MACHINE LEARNING

1. 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. TSS is the sum of square of difference of each data point from the mean value of all the values of target variable
   The explained sum of squares measures how much variation there is in the modelled values and this is compared to the total sum of squares which measures how much variation there is in the observed data, and to the residual sum of squares which measures the variation in the error between the observed data and modelled values.
   RSS It is a measure of the discrepancy between the data and an estimation model, such as a linear regression.
3. Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
4. The Gini impurity index is a measure of impurity or diversity used in decision tree algorithms for classification problems. It measures the probability of misclassifying a randomly chosen element in the dataset if it were labeled randomly according to the distribution of labels in the subset.
5. Yes, unregularized decision trees are prone to overfitting. Overfitting occurs on decision trees because unregularized learn the pattern in training data. In decision trees, this can happen when the tree becomes too complex, and each branch is optimized to perfectly fit the training data, including outliers. This can lead to a high variance in the model and poor performance on new data.
6. Ensemble methods can help to improve the accuracy, robustness, and generalization of machine learning models. They are particularly effective when the individual models are diverse and complementary, which helps to reduce bias and variance and improve the overall performance of the model.
7. Boosting method involves training models sequentially and correcting errors of previous models whereas bagging method trains multiple mode on different subset of training data with replacement.
8. The out-of-bag error provides an estimate of the performance of the random forest on new, unseen data without the need for a separate validation dataset. It can be used to compare the performance of different random forests or to tune the hyperparameters of the model, such as the number of trees or the maximum depth of each tree
9. K-fold cross-validation is a popular technique in machine learning for evaluating the performance of a model on a dataset. The basic idea behind K-fold cross-validation is to divide the original dataset into K equally sized folds, and then use K-1 folds for training the model and the remaining fold for testing the model.
10. Hyperparameter tuning in machine learning is the process of selecting the optimal values for the hyperparameters of a machine learning model, with the goal of improving its performance on new, unseen data. Hyperparameters are parameters of a model that are not learned during training, but are set prior to training and can significantly affect the performance of the model.

11. Using a large learning rate in gradient descent can lead to several issues that can prevent the algorithm from finding the optimal weights for a machine learning model. Some of the issues are:

    Divergence: With a large learning rate, the algorithm can overshoot the minimum of the cost function and start oscillating or diverging. This can lead to unstable and unpredictable behavior, and the algorithm may fail to converge to the optimal weights.

    Slow convergence: A learning rate that is too high can cause the algorithm to take large steps towards the minimum of the cost function, but it may also lead to overshooting the minimum and bouncing back and forth, thereby slowing down the convergence process.
    Overshooting the minimum: When the learning rate is too high, the algorithm can overshoot the minimum of the cost function, and instead of converging to the minimum, the algorithm will oscillate back and forth around it, never reaching the optimal solution.

    Failure to converge: In some cases, using a learning rate that is too high can cause the algorithm to fail to converge to the optimal weights. This can happen when the algorithm keeps overshooting the minimum of the cost function or oscillating around it, without making any significant progress towards the optimal solution.
12. Logistic regression is not suitable for classification of non-linear data as it can only model linear decision boundaries.
    In cases where the data is not linearly separable, using logistic regression can result in poor performance and inaccurate predictions. This is because logistic regression assumes a linear relationship between the features and the response variable, and cannot model complex, non-linear relationships between them.

13. Adaboost is an iterative algorithm that builds a series of weak models (typically decision trees) on different subsets of the data, and combines them into a single strong model. In contrast, Gradient Boosting builds a series of models (again, usually decision trees), each of which tries to correct the errors of the previous model.

    Adaboost can be sensitive to noisy data, as the weights assigned to each training example can be affected by outliers or mislabeled data. In contrast, Gradient Boosting is less sensitive to noisy data, as the model is trained to correct the errors made by the previous model.

14. the bias-variance trade-off is a fundamental concept in machine learning that describes the relationship between the complexity of a model and its ability to generalize to new data. Balancing bias and variance is a critical aspect of model selection and training, and finding the right balance requires careful consideration of the data, the model, and the training method.

SQL

1) Select * FROM movie

2) Select title

FROM movie

WHERE runtime= (Select MAX(runtime) FROM movie)

3) SELECT title

FROM movie

WHERE revenue=(SELECT MAX(revenue) FROM movie)

4) SELECT title

FROM movie

WHERE revenue=(SELECT MAX(revenue/budget) FROM movie)

5) SELECT movie.title, cast.person_name, cast.gender, cast.character_name, cast.cast_order

FROM movies

INNER JOIN cast ON movie.movie.id = cast.movie_id

6) SELECT country.country_name, COUNT(*) AS num_movies

FROM movies

INNER JOIN production ON movie .movie_id = production.movie_id

INNER JOIN country ON production.country_id = country.country_id

GROUP BY country.country_name

ORDER BY num_movies DESC

LIMIT 1

7) SELECT * FROM genre

8)SELECT language.language_name, COUNT(movie_language.movie_id) AS num_movies

FROM language

LEFT JOIN movie_language ON language.language_id=movie_language.language_id

```
GROUP BY language.language_name
9) SELECT
  movie.movie_title,
  COUNT(DISTINCT movie_crew.person_id) AS num_crew_members,
  COUNT(DISTINCT movie_cast.person_id) AS num_cast_members
FROM
  movie
  LEFT JOIN movie_crew ON movie.movie_id = movie_cast.movie_id
  LEFT JOIN movie_cast ON movie.movie_id = movie_cast.movie_id
GROUP BY
  movie.movie_id
10) SELECT title
FROM movie
ORDER BY popularity DESC
LIMIT 10;
11) SELECT
  movie_title,
  revenue
FROM
  movie
ORDER BY
  revenue DESC
LIMIT 1 OFFSET 2
12) SELECT title
FROM movie
```

WHERE movie status= ‘rumored’

13) SELECT movie.movie_title

FROM

movie

JOIN production_country ON movie.movie_id = production_country.movie_id

JOIN country ON production_country.country_id = country.country_id

WHERE

country.country_name = 'United States of America'

ORDER BY

movie.revenue DESC

LIMIT 1

14) SELECT movie_company.movie_id,company.company_name

FROM movie_company

JOIN company ON movie_company.company_id = company.company_id

15)

Statistics

1)d

2)c

3)c

4)b

5)c

6)b

7)a

8)a

9)b

10)a