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

R-squared give better measure of goodness of fit model in regular regression.

Why?

The reason is because R-squared value indicate a higher variability in fit model and explain the variation of the target variable.

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 square is squared differences between the observed dependent variables and the overall mean.

Equation

$$\sum_{i=1}^{n} (\mathrm{yi} - (\overline{\mathrm{y}})_2$$

(ii) Explained sum square – is sum of the squared of the deviations of the predicted values from the mean value of response variable in a standard regression model, for example

$Y = a + b_1x_1 + b_2x_2 + \ldots e$

(iii) Residual sum of squares (RSS) is also known as sum squared estimate of errors (SSE) – measures the discrepancy between the data and an estimation model such as linear regression.

Equation

RSS $= \sum_{i=1}^{n} (\mathrm{yi} - \mathrm{f(xi)})_2$

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

ANSWER

The reason why regularisation is being done in machine learning is to properly fit a model onto our test set.

4. What is Gini–impurity index?

ANSWER

It gives the probability of misclassifying an observation. Then the lower the GINI the better the split and the lower the likelihood of misclassification.

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

ANSWER

No

This is because there will be no pint of perfect classification.

6. What is an ensemble technique in machine learning?

ANSWER

Ensemble method in machine learning usually produce more accurate solution than single model. Or Ensemble method in machine learning technique that

enhance accuracy and resilience in forecasting by merging predictions from multiple models.

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

ANSWER

Bagging is a method that merges the same type of prediction while Boosting merge different type of prediction.

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

ANSWER

Out of bag error in random forest is a method of measuring the prediction error of random forests boosted decision trees, and other machine models and utilising boot strap aggregating. Or is a way of validating the random forest model.

9. What is K-fold cross-validation?

ANSWER

K-fold cross validation is a technique for evaluating predictive models. In this case split your data into training and test set.

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

ANSWER

Hyperparameter turning is essential part of controlling the behaviour of a machine learning model. It consists of finding a set of optimal Hyperparameter values for learning algorithm while applying this optimized algorithm to any data set. The combination of hyperparameters maximises the model performance, minimising a predefined loss function to produce better result with fewer errors.

Why is this being done?

To allow data scientist to tweak model performance for optimal results.

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

ANSWER

when the learning rate is high, the approach will be unstable and the coefficients will explode and get overflow error. Or the gradient descent can suffer from divergence. This means weight increases exponentially, resulting in explosive gradient which can cause problem of instabilities and over high loss values.

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

ANSWER

No

REASON IS BECAUSE

Logistic regression is generalised as linear model because the output data is always dependent on the input data.

13. Differentiate between Adaboost and Gradient Boosting.

ANSWER

Adaboost is the first boosting ensemble model automatically adjusts it parameter to the database actual performance while Gradient boost is a robust machine learning algorithm which are made of Gradient descent and boost.

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

ANSWER

This is a tradeoff between model to minimise bias and variance which is referred to the best solution for selection of value of regularisation of constant.

Bias is the orediction of value by machine learning and variance predict a given data point.

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

ANSWER

A linear kernel is a type of kernel function used in machine learning, including in SVMs (Support Vector Machines). It is the simplest and most commonly used kernel function, and it defines the dot product between the input vectors in the original feature space.

In machine learning, the radial basis function kernel, or 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.

WORKSHEET

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 ANSWER=D

2. Chisquare is used to analyse
a) Score
b) Rank
c) Frequencies ANSWER=C
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 ANSWER=C
d) 8

4. Which of these distributions is used for a goodness of fit testing?
a) Normal distribution
b) Chisqared distribution ANSWER=B
c) Gamma distribution
d) Poission distribution

5. Which of the following distributions is Continuous
a) Binomial Distribution
b) Hypergeometric Distribution
c) F Distribution ANSWER=C
d) Poisson Distribution

6. A statement made about a population for testing purpose is called?
a) Statistic
b) Hypothesis ANSWER=B
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 ANSWER=A
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 ANSWER=A
b) One tailed
c) Three tailed
d) Zero tailed

9. Alternative Hypothesis is also called as?
a) Composite hypothesis
b) Research Hypothesis ANSWER=B
c) Simple Hypothesis
d) Null 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
a) np ANSWER=A
b) n