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?**

   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.
   Linear regression is a measurement that helps determine the strength of the relationship between a dependent variable and one or more other factors, known as independent or explanatory variables.

   The Formula for the Residual Sum of Squares (RSS)
   $RSS = \sum^{n}_{i=1} (y^{i} - f(x_i))^{2}$

   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

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.**

   The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.
   Total SS = $\Sigma(Yi - \text{mean of } Y)^2$.
   Note: Sigma (Σ) is a mathematical term for summation or "adding up." It's telling you to add up all the possible results from the rest of the equation.
Sum of squares is a measure of how a data set varies around a central number (like the mean). You might realize by the phrase that you're summing (*adding up*) squares—but squares of what? You'll sometimes see this formula:

   $y=Y-Y$

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

   This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
   A simple relation for linear regression looks like this. Here Y represents the learned relation and β represents the coefficient estimates for different variables or predictors*(X)*.

$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$RSS = \Sigma yi - B0 - \Sigma B_j x_{ij}$

4. **What is Gini–impurity index?**

   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.

   $Gini = 1 - \Sigma P_i$

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

   Over-fitting is the phenomenon in which the learning system tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the untrained data.
   In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus it ends up with branches with strict rules of sparse data. Thus this effects the accuracy when predicting samples that are not part of the training set.

6. **What is an ensemble technique in machine learning?**
   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. **What is the difference between Bagging and Boosting techniques?**
   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. **What is out-of-bag error in random forests?**
   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. **What is K-fold cross-validation?**
   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. **What is hyper parameter tuning in machine learning and why it is done?**
    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. **What issues can occur if we have a large learning rate in Gradient Descent?**

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

12. **Can we use Logistic Regression for classification of Non-Linear Data? If not, why?**
    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. That can be remedied however if we happen to have a better idea as to the shape of the decision boundary.
13. **Differentiate between Adaboost and Gradient Boosting**
    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. **What is bias-variance trade off in machine learning?**
    Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.
15. **Give short description each of Linear, RBF, Polynomial kernels used in SVM.**
    Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set. ... Training a SVM with a Linear Kernel is Faster than with any other Kernel.
    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.

SQL

1. **Write SQL query to show all the data in the Movie table.**
   SELECT * FROM movie;
2. **Write SQL query to show the title of the longest runtime movie.**
   SELECT title FROM movie
   ORDER BY runtime DESC
   LIMIT 1;
3. **Write SQL query to show the highest revenue generating movie title**
   SELECT title FROM movie
   ORDER BY revenue DESC
   LIMIT 1;
4. **Write SQL query to show the movie title with maximum value of revenue/budget.**
   SELECT title, revenue/budget FROM movie
   ORDER BY revenue/budget DESC
   LIMIT 1;
5. **Write a SQL query to show the movie title and its cast details like name of the person, gender, character name, cast order.**
   SELECT title, person_name, gender, character_name, cast_order
   FROM movie INNER JOIN movie_cast
   ON movie .movie_id = movie_cast .movie_id
   INNER JOIN gender
   ON movie_cast.gender_id = gender.gender_id
   INNER JOIN person
   ON movie_cast.person_id = person.person_id;
6. **Write a SQL query to show the country name where maximum number of movies has been produced, along with the number of movies produced.**
   SELECT country_name, COUNT(movie_id) AS mov_no
   FROM production_country AS pc INNER JOIN country AS con
   ON pc.country_id = con.country_id
   GROUP BY country_name
   ORDER BY mov_no DESC
   LIMIT 1;
7. **Write a SQL query to show all the genre_id in one column and genre_name in second column.**
   SELECT genre_id, genre_name FROM genre;
8. **Write a SQL query to show name of all the languages in one column and number of movies in that particular column in another column.**
   SELECT language_name, COUNT(movie_id) AS mov_no
   FROM movie_languages AS mov_lan INNER JOIN language AS lan
   ON mov_lan.language_id = lan.language_id
   GROUP BY lan.language_id;
9. **Write a SQL query to show movie name in first column, no. of crew members in second column and number of cast members in third column.**

```
SELECT title, COUNT(person_id) AS cast_no
FROM movie AS mov INNER JOIN movie_cast AS mov_cast
ON mov.movie_id = mov_cast.movie_id
GROUP BY mov_cast.movie_id;
```

10. **Write a SQL query to list top 10 movies title according to popularity column in decreasing order.**

```
SELECT title FROM movie
ORDER BY popularity DESC
LIMIT 10;
```

11. **Write a SQL query to show the name of the 3rd most revenue generating movie and its revenue.**

```
SELECT title FROM movie
ORDER BY revenue DESC
LIMIT 3;
```

12. **Write a SQL query to show the names of all the movies which have "rumoured" movie status.**

```
SELECT title FROM movie
WHERE movie_status = "rumoured";
```

13. **Write a SQL query to show the name of the "United States of America" produced movie which generated maximum revenue.**

```
SELECT title, revenue
FROM movie AS mov INNER JOIN production_country AS pro_con
ON mov.movie_id = pro_con.movie_id
INNER JOIN country AS con
ON con.country_id = pro_con.country_id
WHERE country_name = "United States Of America"
ORDER BY revenue DESC
LIMIT 1;
```

14. **Write a SQL query to print the movie_id in one column and name of the production company in the second column for all the movies.**

```
SELECT movie_id, company_name
FROM movie_company AS mov_com INNER JOIN production_company AS prod_com
ON mov_com.company_id = prod_com.company_id;
```

15. **Write a SQL query to show the title of top 20 movies arranged in decreasing order of their budget.**

```
SELECT title, budget FROM movie
ORDER BY budget DESC
LIMIT 20;
```

STATISTICS -5

1. D
2. C
3. C
4. D
5. C
6. B
7. A
8. A
9. B
10. A