ASSIGNMENT

MACHINE LEARNING

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?

Answer: - R-Squared is the ratio of the sum of squares regression (SSR) and the sum of squares total (SST). Sum of Squares Regression (SSR) represents the total variation of all the predicted values found on the regression line or plane from the mean value of all the values of response variables. The sum of squares total (SST) represents the total variation of actual values from the mean value of all the values of response variables. R-squared value is used to measure the goodness of fit or best-fit line**.** The greater the value of R-Squared, the better is the regression model as most of the variation of actual values from the mean value get explained by the 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.

Answer: - Total sum of squares: TSS represents the total sum of squares. It is the squared values of the dependent variable to the sample mean. In other words, the total sum of squares measures the variation in a sample.

Total Sum of squares TSS= $\sum_{i=1}^{n}(y_i - \bar{y})^2$

ESS (Explained sum of square) is a statistical quantity used in modelling of a process. ESS gives an estimate of how well a model explains the observed data for the process.

It tells how much of the variation between observed data and predicted data is being explained by the model proposed. Mathematically, it is the sum of the squares of the difference between the predicted data and mean data.

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

The residual sum of squares (RSS) 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.

$$\mathrm{RSS} = \sum\nolimits^{n}_{i=1} (y^i - f(x_i))^2$$

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

Answer:- 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. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

4. What is Gini–impurity index?

Answer: - Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree.

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

Answer: - Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.

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 usually produces 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.

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

Answer:- Bagging is an acronym for 'Bootstrap Aggregation' and is used to decrease the variance in the prediction model. Bagging is a parallel method that fits different, considered learners independently from each other, making it possible to train them simultaneously. Bagging generates additional data for training from the dataset. This is achieved by random sampling with replacement from the original dataset. Sampling with replacement may repeat some observations in each new training data set. Every element in Bagging is equally probable for appearing in a new dataset.

Boosting is a sequential ensemble method that iteratively adjusts the weight of observation as per the last classification. If an observation is incorrectly classified, it increases the weight of that observation. The term 'Boosting' in a layman language, refers to algorithms that convert a weak learner to a stronger one. It decreases the bias error and builds strong predictive models. Data points mis predicted in each iteration are spotted, and their weights are increased. The Boosting algorithm allocates weights to each resulting model during training. A learner with good training data prediction results will be assigned a higher weight. When evaluating a new learner, Boosting keeps track of learner's errors.

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

Answer:- The *out-of-bag* (OOB) error is the average error for each Zi calculated using predictions from the trees that do not contain Zi in their respective bootstrap sample. This allows the Random Forest Classifier to be fit and validated whilst being trained.

9. What is K-fold cross-validation?

Answer: - K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data. K refers to the number of groups the data sample is split into.
It has a mean validation accuracy of **93.85%** and a mean validation f1 score of 91.69%.

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

Answer: - What is hyper parameter tuning in machine learning and why it is done?

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

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

Answer: - A learning rate that is too large can **cause the model to converge too quickly to a suboptimal solution**.

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

Answer: - No, logistic regression is a linear classifier. It is a type of generalized linear model, which predicts variables with various types of probability distributions by fitting a linear predictor function to some sort of arbitrary transformation of the expected value of the variable.

13. Differentiate between Adaboost and Gradient Boosting.

Answer: -

Ada Boost	Gradient Boost
AdaBoost is the first designed boosting algorithm with a particular loss function. AdaBoost minimises loss function related to any classification error and is best used with weak learners.	Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. Gradient Boosting algorithm is more robust to outliers than AdaBoost.

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

Answer: - The bias–variance trade-off is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

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

Answer: **- Linear Kernel** is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set.

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.