MACHINE LEARNING ANSWERS

1 - A residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model. Ideally, the sum of squared residuals should be a smaller or lower value than the sum of squares from the regression model's inputs

2 - The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.

The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model.

A residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.

3 - regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting. Regularization applies to objective functions in ill-posed optimization problems.

4 - Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. But what is actually meant by 'impurity'? If all the elements belong to a single class, then it can be called pure.

5 – Yes, This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.

6 - Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

7 - Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance,

not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

8 - The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample.

9 - Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into

10 - In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

11 - When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error

12 - Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries.

13 - AdaBoost is the 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 solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

14 - Bias is the simplifying assumptions made by the model to make the target function easier to approximate.

15 - SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form.

The most used type of kernel function is RBF. Because it has localized and finite response along the entire x-axis

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 ANSWERS

1 - d) Expected

2 - d) All of these

3 - b) 12

4 - b) Chisqared distribution

5 - c) F Distribution

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

7 - a) Null Hypothesis

8 - a) Two tailed

9 - b) Research Hypothesis

10 - a) np

SQL ANSWERS

1 - Select * from movie;

2 – select title from movie where runtime = max(runtime);

3 - select title from movie where revenue = max(revenue);

4 - select title from movie where revenue = max(revenue) OR budget=max(budget);

5 - SELECT person_name, gender, character_name, cast_order,title

FROM movie, movie_cast, gender,person

WHERE movie.mId = movie_cast.mId AND movie_cast.rId = person.rId AND movie_cast.dId = gender.dId

ORDER BY title,person_name, gender, character_name, cast_order;

6 – select country_name from country where countryid=max(count(countryid));

7 – select genre_id,genre_name from genre;

8 – select language_name,title from language,movie where language.lid=movie.lid;

9 – select title,cast_order,job from movie,movie_crew,movie_cast where movie.mid=moviecrew.mid=movie_cast.mid;

10 – select title from movie group by popularity order by desc;

11 - SELECT title,revenue FROM movie ORDER BY revenue DESC LIMIT 2,1

12 – select title from movie where movie_status=’rumoured’

13 – select title from movie where production_company=’United states of America’ and revenue=(select max(revenue)from movie)

14 – select movie_id,production_company from movie

15 – select title from movie orderby budget desc