STATISTICS WORKSHEET-5

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
2. Chisquare is used to analyse
   a) Score
   b) Rank
   c) Frequencies
   d) 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
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
5. Which of the following distributions is Continuous
   a) Binomial Distribution
   b) Hypergeometric Distribution
   c) F Distribution
   d) Poisson Distribution
6. A statement made about a population for testing purpose is called?
   a) Statistic
   b) Hypothesis
   c) Level of Significance
   d) TestStatistic
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
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
9. Alternative Hypothesis is also called as?
   a) Composite hypothesis
   b) Research Hypothesis
   c) Simple Hypothesis
   d) Null Hypothesis

FLIP ROBO

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

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?
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.
3. What is the need of regularization in machine learning?
4. What is Gini–impurity index?
5. Are unregularized decision-trees prone to over fitting? If yes, why?
6. What is an ensemble technique in machine learning?
7. What is the difference between Bagging and Boosting techniques?
8. What is out-of-bag error in random forests?
9. What is K-fold cross-validation?
10. What is hyper parameter tuning in machine learning and why it is done?
11. What issues can occur if we have a large learning rate in Gradient Descent?
12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
13. Differentiate between Adaboost and Gradient Boosting.
14. What is bias-variance trade off in machine learning?
15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Answers:

Ans.1 Both R-squared and RSS are measures to assess the goodness of fit of a regression model. R-squared is used when we want to find how well the model fits the data where the values are between 0 and 1. It is exclusively used to compare different models.

Whereas RSS is sued to calculate the absolute goodness of fit and focuses on the measure of the prediction errors. This is used if our goal is to reduce the prediction errors.

Ans. 2 TSS (Total Sum of Squares) is the measure of how a data set varies around a central number.

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The Explained Sum of Squares tells us how much of the variation in the dependent variable our model explains.

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

where $\hat{y}_i$ the value estimated by the regression line

The Residual Sum of Squares tells us how much of the variation in the dependent variable our model did not explain.

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

Where:

y_i = the i^{th} value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

The relationship between TSS, ESS and RSS is as follows:
TSS = ESS + RSS

Ans. 3 Regularization is done in machine learning so that we can reduce the adjusted loss function and prevent under or over fitting of the machine learning models.

Ans. 4 Gini impurity index is the measure of the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

Ans. 6 Ensemble methods in machine learning are the techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly.

Ans. 7 Bagging and Boosting are 2 types of ensemble learning techniques. Bagging is used when the model has a high variance and hence after bagging, the resulting model has a less variance. Whereas boosting is done when the model has a high bias and hence boosting gives a lower bias finally.

Ans. 8 Out-of-bag errors are an estimate of the performance of a random forest classifier or regressor on unseen data. This is computed using the samples that were not used in training of the trees. For Random Forests, this error is calculated by finding the average of each of the individual trees.

Ans. 9 K-fold cross-validation is a method used for evaluating predictive models. The dataset is divided into k subsets, hence it is called k-fold cross validation method. The model is trained and evaluated k times, using a different subset as the validation set each time. Performance metrics from each subset are averaged to estimate the model's generalization performance. This method helps in

model assessment, selection, and hyper-parameter tuning, providing a more reliable measure of a model's effectiveness.

Ans. 10 Hyperparameter tunig is a method of finding a set of optimal hyper parameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyper parameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

Ans. 11 If the learning rate in gradient is high, it can have a high divergence, i.e. the model may undergo over fitting.

Ans. 12 We cannot use logistic regression for the classification of non-linear data as logistic regression demands the assumption that there should exist a linearity between the dependent and independent variables.

Ans. 13 Both Adaboost and gradient boost are machine learning ensemble techniques used for creating a stronger model. Here are some differences between the 2 techniques:

The weights of the sample are adjusted at each iteration.	No reweighting takes place during iterations.
The next tree is built using the same training data but using the newly weighted training samples until the desired performance.	Gradient descent is used to lit new learners to the previous ones so that the loss functions is minimized.
$Prediction = sign(\sum_{m=1}^{M} \alpha_m * F_m(x))$ where, $F_m(x)$ is the output of each model and α_m are the weight computed by the boosting algorithm, m is the number of iteratio	$Prediction = \hat{y} + \eta * \sum_{m=2}^{M} \hat{r}_{m-1}$ where, $\hat{y}$ is the prediction from the first tree, η is the learnin rate, $\hat{r}_i$ is the prediction of residuals, m is the no. of iteratio

Ans. 14 Getting a balance between the accuracy and ability of the model to make predictions in the testing data is called the bias-variance tradeoff.

FLIP ROBO