ASSIGNMENT - 5

MACHINE LEARNING

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: r2 is better as compared to residual sum of square. because: 1. it calculates the variance of independent variable that is explained by dependent variable(x). This shows how the x is creating the variance in the label or target.
2.it is easy to interpret and can be standardised.
3.it takes into account both total sum of square (TSS) and RSS (residual sum of square)

for eg: if r square is 0.8, that shows the 80% of the variance is explained by predictors.
value of r square ranges from 0-1, 1 shows the 100% goodness of fit of the 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.

Ans: TSS: shows the variance in the response / target variable.

ESS: shows the sum of difference between predicted variable from the mean. that Is variance from the predicted value.

RSS: refers to the unexplained variance resulting from sum of difference between predicted value and actual value.

relation between the three: TSS = RSS + ESS.

3 . What is the need of regularization in machine learning?
Ans: regularisation helps to prevent the overfitting of data in a model.
this helps in reducing the collinearity of variables resulting in better prediction.
2.aslo helps in measuring the coefficients better, helps in better fit of model

3.prevention of learning the noise and random fluctuation in data causing the model give wrong prediction.

Thus, it is necessary to use regularisation of data.

4. What is Gini–impurity index?
Ans: in decision tree algorithms, Gini impurity measures the degree of impurity or uncertainty in a set of labels/classes. It ranges from 0 to 0.5, where 0 indicates that all elements belong to a single class, and 0.5 indicates an equal distribution across all classes.

5. 5. Are unregularized decision-trees prone to overfitting? If yes, why?
Ans : yes , undervaluation may cause overfitting of data . due to :

1.memorisation : learning each data points without understanding the pattern or relationship does not meet std of goodness of fit of model , it fails to capture specific pattern and characteristics of data.

2.less fitting : the decision tree model unregularized data learns the training data but poorly perform on new data given , thus making it less robust in real-life scenario.
3.high variance : model can capture noises and outliers , and can be sensitive to small fluctuation .

6. What is an ensemble technique in machine learning?
Ans :

Ensemble technique : this refers to method of combining different models to improve the actual prediction. with this model , lower strength of one model can be compensated by another model. There are different ensemble method : 1.baggig , 2.boosting , 3.stacking , 4.voting etc.

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

Ans : bagging : types of ensemble technique where different subsets is trained on same base algorithm , and later averaging the prediction.eg : random forest. 2. Boosting : refers to the building of sequence of model where each subsequent model does reduce the weakness of the previous model. eg : Adaboost , GBM (gradient boosting machine)

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

Ans : it refers to average prediction error (classification error or MAE) across all trees in random forest based on their respective out of bag data.it provides unbiased estimate of the model's performance without the need for cross-validation or a separate validation set.

9. What is K-fold cross-validation?
Ans : K-fold cross-validation is a popular technique used in machine learning for evaluating the performance of a model, especially when the dataset is limited in size. It helps in assessing how well a model generalizes to new data by partitioning the original dataset into K equal-sized subsets (or folds).K-fold cross-validation provides a more accurate estimate of model performance because it uses multiple train-test splits of the data.

10. What is hyper parameter tuning in machine learning and why it is done?
Ans : hyperparameter tuning in machine learning refers to the process of finding the optimal set of hyperparameters for a learning algorithm. Hyperparameters are parameters that are set before the learning process begins, and they control aspects of the learning process itself. Unlike model parameters, which are learned during training, hyperparameters are typically set based on heuristics, prior knowledge, or grid/random search.

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

Ans : Having a large learning rate in machine learning can lead to several issues.
Firstly, it can cause the optimization process to diverge, meaning the model parameters fail to converge to a minimum or oscillate wildly. This instability prevents effective learning and model refinement.

Secondly, a large learning rate can result in overshooting the minimum of the loss function during optimization, leading to erratic updates and difficulty in achieving a stable model configuration.

Thirdly , models trained with a large learning rate may struggle to generalize well to new data, as they may be overfitted to the training set. Finally, a large learning rate complicates the process of hyperparameter tuning, as it becomes harder to find an optimal combination that maximizes model performance.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
Ans : no logistic regression will not be suitable to predict non-linear data because for classification purpose , decision boundary are not linear.it will fail to capture the non-linearity effectively , might lead to underfitting and ultimately it cannot inherently learn complex non-linear decision boundaries without additional modifications or transformations of the input features.

14. What is bias-variance trade off in machine learning?
Ans : Bias-variance trade-off is a critical concept in machine learning that highlights the need to find a suitable balance between model simplicity (bias) and flexibility (variance) to achieve good generalization performance on unseen data.it helps to generalise the model and help to reduce the complexity of the model.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.
Ans : Polynomial kernel : The polynomial kernel calculates the similarity between two vectors after transforming them into a higher-dimensional space using a polynomial function.

The RBF : The RBF kernel measures the similarity between two data points based on the Euclidean distance in the feature space.

linear kernel : It computes the inner product between two vectors , Suitable for linearly separable data or when the number of features is very large.