

MACHINE LEARNING

Answer Keys

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

11. An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal.

For eg: While analyzing salaries of all the people in India, salary of Mukesh Ambani or Ratan Tata might be an outlier.

Inter Quartile Range (IQR):

$IQR = Q3 - Q1$

Where, $Q1 = 25^{th}$ % ile of the data

$Q2 = 50^{th}$ % ile (a.k.a. median)

$Q3 = 75^{th}$ % ile of the data.

Upper bound = $Q3 + 1.5 * Q3$

Lower Bound = $Q1 - 1.5 * Q1$

Any data point lying above than upper bound and lower than lower bound is considered as an outlier.

12. **Bagging:** Bagging is also known as bootstrap aggregating sits on top of the majority voting principle. The samples are bootstrapped each time when the model is trained. When the samples are chosen, they are used to train and validate the predictions. The samples are then replaced back into the training set. The samples are selected at random. This technique is known as bagging. To sum up, base classifiers such as decision trees are fitted on random subsets of the original training set. Subsequently, the individual predictions are aggregated (voting or averaging etc.). The final results are then used as predictions. It reduces the variance of a black box estimator. Due to this the chances of overfitting is ruled out.

Boosting: The concept of Adaptive Boost revolves around correcting previous classifier mistakes. Each classifier gets trained on the sample set and learns to predict. The misclassification errors are then fed into the next classifier in the chain and are used to correct the mistakes until the final model predicts accurate results. When a weak-classifier misclassifies a training sample, the algorithm then uses these very samples to improve the performance of the ensemble.

13. Adjusted R^2 and R^2 both represent that how well the model fits the data points. But adjusted R^2 penalizes the model for using more features. In case we increase the number of features in training data the R^2 will increase but adjusted R^2 will only increase if the new feature adds value to our model. Due to this reason adjusted R^2 is considered as a better evaluation metric than R^2 . Adjusted R^2 is always less than or equal to R^2 . The formula to calculate adjusted R^2 is as follows:

$$R^2_{adj} = \left[1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

Where, n = number of data points in the dataset

K = Number of features in the dataset excluding the constant term

MACHINE LEARNING

14. In Normalization a dataset is scaled in such a way that all the data points lie between 0 and 1. Normalization is often called min-max scaling. Formula for Normalization is as follows:

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Whereas, In Standardization a dataset is scaled in such a way that the mean of data points becomes 0 and standard deviation is 1. The transformed data may be positive as well as negative in standardization. The formula for standardization is as follows:

$$x_{new} = \frac{x_i - \bar{x}}{s}$$

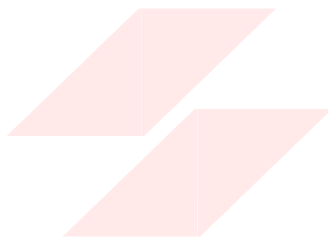
Where, x_i = i^{th} data point

\bar{x} = sample mean

s = sample standard deviation

15. Cross validation is a technique to fit a model on data set. In cross validation the data set is divided into 'k' number of sets where 'k-1' sets are used for training and 1 set is used as validation set. And this is done for all the set one by one and the final score of model is taken as average score of all the 'k' number of fits.

Advantage of using Cross validation is that, there is no need of separate validation data, cross validation reduces chances of overfitting and gives a more generic model. Cross validation has a disadvantage that it takes more time to fit the model over a large dataset and the model built is more complex than the basic model.



FLIP ROBO