Machine learning

1) The RSS is just the absolute amount of explained variation, the R squared is the (RSS/SST), i.e. the absolute amount of variation as a proportion of total variation

R-squared is a goodness-of-**fit measure** for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.

It takes into account the strength of the relationship between the model and the dependent variable. Its convenience is measured on a scale of 0 – 100%.

2) Explained sum of squares (ESS): Also known as the explained variation, the ESS is the portion of total variation that measures how well the regression equation explains the relationship between X and Y.

You compute the ESS with the formula

$$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

Residual sum of squares (RSS): This expression is also known as unexplained variation and is the portion of total variation that measures discrepancies (errors) between the actual values of Y and those estimated by the regression equation.

You compute the RSS with the formula

$$RSS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Total sum of squares (TSS):

The sum of RSS and ESS equals TSS.

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

R2 is the ratio of explained sum of squares (ESS) to total sum of squares (TSS):

$$R^2 = \frac{ESS}{TSS}$$

You can also use this formula:

$$R^2 = 1 - \frac{RSS}{TSS}$$

3) Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

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) No,My understanding is that different regularization technique is adding a term to cost functions such as cross-entropy to reduce accuracy/overfitting to training data.

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)

1. Bagging technique can be an effective approach to reduce the variance of a model, to prevent over-fitting and to increase the accuracy of unstable models.
2. On the other hand, Boosting enables us to implement a strong model by combining a number of weak models together.
3. In contrast to bagging, samples drawn from the training dataset are not replaced back into the training set during the boosting exercise.
4. If you analyse the decision boundaries, known as stumps, that are computed by the Adaptive Boosting algorithm when compared with the Decision trees, you will note that the decision boundaries computed by AdaBoost can be very sophisticated.
5. Although this can help us implement a strong predictive model, the ensemble learning increases the computational complexity compared to individual classifiers.

8) Random Forest, is a powerful ensemble technique for machine learning, but most people tend to skip the concept of OOB_Score while learning about the

algorithm and hence fail to understand the complete importance of Random forest as an ensemble method.

This blog will walk you through the OOB_Score concept with the help of examples.

9) K-Fold CV is where a given data set is split into a ***K*** number of sections/folds where each fold is used as a testing set at some point. Lets take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

10) **Hyper Parameters** are those parameters of a machine learning algorithm that controls the learning process and efficiency of the machine learning algorithm in its training phase and of course can be set and optimized manually. It basically determines how the algorithm is going to take the different learning approaches in the different steps of its learning process. The wrong choice of Hyper Parameters can make your machine learning model vulnerable to overfitting or underfitting, resulting in poor performance. Therefore determining hyperparameter setting wisely can certainly make your machine learning model better.

11)When the learning rate is too large,gradient descent nadvertently increase rather than decrease the training error. […] when the learning rate is too small,training is not only slower, but may become permanently stuck with a high training error.

12) The short answer is: Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters. Or in other words, the output cannot depend on the product (or quotient, etc.) of its parameters!

13) Both are boosting algorithms which means that they convert a set of weak learners into a single strong learner. They both initialize a strong learner (usually a decision tree) and iteratively create a weak learner that is added to the strong learner. They differ on how they create the weak during the iterative process.

At each iteration, adaptive boosting changes the sample distribution by modifying the weights attached to each of the instances. It increases the weights of the wrongly predicted instances and decreases the ones of the correctly predicted instances. The weak learner thus focuses more on the difficult instances. After being trained, the weak learner is added to the strong one accorging to his performance (so-called alpha weight). The higher it performs, the more it contributes to the strong learner.

On the other hand, gradient boosting doesn't modify the sample distribution. Instead of training on a newly sample distribution, the weak learner trains on remaing errors (so-called pseudo-residuals) of the strong learner. It is another way to give more importance to the difficult instances. At each iteration, the pseudo-residuals are computed and a weak learner is fitted to these pseudo-residuals. Then, the contribution of the weak learner (so-called multiplier) to the strong one isn't computed according to his performance on the newly distribution sample but using a gradient descent optimisation process. The computed contribution is the one minimizing the overall error of the strong learner.

14)

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) 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. ... Training a SVM with a Linear kernel is Faster than with any other kernal

In general, the polynomial kernel is defined as ;

$$K(X_1, X_2) = (a + X_1^T X_2)^b$$

b = degree of kernel & a = constant term.

in the polynomial kernel, we simply calculate the dot product by increasing the power of the kernel.

Gaussian RBF(Radial Basis Function) is another popular Kernel method used in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point. Gaussian Kernel is of the following format;

$$K(X_1, X_2) = exponent(-\gamma \|X_1 - X_2\|^2)$$

sql

```
1)select * FROM  movie;

2)select max(runtime) FROM movie;

3)select max(revenue) FROM movie;

4)select tiltle from movie where max(revenue) FROM movie;

5)select
movie.title,movie_cast.cast_order,movie_cast.gender_id,movie_cast.caste_id
FROM movie,movie_cast;

6) select country_id ,title

FROM country, movie

Group by max(titles),titles as max_no_movies_produced,no of movies;

7) select genere_id,genere_name

genere_id as genere_id,

Genere_name as genere_name

From genere;

8) select language_name,movie_id
```

```
Where language_name=movie_id;

Language name as languages, movie_id as no of movies

From language,movie_languages;

9)  select title,person_id,person_id

title as movie_name

Person_id as crewmember,

Person_id as cast_member_cast

From movie,movie_cast,movie_crew;

10) select popularity

 Where popularity=title

From movie

And title.movie>=10

Order by title desc;

11)select revenue,title

Where title=revenue >3

From movie;

12)select movie_status

Where movie_status

From movie_status=rumoured

From movie;

13)select title,revenue

Where title= United States of America
```

And title=revenue

From movie;

14)select movie_id,companyname

Where move_idi=company_name

Movie_id as move_id

Company_name as production_name

From movie,production_company;

15)select title

Where title>=20

From movie;

Statistics

1)d

2)d

3)c

4)d

5)c

6)b

7)a

8)a

9)b

10)a