Machine learning

1)d

2)b

3)c

4)a

5)a

6)a,d

7)b

8)d

9)d

10) The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

The adjusted R2 will penalize you for adding independent variables (K in the equation) that do not fit the model. Why? In regression analysis it can be tempting to add more variables to the data as you think of them. Some of those variables will be significant, but you can't be sure that significance is just by chance. The adjusted R2 will compensate for this by that penalizing you for those extra variables.

While values are usually positive, they can be negative as well. This could happen if your R2 is zero; After the adjustment, the value can dip below zero. This usually indicates that your model is a poor fit for your data. Other problems with your model can also cause sub-zero values, such as not putting a constant term in your model.

11) Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 where as lasso go with l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it's the sum of the absolute values of the coefficients. It's a shrinkage towards

zero using an absolute value (l1 penalty) rather than a sum of squares(l2 penalty).

As we know that ridge regression can't zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficients of collinear variables. Here it helps to select the variable(s) out of given n variables while performing lasso regression

12) A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are usually calculated by software, as part of regression analysis. You'll see a VIF column as part of the output. VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. "i" is the predictor you're looking at (e.g. x1 or x2):

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.
A rule of thumb for interpreting the variance inflation factor:

1 = not correlated.

Between 1 and 5 = moderately correlated.

Greater than 5 = highly correlated.

Exactly how large a VIF has to be before it causes issues is a subject of debate. What is known is that the more your VIF increases, the less reliable your regression results are going to be. In general, a VIF above 10 indicates high

correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above.

Sometimes a high VIF is no cause for concern at all. For example, you can get a high VIF by including products or powers from other variables in your regression, like x and x2. If you have high VIFs for dummy variables representing nominal variables with three or more categories, those are usually not a problem.

13) Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful.

The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature scaling helps to weigh all the features equally.

Formally, If a feature in the dataset is big in scale compared to others then in algorithms where Euclidean distance is measured this big scaled feature becomes dominating and needs to be normalized.

14) The sum of squares due to error (SSE)

R-square

Adjusted R-square

Root mean squared error (RMSE)

15) Accuracy: Overall, how often is the classifier correct?

(TP+TN)/total = (1200+1000)/2500 = 0.88

True Positive Rate: When it's actually yes, how often does it predict yes?

TP/actual yes = 1200/1450 = 0.827

also known as "Sensitivity" or "Recall"

True Negative Rate: When it's actually no, how often does it predict no?

TN/actual no = 1000/1050 = 0.95

equivalent to 1 minus False Positive Rate

also known as "Specificity"

Sql

1)c,d

2)a,c

3)b

4)c

5)b

6)b

7)a

8)c

9)d

10)a

11) Denormalization is a database optimization technique where we add redundant data in the database to get rid of the complex join operations. This is done to speed up database access speed. Denormalization is done after normalization for improving the performance of the database. The data from one table is included in another table to reduce the number of joins in the query and hence helps in speeding up the performance.

12) A database cursor can be thought of as a pointer to a specific row within a query result. The pointer can be moved from one row to the next. Depending on the type of cursor, you may be even able to move it to the previous row.

Think of it this way: a SQL result is like a bag, you get to hold a whole bunch of rows at once, but not any of them individually; whereas, a cursor is like a pair of tweezers. With it, you can reach into the bag and grab a row, and then move onto the next.

13) Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table. If there is any violation between the constraint and the data action, the action is aborted.

14)Constraints can be column level or table level. Column level constraints apply to a column, and table level constraints apply to the whole table.

15) Auto-increment allows a unique number to be generated automatically when a new record is inserted into a table.

Often this is the primary key field that we would like to be created automatically every time a new record is inserted.

Statistics

1)d

2)c

3)a

4)c

5)a

6)a

7)c

8)c

9)b

10) **Histograms** and box plots are graphical representations for the frequency of numeric data values. ... **Histograms** are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing **between** several data sets.

11) Good **metrics** are important to your company growth and objectives. Your key **metrics** should always be closely tied to your primary objective. ...

Good **metrics** can be improved. Good **metrics** measure progress, which means there needs to be room for improvement. ...

Good **metrics** inspire action.

12)steps to calculate statistical signiance

Create a null hypothesis.

Create an alternative hypothesis.

Determine the **significance** level.

Decide on the type of test you'll use.

Perform a power analysis to find out your sample size.

Calculate the standard deviation.

pUse the standard error formula.

13)examples: Allocation of wealth among individuals

Values of oil reserves among oil fields (many small ones, a small number of large ones)

14) The median is preferable to the mean when there is an outlier or two. Take this example:

35, 42, 81, 27, 30, 38

81 is a clear outlier, and will mess up the mean and make the mean not truly representative of the whole data set. Most of the data is in the range of 27-42, but that 81 is way outside the range.

15) Let X1, X2, ..., Xn have a joint density function f(X1, X2, ..., Xn|θ). Given X1 = x1, X2 = x2, ..., Xn = xn is observed, the function of θ defined by:

L(θ) = L(θ|x1, x2, ..., xn) = f(x1, x2, ..., xn|θ) is the likelihood function.

• The likelihood function is not a probability density function.

• It is an important component of both frequentist and Bayesian analyses

• It measures the support provided by the data for each possible value of the parameter. If we compare the likelihood function at two parameter points and find that L(θ1|x) > L(θ2|x) then the sample we actually observed is more likely to have occurred if θ = θ1 than if θ = θ2. This can be interpreted as θ1 is a more plausible value for θ than θ2.