Batch- DS2311

Name- Akash yadav

ASSESMENT-2

MACHINE LEARNING

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 and RSS both measures of goodness of fit model in regression , but they capture different aspect of model performance.

If we talk about R squared – it measure variations in the dependent variable (that is dependent on independent variables). It indicate how well the model fits the data(with a range from 0-1). Higher the value of R-squared better the data fits to model.

If we talk about RSS- it measures the total sum of squared difference between the actual values of dependent variables and the predicted values . It represents the unexplained variations in the data. Lower the values of RSS better fits the data .

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- tells how much variation there is in the dependent variable.

ESS- tells you how much of the variation in the dependent variable your model explained.

RSS- tells you how much of the dependent variable's variation your model did not explain.

Eq- **TSS = ESS + RSS**

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

Ans. Whenever we are training the machine learning model for data analyses, at that time the model may go under overfitting or underfitting condition. To avoid this , we use regularization in machine learning model to properly fit a model for test data set.

4. What is Gini–impurity index?

Ans. Gini Impurity tells us the probability of misclassifying an observation. lower the Gini the better the split. In other words the lower the likelihood of misclassification. The Gini index has a maximum impurity is 0.5 and maximum purity is 0.

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

Ans. Yes, decision tree prone to overfitting because the decision tree keep growing and become more complex until they get perfectly classify the training data. Due to this the model captures the noise in data and thus perform poorly on new data. The regularization techniques such pruning sets a minimum number of samples required to split a node or limits the maximum number of depth of the tree can leads to mitigate overfitting in decision tree.

6. What is an ensemble technique in machine learning?

Ans. Ensemble technique in machine learning are the technique that help in creating multiple models. Which then combine together to produce improved results. It produce more accurate solutions than a single model.

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

Ans. Bagging and boosting techniques are two different ensemble techniques that use multiple models to reduce error and optimize the model.

If we talk about bagging technique, it combines the models trained on different subset of data, whereas the boosting trains the model sequentially.

These two model technique focus on error made by previous model (so that error can be reduced).

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

Ans. Out-of-bag (OOB) error is a method of measuring the prediction error of random forests. It is the average error for each calculated data using predictions from the trees.

9. What is K-fold cross-validation?

Ans. 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?

Ans. When we are training machine learning models, each dataset or model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through model. This is called hyperparameters tuning.

This is done in ML model just to maximizes the model’s performance and minimize the errors.

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

Ans. 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.

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

Ans. No we cannot use logistics regression for classification of non-linear data. In non linear data the classes cannot be separated by linear boundaries. If we apply Logistic Regression for classification of Non-Linear Data then if will cause the errors and the model will not run perfectly.

13. Differentiate between Adaboost and Gradient Boosting

Ans. Adaboost is computed with a specific loss function and become more rigid when comes to few iterations.

But in Gradient boosting, it assists in finding the proper solution to all the additional iteration of model problem.

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

Ans. In machine learning , whenever we trying to minimize one component of error i.e bias, the other component i.e variance increases and vice versa. The process in which we can balance the bias and variance to create the effective and accurate model is called bias-variance trade off.

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

Ans. Linear- An SVM with a linear kernel learns a linear decision boundary in the original feature space.

RBF- In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms.

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.

STATISTICS WORKSHEET

1. Using a goodness of fit,we can assess whether a set of obtained frequencies differ from a set of frequencies

Ans. Expected

2. Chisquare is used to analyse

Ans. Frequencies.

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

Ans. 6

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

Ans. Chisqared distribution

5. Which of the following distributions is Continuous

Ans . F Distribution

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

Ans Statistic

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

Ans . Null Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?

Ans Two tailed

9. Alternative Hypothesis is also called as?

Ans 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

Ans. np