Q1. The value of correlation coefficient will always be:

Ans. (C) between -1 and 1 (indicating perfect correlation)

Q2. Which of the following cannot be used for dimensionality reduction?

Ans. (D) Ridge Regularisation

Q3. Which of the following is not a kernel in Support Vector Machines?

Ans. (C) Hyperplane

Q4. . Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Ans. (A) Logistic Regression

Q5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be?

Ans. (C) old coefficient of ‘X’ ÷ 2.205

Q6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

Ans. (C) Decreases (n\_ estimators - Number of weak learners to train iteratively.)

Q7. Which of the following is not an advantage of using random forest instead of decision trees?

Ans. (C) Random Forests are easy to interpret

**Q8. Which of the following are correct about Principal Components?**

A) Principal Components are calculated using supervised learning techniques   
B) Principal Components are calculated using unsupervised learning techniques   
C) Principal Components are linear combinations of Linear Variables.   
D) All of the above

Ans. B, C

**Q9. Which of the following are applications of clustering?**   
A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index   
B) Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts.   
C) Identifying spam or ham emails   
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Ans. A, B, C, D

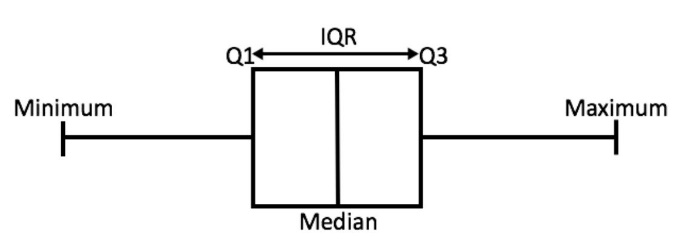
**Q10. Which of the following is(are) hyper parameters of a decision tree?**   
A) max\_depth   
B) max\_features   
C) n\_estimators   
D) min\_samples\_leaf

Ans. A, B, D

**Q11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.**

**Ans.**

**Outlier** - is a data point which differs significantly from other observations.



**IQR Method – (**Inter-Quartile Range): To detect the outliers using IQR method

* The median is the median (or centre point), also called second quartile, of the data (resulting from the fact that the data is ordered).
* Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.
* Q3 is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3.

We define range here, let’s call it decision range, and any data point lying outside this range is considered as outlier and is accordingly dealt with -

**IQR = Q3 - Q1**

**Lower Bound: (Q1 - 1.5 \* IQR)**

**Upper Bound: (Q3 + 1.5 \* IQR)**

Some time when features are very-very important and can’t take risk then we can use extreme boundary (Extreme outliers) also –

**Lower Bound: (Q1 - 3 \* IQR)**

**Upper Bound: (Q3 + 3 \* IQR)**

Note: - Any data point less than the *Lower Bound or*more than the *Upper Bound*is considered as an outlier.

**Q12. What is the primary difference between bagging and boosting algorithms?**

**Ans.**

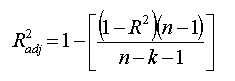
|  |  |
| --- | --- |
| **Bagging** | **Boosting** |
| 1. Simplest way of combining predictions that belong to the same type. | A way of combining predictions that belong to the different types. |
| 1. Aim to decrease variance, not bias. | Aim to decrease bias, not variance. |
| 1. Each model receives equal weight. | Models are weighted according to their performance |
| 1. Each model is built independently. | New models are influenced by performance of previously built models. |
| 1. Different training data subsets are randomly drawn with replacement from the entire training dataset. | Every new subsets contains the elements that were misclassified by previous models |
| 1. Bagging tries to solve over-fitting problem. | Boosting tries to reduce bias. |
| 1. If the classifier is unstable (high variance), then apply bagging. | If the classifier is stable and simple (high bias) the apply boosting. |
|  |  |

**Q13. What is adjusted R2 in logistic regression. How is it calculated?**

**Ans.**

R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model.

* If you add more and more **useless** [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)to a model, adjusted r-squared will decrease.
* If you add more **useful** variables, adjusted r-squared will increase.
* Adjusted R2 will always be less than or equal to R2.



here:

* N is the number of points in your data sample.
* K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

**Adjusted R2**:

* Both R2 and the adjusted R2 give you an idea of how many data points fall within the line of the [regression equation](https://www.statisticshowto.com/what-is-a-regression-equation/).
* There is one main difference between R2 and the adjusted R2: R2 assumes that every single variable explains the variation in the [dependent variable](https://www.statisticshowto.com/dependent-variable-definition/).
* The adjusted R2 tells you the percentage of variation explained by only the [independent variables](https://www.statisticshowto.com/independent-variable-definition/) that actually affect the dependent variable.

**How Adjusted R2 Penalizes:**

* The adjusted R2 will penalize you for adding independent variables (K in the equation) that do not fit the model.
* Why? In [regression analysis](https://www.statisticshowto.com/probability-and-statistics/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](https://www.statisticshowto.com/what-is-statistical-significance/), 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.

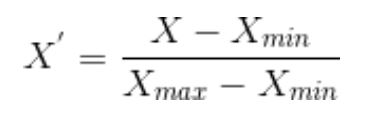
**Problems with R2 that are corrected with an adjusted R2:**

* R2 increases with every predictor added to a model. As R2 always increases and never decreases, it can appear to be a better fit with the more terms you add to the model. This can be completely misleading.
* Similarly, if your model has too many terms and too many [high-order polynomials](https://calculushowto.com/polynomial-function-degrees/) you can run into the problem of over-fitting the data. When you over-fit data, a misleadingly high R2 value can lead to misleading projections.

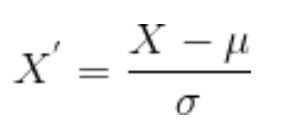
**Q14. What is the difference between standardisation and normalisation?**

**Ans.**

**Normalization:** Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

****

**Standardization:** Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.



**Difference between Normalize or Standardize:**

* **Normalization** is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
* **Standardization** can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

*Note: - You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.*

**Q15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.**

**Ans.**

**Cross Validation**: CV in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:

**Advantages:**

* **Reduces Overfitting:** In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.  
    
  **Note:** Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.
* **Hyperparameter Tuning:** Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

**Disadvantages:**

* **Increases Training Time:** Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.   
    
  For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.
* **Needs Expensive Computation:** Cross Validation is computationally very expensive in terms of processing power required.