MACHINE LEARNING WORKSHEET 4

1Q. C

2Q. D

3Q. C

4Q. B

5Q. D

6Q. B

7Q. C

8Q. D

9Q. A,C

10Q. C,D

11Q. Outliers:
The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.
However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.
Significance of outliers:
Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results.
Most machine learning algorithms do not work well in the presence of outlier. So it is desirable to detect and remove outliers.
Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.
IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3

called first, second and third quartiles are the values which separate the 4 equal parts.

Q1 represents the 25th percentile of the data.
Q2 represents the 50th percentile of the data.
Q3 represents the 75th percentile of the data.

If a dataset has 2n / 2n+1 data points, then
Q1 = median of the dataset.
Q2 = median of n smallest data points.
Q3 = median of n highest data points.
IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.
Example:
Assume the data 6, 2, 1, 5, 4, 3, 50. If these values represent the number of chapatis eaten in lunch, then 50 is clearly an outlier.

12Q. 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.
In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.
Models are built independently in Bagging. New models are affected by a previously built model’s performance in Boosting.
In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. In Boosting, every new subset comprises the elements that were misclassified by previous models.
Bagging is usually applied where the classifier is unstable and has a high variance. Boosting is usually applied where the classifier is stable and simple and has high bias.

13Q. R-squared (R^2) is an important statistical measure which is a regression model that represents the proportion of the difference or variance in statistical terms for a dependent variable which can be explained by an independent variable or variables. In short, it determines how well data will fit the regression model.

R Squared Formula

For the calculation of R squared, you need to determine the Correlation coefficient, and then you need to square the result.

R Squared Formula = r^2

Where r the correlation coefficient can be calculated per below:

$$r = n\ (\Sigma xy) - \Sigma x\ \Sigma y\ /\ \surd\ [n^{*}\ (\Sigma x^2 - (\Sigma x)^2)]\ *\ [n^{*}\ (\Sigma y^2 - (\Sigma y)^2)]$$

14Q. Normalization
Standardization

Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

15Q. In Machine Learning, Cross-validation is a statistical method of evaluating generalization performance that is more stable and thorough than using a division of dataset into a training and test set. In this article, I’ll walk you through what cross-validation is and how to use it for machine learning using the Python programming language.

Cross Validation 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 of Cross Validation :

1. 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.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation :

1. 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.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.