MACHINE LEARNING

ASSIGNMENT - 5

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?

Ans: R-squared is the better measure of goodness-of-fit measure in Regression.R-squared is defined as a number that tells you how well the independent features in a statistical model explain the variation in the dependent feature. It goes from 0 to 1, where 1 indicates a perfect fit of the model to the data. Every feature added to a model increases R-squared and never decreases it. 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. So, it has several limitations like more calculations, give weightage to outliers, its susceptibility to outliers, interpretation difficulties, and inability to address model assumptions. RSS is useful for evaluating the fit of a single model, comparing the fit across multiple models using RSS alone can be tough.

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.

Ans: T**otal sum of squares (TSS)** is the sum of squared differences between the observed *dependent variables* and the overall **mean**.

$$TSS = \Sigma(y_i - \bar{y})^2$$

Explained sum of squares (ESS) is the sum of the differences between the *predicted value* and the **mean** of the *dependent variable*. In other words, it describes how well our line fits the data.

$$ESS = \Sigma(\hat{y}i - yi)2$$

Residual sum of squares (RSS, where residual means remaining or unexplained**)** is the difference between the *observed* and *predicted* values. It is used to 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).

$$RSS = \Sigma(\hat{y}i - y)2$$

Relationship between TSS, ESS, and RSS

Mathematically, **TSS = RSS + ESS**

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

Ans: Regularization is a technique used to reduce errors by fitting the model appropriately on the given training set and avoiding overfitting. It helps to reduce the variance of the model without a substantial increase in the bias. If there is variance in the model that means the model won't fit well for dataset different from training data. The tuning parameter controls this bias and variance trade off. It improves model performance by preventing excessive weightage to outliers or irrelevant features. It makes models stable across different subsets of the data. It reduces the sensitivity of model outputs to minor changes in the training set. It promotes consistent model performance across different datasets.

4. What is Gini–impurity index?

Ans: Gini -Impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset. Its value lies between 0 and 1, where 0 being no impurity and 1 denoting maximum. The node for which the Gini Impurity is least is selected as the root node to split. In simple words it tells us how impure is feature. Lower the impurity better the feature

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

Ans: Yes, unregularized decision trees are prone to overfitting because decision tree doesn't use regularization to fight against overfitting. Instead, it employs tree pruning decision trees by their very nature are prone to overfitting, especially when they are deep. Overfitting occurs when a model captures variations in the training data that do not represent the underlying data distribution. A tree that is too complex might achieve a perfect accuracy score on the training data but perform poorly on new data. Such a tree has low bias but high variance, and its predictions can be unstable. Regularization techniques add constraints to the learning algorithm, reducing its freedom and, hence, its capacity to overfit. Regularization is achieved by controlling the tree's depth and complexity. The **regularization techniques of decision trees include:** Maximum Depth (max_depth), Minimum Samples Split (min_samples_split), Minimum Samples Leaf (min_samples_leaf). The key challenge in regularization is striking a balance between underfitting and overfitting. Under-regularized trees might be too deep and capture noise (overfit), while over-regularized trees might be too shallow and miss important patterns (underfit). Regularization parameters should ideally be determined using cross-validation or a separate validation dataset, ensuring that the settings generalize well to new data. Selecting the right hyperparameters (tree depth and leaf size) also requires experimentation, e.g. doing cross-validation with a hyperparameter matrix.

6. What is an ensemble technique in machine learning?

Ans: Ensemble technique in machine learning is a technique that combines the predictions from multiple individual models to obtain a better predictive performance rather than any single model. This can lead to improved performance and generalization. Most commonly used ensembles techniques are Bagging which is used to generate Random Forest algorithms and Boosting which is used to generate algorithms such as AdaBoost, XGBoost etc. In other words, an ensemble technique takes multiple small models and combine their predictions to obtain a more powerful predictive power.

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

Bagging	Boosting
Training Data Subsets are drawn randomly with replacement from the entire training dataset.	Each new subset contains the components that were misclassified by previous methods.
Bagging attempts to tackle the overfitting issue.	Boosting tries to reduce bias
Every model receives equal weightage.	Models are weighted by their performance.
Its objective to decrease variance, not bias.	Its objective to decrease bias, not variance.
Every model is built independently.	New models are affected by the performance of the previously developed model.

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

Ans: OOB (out-of-bag) errors are an estimate of the performance of a random forest classifier or regressor on unseen data. It can be obtained using the out of bag score attribute. It is computed using the samples that were not included in the training of the individual trees. This is different from the error computed using the usual training and validation sets, which are used to tune the hyperparameters of the random forest. It can be useful for evaluating the performance of the random forest on unseen data. It is not always a reliable estimate of the generalization error of the model, but it can provide a useful indication of how well the model is performing.

9.What is K-fold cross-validation?

Ans: K-fold cross-validation is a technique for evaluating predictive models. In this, the dataset is divided into k subsets. The model is trained and evaluated k times, using a different fold as the validation set each time. The idea is simple, divide the whole dataset into 'k' sets preferably of equal sizes. Then the first set is selected as the test set and the 'k-1' sets are used to train the data. Error is calculated for this particular dataset. Then the steps are repeated i.e. the second set is selected as the test data and the remaining 'k-1' sets are used as the training data. Again, the error is calculated. Similarly, the process is continues for 'K' times. Performance metrics from each fold are averaged to estimate the model's generalization performance. This method aids in model assessment, selection, and hyperparameter tuning, providing a more reliable measure of a model's effectiveness.

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

Ans: Hyperparameter tuning is the process of selecting the optimal values for machine learning model hyperparameters. Hyperparameters are settings that control the learning process of the model, such as the learning rate. The goal of hyperparameter tuning is to find the values that lead to the best performance on a given task. It is done because they are often used to tune the performance of a model and have a significant impact on the model's accuracy, generalization, and other metrics. Hyperparameters tuning is crucial as they control the overall behaviour of a machine learning model. Its example are GridSearchCV, RandomisedSearchCV.

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

Ans: Gradient Descent is too sensitive to the learning rate. If it is too big, the algorithm may bypass the local minimum and overshoot. Gradient descent algorithms multiply the gradient by a scalar known as the learning rate to determine the next point. For example, if the gradient magnitude is 2.5 and the learning rate is 0.01, then the gradient descent algorithm will pick the next point 0.025 away from the previous point. If it is too large, model will converge as our pointer will shoot and we'll not be able to get to minima. A higher learning rate will lead to overshooting of slope.

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

Ans: No, we cannot use Logistic Regression for classification of Non-Linear Data because Logistic regression is one such type of classification model which is used to classify the dependent variable into two or more categories. It calculates the probability that a given value belongs to a specific class. It acts as a binary classifier. Logistic regression is an exercise in predicting discrete outcomes from a

continuous and/or categorical set of observations. Each observation is independent and the probability *p* that an observation belongs to the class of the features describing that observation.

13.Differentiate between Adaboost and Gradient Boosting.

Ans:

Adaboost	Gradient Boosting
1. AdaBoost is the first designed boosting algorithm with a particular loss function.	1. Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem.
2. Less Flexible	2. More Flexible
3. Algorithm is less robust to outliers.	3. Algorithm is more robust to outliers.
4. Each classifier has different weights assigned to the final prediction based on its performance.	4. All classifiers are weighed equally and their predictive capacity is restricted with learning rate to increase accuracy.
5. It gives weights to both classifiers and observations thus capturing maximum variance within data.	5. It builds trees on previous classifier's residuals thus capturing variance in data.

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

Ans: Bias-variance trade-off is all about finding the right balance between simplicity and complexity in a machine learning model. A model that is too simple has high bias and low variance but lacks the capacity to learn from the data. In contrast, a highly complex model has low bias and high variance which causes model to overfit and perform poorly on new data. The goal of machine learning is to strike a balance between bias and variance, so the model generalizes well to unseen data. This process is typically achieved through model selection, hyperparameter tuning or sometimes by regularization to prevent overfitting.

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

Ans: **SVM (Support Vector Machine) algorithms** use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. So, Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. Basically, It returns the inner product between two points in a standard feature dimension. Support Vector Machine (SVM) is a relatively simple **Supervised Machine Learning Algorithm** used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data.

Different SVM algorithms use different types of kernel functions. These functions can be different types. For example, *linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.*

Linear SVM: When the data is perfectly linearly separable only then we can use Linear SVM. Perfectly linearly separable means that the data points can be classified into 2 classes by using a single straight line.

Radial Basis Function (RBF) Kernel: RBF kernels are the most generalized form of kernelization and is one of the most widely used kernels due to its similarity to the Gaussian distribution. The RBF kernel function for two points X_1 and X_2 computes the similarity or how close they are to each other. This kernel can be mathematically represented as follows:

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

Polynomial kernel A polynomial kernel is a kind of SVM kernel that uses a polynomial function to map the data into a higher-dimensional space. It does this by taking the dot product of the data points in the original space and the polynomial function in the new space. In this, the data is mapped into a higher-dimensional space using a polynomial function. The dot product of the data points in the original space and the polynomial function in the new space is then taken. The polynomial kernel is often used in SVM classification problems where the data is not linearly separable. By mapping the data into a higher-dimensional space, the polynomial kernel can sometimes find a hyperplane that separates the classes.The polynomial kernel has a number of parameters that can be tuned to improve its performance, including the degree of the polynomial and the coefficient of the polynomial.