MACHINE LEARNING ASSIGNMENT

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 the percentage of the response variable variation that is explained by a linear model. It is always between 0 and 100%. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

In general, the higher the R-squared, the better the model fits your data. However, there are important conditions for this guideline that I discuss elsewhere. Before you can trust the statistical measures for goodness-of-fit, like R-squared, you should check the residual plots for unwanted patterns that indicate biased results.

The residual sum of squares (RSS) measures the difference between your observed data and the model's predictions. It is the portion of variability your regression model does not explain, also known as the model's error. Use RSS to evaluate how well your model fits the data.

In least squares regression, the concept of the sum of squares (SS) is pivotal in quantifying several types of variability relating to a regression model. This statistical method assesses how effectively a dataset aligns with a specific function. The various kinds of SS are essential mathematical tools in identifying the function that most accurately mirrors the data.

Within this framework, the residual sum of squares is one of several types of SS. After fitting the model, it evaluates explicitly the amount of error between the dataset and the regression function. Learn about the other types of sums of squares.

The residual sum of squares formula is the following

$$\sum_{i=1}^{n} (y_i - \hat{y}_i$$

R-squared is the standard goodness-of-fit measure for linear models. Its formula incorporates RSS. Typically, you'll evaluate R 2 rather than the RSS because it avoids some of RSS's limitations, making it much easier to interpret.

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: The residual sum of squares is used to help you decide if a statistical model is a good fit for your data. It measures the overall difference between your data and the values predicted by your estimation model (a "residual" is a measure of the distance from a data point to a regression line). In ANOVA, Total SS is related to the total sum and explained sum with the following formula:
Total SS = Explained SS + Residual Sum of Squares.

What is the Total Sum of Squares:The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.
Total SS = $\Sigma(Yi - \text{mean of } Y)^2$.
Note: Sigma (Σ) is a mathematical term for summation or "adding up." It's telling you to add up all the possible results from the rest of the equation.

Sum of squares is a measure of how a data set varies around a central number (like the mean). You might realize by the phrase that you're summing (*adding up*) squares—but squares of what? You'll sometimes see this formula:

$$y = Y - \overline{Y}$$

Sum of Sq. in ANOVA and Regression**:** As you can probably guess, things get a little more complicated when you're calculating sum of squares in regression analysis or hypothesis testing. It is rarely calculated by hand; instead, software like Excel or SPSS is usually used to calculate the result for you.

For reference, sum of squares in regression uses the equation:

$$\Sigma(y - \overline{y})^2 = \Sigma(\hat{y} - \overline{y})^2 + \Sigma(y - \hat{y})^2$$

And in ANOVA it is calculated with:
The total SS = treatment sum of squares (SST) + SS of the residual error (SSE)

What is the Explained Sum of Squares:The Explained SS tells you how much of the variation in the dependent variable your model explained.
Explained SS = $\Sigma(\text{Y-Hat} - \text{mean of } Y)^2$.

What is the Residual Sum of Squares: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

If all those formulas look confusing, don't worry! It's very, very unusual for you to want to use them. Finding the sum by hand is tedious and time-consuming. It involves a *lot* of subtracting, squaring and summing. Your calculations will be prone to errors, so you're much better off using software like Excel to

do the calculations. You won't even need to know the actual formulas, as Excel works them behind the scenes.

Sum of Squares Within:Within-group variation is reported in ANOVA output as SS(W) or which means Sum of Squares Within groups or SSW: Sum of Squares Within. It is intrinsically linked to between group variation (Sum of Squares between), variance difference caused by how groups interact with each other.

SSW is one component of total sum of squares (the other is between sum of squares). Within sum of squares represents the variation due to individual differences in the score. In other words, it's the variation of individual scores around the group mean

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

Answer: Regularization in machine learning serves as a method to forestall a model from overfitting. Overfitting transpires when a model not only discerns the inherent pattern within the training data but also incorporates the noise, potentially leading to subpar performance on fresh, unobserved data. The employment of regularization aids in mitigating this issue by augmenting a penalty to the loss function employed for Model Training Here are the key points about regularization:

1. Purpose: The primary goal of regularization is to reduce the model's complexity to make it more generalizable to new data, thus improving its performance on unseen datasets.

2. Methods: There are several types of regularization techniques commonly used:

- L1 Regularization (Lasso): This adds a penalty equal to the absolute value of the magnitude of coefficients. This can lead to some coefficients being zero, which means the model ignores the corresponding features. It is useful for feature selection.
- L2 Regularization (Ridge): Adds a penalty equal to the square of the magnitude of coefficients. All coefficients are shrunk by the same factor, and none are eliminated, as in L1.
- Elastic Net: This combination of L1 and L2 regularization controls the model by adding penalties from both L1 and L2, which can be a useful middle ground.

3. Impact on Loss Function: Regularization modifies the loss function by adding a regularization term.

4. Choice of Regularization Parameter: The choice of λ (also known as the regularization parameter) is crucial. It is typically chosen via cross-validation to balance fitting the training data well and keeping the model simple enough to perform well on new data.

4. What is Gini–impurity index?

Answer: The Gini Index is a way of quantifying how messy or clean a dataset is, especially when we use decision trees to classify it. It goes from 0 (cleanest, all data points have the same label) to 1 (messiest, data points are split evenly among all labels).

Think of a dataset that shows how much money people make. A high Gini Index for this data means that there is a huge difference between the rich and the poor, while a low Gini Index means that the income is more balanced.

When we build decision trees, we want to use the Gini Index to find the best feature to split the data at each node. The best feature is the one that reduces the Gini Index the most, meaning that it creates the purest child nodes. This way, we can create a tree that can distinguish different labels based on the features.

What Does a Decision Tree do:A decision tree is a ML algorithm used for both classification and regression tasks. It resembles a tree-like structure with branches and leaves. Each branch represents a decision based on a specific feature of the data, and the leaves represent the predicted outcome.

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

Answer: It is easy to go too deep in the tree, and to fit the parameters that are specific for that training set, rather than to generalize to the whole dataset. This is overfitting. In other words, the more complex the model, the higher the chance that it will overfit. The overfitted model has too many features.

If a decision tree is fully grown, it may lose some generalization capability. This is a phenomenon known as overfitting.

Overfitting can be one problem that describes if your model no longer generalizes well.

Overfitting happens when any learning processing overly optimizes training set error at the cost test error. While it's possible for training and testing to perform equality well in cross-validation, it could be as the result of the data being very close in characteristics, which may not be a huge problem. In the case of decision tree's 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 — ie: overfitting.

6. What is an ensemble technique in machine learning?

Answer: Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods in machine learning usually produce more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods. In the popular Netflix Competition, the winner used an ensemble method to implement a powerful collaborative filtering

algorithm. Another example is KDD 2009 where the winner also used ensembling. You can also find winners who used these methods in Kaggle competitions.
It is important that we understand a few terminologies before we continue with this article. Throughout the article I used the term "model" to describe the output of the algorithm that trained with data. This model is then used for making predictions. This algorithm can be any ML algorithm such as logistic regression, decision tree, etc. These models, when used as inputs of ensemble methods, are called "base models," and the end result is an ensemble model.

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

Answer:

Difference Between Bagging and Boosting: Bagging vs Boosting

	Bagging	**Boosting**
Basic Concept	Combines multiple models trained on different subsets of data.	Train models sequentially, focusing on the error made by the previous model.
Objective	To reduce variance by averaging out individual model error.	Reduces both bias and variance by correcting misclassifications of the previous model.
Data Sampling	Use Bootstrap to create subsets of the data.	Re-weights the data based on the error from the previous model, making the next models focus on misclassified instances.
Model Weight	Each model serves equal weight in the final decision.	Models are weighted based on accuracy, i.e., better-accuracy models will have a higher weight.
Error Handling	Each model has an equal error rate.	It gives more weight to instances with higher error, making subsequent model focus on them.
Overfitting	Less prone to overfitting due to average mechanism.	Generally not prone to overfitting, but it can be if the number of the model or the iteration is high.
Performance	Improves accuracy by reducing variance.	Achieves higher accuracy by reducing both bias and variance.
Common Algorithms	Random Forest	AdaBoost, XGBoost, Gradient Boosting Mechanism
Use Cases	Best for high variance, and low bias models.	Effective when the model needs to be adaptive to errors, suitable for both bias and variance errors.

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

Answer: Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging).

Bagging uses subsampling with replacement to create training samples for the model to learn from. OOB error is the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.

Bootstrap aggregating allows one to define an out-of-bag estimate of the prediction performance improvement by evaluating predictions on those observations that were not used in the building of the next base learner.

9. What is K-fold cross-validation?

Answer: 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.

A technique used in machine learning to assess the performance and generalizability of a model. The basic idea is to partition the dataset into "K" subsets (folds) of approximately equal size. The model is trained K times, each time using K-1 folds for training and the remaining fold for validation. This process is repeated K times, with a different fold used as the validation set in each iteration. K-Fold Cross-validation helps in obtaining a more reliable estimate of a model's performance by reducing the impact of the specific data split on the evaluation. It is particularly useful when the dataset is limited or when there is a concern about the randomness of the data partitioning.

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

Answer: Hyperparameters directly control model structure, function, and performance. 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.

For example, assume you're using the learning rate of the model as a hyperparameter. If the value is too high, the model may converge too quickly with suboptimal results. Whereas if the rate is too low, training takes too long and results may not converge. A good and balanced choice of hyperparameters results in accurate models and excellent model performance.

As previously stated, hyperparameter tuning can be manual or automated. While manual tuning is slow and tedious, a benefit is that you better understand how hyperparameter weightings affect the model. But in most instances, you would normally use one of the well-known hyperparameter learning algorithms.

The process of hyperparameter tuning is iterative, and you try out different combinations of parameters and values. You generally start by defining a target variable such as accuracy as the primary metric, and you intend to maximize or minimize this variable. It's a good idea to use cross-validation techniques, so your model isn't centered on a single portion of your data.

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

Answer: 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. Overfitting: Gradient descent can overfit the training data if the model is too complex or the learning rate is too high.

It determines the step size taken into the gradient direction in backpropagation. Too small learning rate will lead to very slow learning or even inability to learn at all, while too large learning rate can lead to exploding or oscillating performance over the training epochs and to a lower final performance. If the step size is too large, however, we may never converge to a local minimum because we overshoot it every time. Large step size diverges. If we are lucky and the algorithm converges anyway, it still might take more steps than it needed. Large step size converges slowly.

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

Answer: 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.

Logistic regression is simple and easy to implement, but it also has some drawbacks. One of them is that it assumes a linear relationship between the input features and the output. This means that it cannot capture the complexity and non-linearity of the data

If you use a linear predictor function, then it's a "linear model". But you can also use a nonlinear predictor function, in which case it is a "non-linear model". The predictor function is linket to the expected value (μ) by a link function.

13. Differentiate between Adaboost and Gradient Boosting.

Answer: The most significant difference is that gradient boosting minimizes a loss function like MSE or log loss while AdaBoost focuses on instances with high error by adjusting their sample weights adaptively.

Gradient boosting models apply shrinkage to avoid overfitting which AdaBoost does not do. Gradient boosting also performs subsampling of the training instances while AdaBoost uses all instances to train every weak learner.

Overall gradient boosting is more robust to outliers and noise since it equally considers all training instances when optimizing the loss function. AdaBoost is faster but more impacted by dirty data since it fixates on hard examples.

Gradient boosting and AdaBoost are both powerful ensemble machine learning algorithms based on boosting techniques. When comparing their performance, gradient boosting generally achieves higher accuracy.

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

Answer: 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. In general, as we increase the number of tunable parameters in a model, it becomes more flexible, and can better fit a training data set. It is said to have lower error, or bias. However, for more flexible models, there will tend to be greater variance to the model fit each time we take a set of samples to create a new training data set. It is said that there is greater variance in the model's estimated parameters.

The bias–variance dilemma or bias–variance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set

- The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting).

The bias–variance decomposition is a way of analyzing a learning algorithm's expected generalization error with respect to a particular problem as a sum of three terms, the bias, variance, and a quantity called the irreducible error, resulting from noise in the problem itself.

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

Answer: The **Linear** is one of the most straightforward models in machine learning. It is the building block for many complex machine learning algorithms, including deep neural networks. Linear models predict the target variable using a linear function of the input features.

The linear model is one of the most simple models in machine learning. It assumes that the data is linearly separable and tries to learn the weight of each feature. Mathematically, it can be written as $Y=WTXY=WTX$, where X is the feature matrix, Y is the target variable, and W is the learned weight vector. We apply a transformation function or a threshold for the classification problem to convert the continuous-valued variable Y into a discrete category.

RBF: In machine learning, a radial basis function (RBF) is a type of function that is used to approximate complex, nonlinear relationships between variables. RBFs are often used in conjunction with other machine learning algorithms, such as support vector machines (SVMs) and neural networks, to improve their performance.

An RBF is defined as a function that measures the similarity between a point in space and a central point, known as the "center." The function assigns a higher value to points that are closer to the center and a lower value to points that are farther away. The shape of the function is radially symmetrical, meaning that it looks the same in all directions from the center.

RBFs are used in a variety of machine learning applications, including classification, regression, and clustering. They are particularly useful for handling high-dimensional data and for modeling complex, nonlinear relationships between variables.
One common form of an RBF is the Gaussian function, which is defined as follows:
f(x) = exp(-gamma * |x - c|^2)

Polynomial kernels: 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.

the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. In the context of regression analysis

STATISTICS WORKSHEET

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: Expected

2. Chisquare is used to analyse

a) Score b) Rank c) Frequencies **d) All of these**

Answer: All of these

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

a) 4 b) 12 c) **6** d) 8

Answer:6

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

a) Normal distribution b) **Chisqared distribution** c) Gamma distribution d) Poission distribution

Answer: Chisqared distribution

5. Which of the following distributions is Continuous

a) Binomial Distribution b) Hypergeometric Distribution c) **F Distribution** d) Poisson Distribution

Answer: F Distribution

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

a) Statistic b**) Hypothesis** c) Level of Significance d) TestStatistic

Answer: Hypothesis

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: Null Hypothesis

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: Two tailed

9. Alternative Hypothesis is also called as?

a) Composite hypothesis b) **Research Hypothesis** c) Simple Hypothesis d) Null Hypothesis

Answer:Research 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** b) n

Anwer: np