is the need of regularization in machine learning?**

Answer: Regularization in Machine Learning is technique uses in ML to avoid overfitting which can leads to bad performance on the new set of data. Overfitting occurs when a model learns too much from the Training(Train_test_split) data, including (error)noise or irrelevant pattern and can't form a new data. Then Regularization comes into ML for helping model to focus on the given data which can leads model to improve their ability to make accurate predictions on new data.

4. **What is Gini–impurity index?**

Answer: In Decision Tree which is a type of Supervised Machine Learning , uses hierarchal tree to predict based on previously trained data. By looking at leaf nodes we says that the Gini Impurity is 0 or anything for all the leaf node . This means these are the Pure node and not needed to divided further . Gini Impurity is Higher at the Root Node and 0 at the leaf Node. The range of value Gini Impurity can have is between 0 to 0.5, The lesser the Gini Impurity, the better the split is. A Gini Impurity of 0 denotes a pure node and 0.5 denotes a most impure node.

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

Answer: Yes, an unregularized decision trees prone to overfitting. Overfitting can be problem that describes if our model no longer generalizes properly.It prone to overfitting , especially when a tree is particularly big. This is due to the amount of specificity we look at

leading to a smaller sample of events that meets the previous assumption. In the case of decision trees, they can learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is the behavior of this model that makes it prone to learning every point extremely well to the point of perfect classification i.e.: overfitting.

6. **What is an ensemble technique in machine learning?**

Answer: In Machine Learning Ensemble Technique is the combination of multiple models to improve performance and reduces the Overfitting in the model.These techniques that aim at improving the accuracy of results in models by combining multiple models instead of using singe one. Ensemble methods which are BAGGING / RANDOM FOREST/ BOOSTING etc , that allows for weighting different aspects of decision trees with bootstrapping, with random variable selection, with weighting of weak learners, with sample weight selection. these are the main aspects that different ensembles mix and match to generalize better results.

7. **What is the difference between Bagging and Boosting techniques?**

Answer: These two Ensemble techniques decrease the variance of a single estimate Umbrella break up by several estimates from different models. So, the result may be a model with higher stability.

Bagging:

- Combines multiple models trained on different subsets of data.
- Its objective is to reduce variance by averaging out individual model error.
- In Data sampling Use Bootstrap to create subsets of the data.
- Each model serves equal weight in the final decision.
- Each model has an equal error rate.
- Less prone to overfitting due to average mechanism.
- Improves accuracy by reducing variance.

Boosting:

- Train models sequentially, focusing on the error made by the previous model.
- Its Objective is to reduces both bias and variance by correcting misclassifications of the previous model.
- Re-weights the data based on the error from the previous model, making the next models focus on misclassified instances.

- Models are weighted based on accuracy, i.e., better-accuracy models will have a higher weight.
- It gives more weight to instances with higher error, making subsequent models focus on them.
- Generally, not prone to overfitting, but it can be if the number of the model or the iteration is high.
- Achieves higher accuracy by reducing both bias and variance.

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

Answer: Out-of-bag error (OOB) is a method of measuring the prediction accuracy of Random Forests and other models of Machine Learning that use Bootstrap. Like Bagging, it's a statistical resampling technique that uses subsampling with replacement to create training samples for the model to learn from. its means that some samples are likely to be selected multiple times, and others may not be selected at all. The unselected samples are called out-of-bag samples, and the error achieved on these samples is called an OOB error. OOB error allows the Random Forest Classifier to be fit and validated while it's being trained. It can also be used to approximate a suitable value of n_estimators, which is the total number of trees in the random forest, at which the error stabilizes.

9. What is K-fold cross-validation?

Answer: Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set. The main purpose of cross-validation is to prevent overfitting, which occurs when a model is trained too well on the training data and performs poorly on new.

K-Fold Cross Validation:

In K-Fold Cross Validation, we split the dataset into k number of subsets known as folds then we perform training on all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purposes each time.

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

Answer: There is Hyper Parameter Tuning in Machine Learning which refers to the process that involves finding the best parameter values for learning algorithm to optimize model performance. Hyperparameters are an adjustable parameter that controls the behavior of a model or algorithm. A higher learning rate leads to a model converging too quickly, while a low rate might make training take too long.
Hyperparameter tuning is done because it can help balance the simplicity and complexity which can easily improves generalization and avoid overfitting or underfitting(focussing on a particular feature).
It gives the outputs of these several questions:

- What degree of polynomial features should be used for a linear model?
- What should be the maximum depth allowed for a decision tree?
- How many neurons should be in a neural network layer?

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

Answer: Gradient descent can overfit the training data if the model is too complex or the learning rate is too high. If the learning rate is large in Gradient Descent it can suffer from divergence. This means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values.

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

Answer: We can't use Logistic Regression for Classification of Non-Linear Data because it has a linear decision surface.

13. Differentiate between Adaboost and Gradient Boosting.

Answer: Gradient Boosting and Adaboost both are ensemble techniques that combines multiple weak learners to create a strong learner.

- AdaBoost (Adaptive Boosting): In AdaBoost, the algorithm assigns weights to the training instances and focuses on the misclassified instances in each iteration. It increases the weight of misclassified instances so that subsequent weak learners focus more on these instances.

- Gradient Boosting: In gradient boosting, the algorithm fits the new weak learner to the residual errors made by the existing ensemble. It uses the gradient of the loss function with respect

to the predictions of the ensemble to update the model in the direction that minimizes the loss.

- AdaBoost: AdaBoost typically uses the exponential loss function, which puts more emphasis on the misclassified instances.

- Gradient Boosting: Gradient boosting can work with various loss functions, such as squared error loss (for regression problems) or log loss (for classification problems). The choice of loss function can be tailored to the specific problem being addressed.

- AdaBoost: AdaBoost focuses on building a sequence of weak learners, where each weak learner tries to correct the mistakes of the previous ones.
- Gradient Boosting: Gradient boosting typically uses decision trees as weak learners, building trees sequentially where each new tree is trained to predict the residual errors of the existing ensemble.

- AdaBoost: AdaBoost uses a learning rate parameter that controls the contribution of each weak learner to the final prediction.

- Gradient Boosting: Gradient boosting also uses a learning rate, but it is typically lower than in AdaBoost. The learning rate in gradient boosting scales the contribution of each tree, helping to prevent overfitting.

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

Answer: Bias- Variance trade-off is a fundamental concept describes the relationship between a model's complexity, its accuracy, and how well it can predict unseen data. It's a property of all supervised machine learning models, and it involves managing two types of errors: bias and variance.

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

Answer: **Linear Kernel**- A linear kernel is a type of kernel function used in machine learning, including in SVMs (Support Vector Machines). It is the simplest and most commonly used kernel function, and it defines the dot product between the input vectors in the original feature space. K(x,y)=x.y, where x and y are the input feature vectors. The dot product of the input vectors is a measure of their similarity or distance in the original feature space.

Polynomial Kernel
A particular kind of kernel function utilised in machine learning, such as in SVMs, is a polynomial kernel (Support Vector Machines). It is a nonlinear kernel function that employs polynomial functions to transfer the input data into a higher-dimensional feature space. The degree of nonlinearity in the decision boundary is determined by the degree of the polynomial.

RBF(Gaussian) Kernel
The Gaussian kernel, also known as the radial basis function (RBF) kernel, is a popular kernel function used in machine learning, particularly in SVMs (Support Vector Machines). It is a nonlinear kernel function that maps the input data into a higher-dimensional feature space using a Gaussian function.

STATISTICS WORKSHEET (SET- 2)

Q1 to Q10 are MCQs with only one correct answer. Choose the correct option.

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 (**ANSWER)**

2. Chisquare is used to analyse
a) Score
b) Rank
c) Frequencies (**ANSWER)**
d) All of these

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?
a) 4
b) 12
c) 6 (**ANSWER)**
d) 8

4. Which of these distributions is used for a goodness of fit testing?
a) Normal distribution
b) Chisqared distribution (**ANSWER)**
c) Gamma distribution
d) Poission distribution

5. Which of the following distributions is Continuous
a) Binomial Distribution
b) Hypergeometric Distribution
c) F Distribution (**ANSWER)**
d) Poisson Distribution

6. A statement made about a population for testing purpose is called?
a) Statistic
b) Hypothesis (**ANSWER)**
c) Level of Significance
d) TestStatistic

7. If the assumed hypothesis is tested for rejection considering it to be true is called?
a) Null Hypothesis (**ANSWER)**
b) Statistical Hypothesis
c) Simple Hypothesis
d) Composite Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?
a) Two tailed (**ANSWER)**

b) One tailed
c) Three tailed
d) Zero tailed

9. Alternative Hypothesis is also called as?
a) Composite hypothesis
b) Research Hypothesis (**ANSWER)**
c) Simple Hypothesis
d) Null Hypothesis

10. 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 (**ANSWER)**
b) n