WORKSHEET 5

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 (R^2) and Residual Sum of Squares (RSS) are both commonly used measures to assess the goodness of fit of a regression model, but they capture different aspects of model performance, and the choice between them depends on the context and what you want to evaluate.

1. R-squared (R^2):

- R-squared is a measure of the proportion of the variance in the dependent variable that is explained by the independent variables in your regression model.

- R-squared ranges from 0 to 1, with higher values indicating a better fit. A value of 1 means that the model perfectly explains the variance in the dependent variable, while a value of 0 means that the model provides no explanatory power.

- R-squared is a relative measure, and it does not provide information about the absolute goodness of fit or the quality of the model's predictions.

R-squared can be useful for comparing different models or assessing how much of the variation in the dependent variable can be attributed to the predictors. However, it has limitations. For example, it can be artificially inflated by adding more predictors to a model, even if those predictors do not have a meaningful relationship with the dependent variable.

2. Residual Sum of Squares (RSS):

- RSS measures the total squared difference between the observed values of the dependent variable and the predicted values from the regression model. It quantifies the overall error or "residuals" in the model's predictions.

- A smaller RSS indicates a better fit because it means that the model's predictions are closer to the actual observed values.

RSS is an absolute measure of the goodness of fit. It tells you how well the model fits the data in terms of minimizing prediction errors. Unlike R-squared, RSS does not provide information about the proportion of variance explained but gives you a direct measure of the model's prediction accuracy.

Which Measure to Use:

- R-squared is often used when you want to understand the proportion of variance explained by the model, especially in the context of comparing different models. It can provide insights into how well the predictors collectively contribute to explaining the variation in the dependent variable.

- RSS is useful when you want to evaluate the absolute goodness of fit, focusing on the magnitude of the prediction errors. If minimizing prediction errors is a primary concern, RSS is a more appropriate choice.

In practice, both measures can be valuable. R-squared provides a high-level summary of the explanatory power of your model, while RSS gives you a detailed view of the model's prediction accuracy. The choice between them should align with your specific goals and the questions you want to answer about your regression model.

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.

Residual sum of squares (**RSS**), also known as the **sum of squared residuals** (**SSR**) or the **sum of squared estimate of errors** (**SSE**), is the sum of the squares of residential (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model, such as a linear regression. A small RSS indicates a tight fit of the model to the data. It is used as an optimal criterion in parameter selection and model selection.

3. What is the need of regularization in machine learning?
Regularization is the process of adding tuning parameter to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

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.

4. **What is Gini–impurity index?**

The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits.

Gini impurity measures how often a randomly chosen element of a set would be incorrectly labeled if it were labeled randomly and independently according to the distribution of labels in the set. It reaches its minimum (zero) when all cases in the node fall into a single target category.

5). **Are unregularized decision-trees prone to overfitting? If yes, why?**
Decision trees are prone to overfitting when they capture noise in the data. Pruning and setting appropriate stopping criteria are used to address this assumption

6. **What is an ensemble technique in machine learning?**
Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

7). **What is the difference between Bagging and Boosting techniques?**
Bagging is a method of merging the same type of predictions. Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions.
Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.
Bagging is also an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa. Boosting in general decreases the bias error and builds strong predictive models. However, they may over fit on the training data.

8). **What is out-of-bag error in random forests?**
Out-of-Bag Error, also known as OOB Error, is a concept used in ensemble machine learning algorithms such as random forests. When building a random forest model, each tree is trained using a subset of the original data, known as the bootstrap sample. During the training process, some observations are left out or "out-of-bag" (OOB) for each tree.

The OOB observations that were not used in the training of a particular tree can be considered as a validation set for that tree. The model's prediction accuracy on the OOB observations can then be calculated and averaged across all the trees to obtain the OOB Error.

9). **What is K-fold cross-validation?**
K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model's generalization performance

10. **What is hyper parameter tuning in machine learning and why it is done?**
A hyperparameter is a parameter whose value is set before the learning process begins. It determines how a network is trained and the structure of the network (such as the number of hidden units, the learning rate, epochs, etc.).
Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success. Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

11). **What issues can occur if we have a large learning rate in Gradient Descent?**
The learning rate is an important hyperparameter that greatly affects the performance of gradient descent. It determines how quickly or slowly our model learns, and it plays an important role in controlling both convergence and divergence of the algorithm. When the learning rate is too large, gradient descent 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. On the other hand, if the learning rate is too small, then gradient descent can suffer from slow convergence or even stagnation—which means it may not reach a local minimum at all unless many iterations are performed on large datasets.

In order to avoid these issues with different learning rates for each parameter/variable, we use adaptive techniques such as Adagrad and Adam which adjust their own learning rates throughout training based on real-time observations of parameters during optimization (i.e., they control exploration/exploitation trade-offs). These adaptive measures ensure better results than standard gradient descent while avoiding potential pitfalls in

terms of either massive gains or slow losses due to misconfigured static global learning rates like those used with traditional gradient descent algorithms.
A too high learning rate will make the learning jump over minima

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

Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.

13). **Differentiate between Adaboost and Gradient Boosting.**
Adaboost is computed with a specific loss function and becomes more rigid when comes to few iterations. But in gradient boosting, it assists in finding the proper solution to additional iteration modeling problem as it is built with some generic features.

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

In statistics and machine learning, the bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.

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

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

An SVM with a linear kernel learns a linear decision boundary in the original feature space. A kernel SVM, on the other hand still learns a linear decision boundary, but in a transformed space. For example, a radial basis function SVM learns a linear boundary in an infinite dimensional space.

Statistics Worksheet

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
Anwer d

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

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

4. 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
Answer b

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

6. A statement made about a population for testing purpose is called?
a) Statistic

b) Hypothesis
c) Level of Significance
d) TestStatistic
Answer b

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

8. 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
Answer a

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

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
b) n
Answer a