WORKSHEET

STATISTICS WORKSHEET-5

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. Chi square 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) Chi sqared 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

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

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:

Typically, however, a smaller or lower value for the RSS is ideal in any model since it means there's less variation in the data set. In other words, the lower the sum of squared

residuals, the better the regression model is at explaining 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.

ANSWER:

The ESS is then: where the value estimated by the regression line. In same cases: total sum of squares(TSS) = Explained sum of squares(ESS) + residual sum of square(RSS).

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 reduced the errors in it.

4. What is Gini–impurity index?

ANSWER:

More precisely, the Gini impurity of a dataset is a number between 0-0.5, which indicates the likehood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

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

ANSWER:

Yes, decision trees are prone to overfitting, especially where a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of event 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 is a machine learning technique that combines several base models in order to producer one optimal predictive model. To better understand this definition lets take a step back into ultimate goal of ml the model building.

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

ANSWER:

Very roughly we can say that bagging will mainly focus at getting an ensemble model with less variance then its components where as bosting will mainly try to produce strong models less biased then there components (even if variances can also be reduced).

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

ANSWER:

The out – of – bag (oob) error is the average error for each calculated using prediction from the trees that do not contain

their respective boot step 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 where the data set is split into a K number of foldes and is used to evaluate the model's ability where given new data K refers to the no of groups the data sample split into.

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

ANSWER:

Hyper parameter tuning consist of finding a set of optical hyper parameter values for a learning algorithm while applying this optimized algorithm to any data set. The combination of hyper parameters maximize the model performance, minimizing predfined loss function to produced better results with few errors.

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

ANSWER:

A learning rate that is to large can cause the model to converge too quickly to a suboptimal solution where as a leaning rate that is too small can cause the process to get stuck.

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

ANSWER:

It can only be used to predict discrete functions. Hence the dependent variables of logistic regression is bound to the discrete number set. It is very fast at classifying unknow records. Known linear problem can't be solved with logistic regression because it has a linear decision surface.

13. Differentiate between Adaboost and Gradient Boosting.

ANSWER:

Adaboost is a first designed boosting algorithm with a particular loss function. On the other hand , Gradient Boosting is a generic algorithm that assists in searching the approximate solution to the additive modelling problem. This makes Gradient Boosting more flexible than Adaboosting.

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

ANSWER:

In machine learning, 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 – It is useful when dealing with large sparse data vectors. It is often used in text categorization.

RBF – It is general – purpose kernel; used when there is no prior knowledge about the data.

Polynomial – It is popular in image processing.