Mrs. Swati Amit Motugade

MACHINE LEARNING ASSIGNMENT 4

Q.1. C

Q.2. B

Q.3. C

Q.4. B

Q.5. C

Q.6. C

Q.7. C

Q.8. B & C

Q.9. A, B, D

Q.10. A, B, D

Q.11.What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Ans. An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value**.** When you graph an outlier, it will appear not to fit the pattern of the graph.

Or in other words an outlier is a value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.

For example, in the scores 25,29,3,32,85,33,27,28 both 3 and 85 are "outliers".

Some outliers are due to printing mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening.

InterQuartile Range (IQR):

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.

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

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.

We can find the outlier of this dataset using python.

For this we need to perform the following steps:

Step 1: Import necessary libraries.

```
import numpy as np
import seaborn as sns
```

Step 2: Take the data and sort it in ascending order.

```
data = [6, 2, 3, 4, 5, 1, 50]
sort_data = np.sort(data)
sort_data
```

Step 3: Calculate Q1, Q2, Q3 and IQR.

```
Q1 = np.percentile(data, 25)
Q2 = np.percentile(data, 50)
Q3 = np.percentile(data, 75)

print('25th percentile of the given data is, ', Q1)
print('50th percentile of the given data is, ', Q2)
print('75th percentile of the given data is, ', Q3)

IQR = Q3 - Q1
```

```
print('Interquartile range is', IQR)
```

Output:

```
25th percentile of the given data is 2.5

50th percentile of the given data is 4.0

75th percentile of the given data is 5.5

Interquartile range is 3.0
```

Step 4: Find the lower and upper limits as Q1 – 1.5 IQR and Q3 + 1.5 IQR, respectively.

```
low_lim = Q1 - 1.5 * IQR
up_lim = Q3 + 1.5 * IQR
print('low_limit is', low_lim)
print('up_limit is', up_lim)
```

Output:

```
low_limit is -2.0
up_limit is 10.0
```

Step 5: Data points greater than the upper limit or less than the lower limit are outliers

```
outlier =[]
for x in data:
    if ((x> up_lim) or (x<low_lim)):
        outlier.append(x)
print(' outlier in the dataset is', outlier)
```

Output:

```
outlier in the dataset is [50]
```

Q. 12. What is the primary difference between bagging and boosting algorithms?

Ans.

1. **Bagging**: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.
2. **Boosting**: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

Bagging

Bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It decreases the and helps to avoid overfitting. It is usually applied to decision tree methods. Bagging is a special case of the model averaging approach.

Boosting

Boosting is an ensemble modelling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

The basic difference between Bagging and Boosting are

Sr.No.	Bagging	Boosting
1	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3	Each model receives equal weight.	Models are weighted according to their performance.
4	Each model is built independently.	New models are influenced by the performance of previously built models.
5	Different training data subsets are selected using row sampling with replacement and random	Every new subset contains the elements that were misclassified by previous models.

	sampling methods from the entire training dataset.	
6	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8	In this, base classifiers are trained parallelly.	In this, base classifiers are trained sequentially.
9	Example: The Random Forest model uses Bagging.	Example: The AdaBoost uses Boosting techniques

Q. 13. What is adjusted R2 in linear regression. How is it calculated? Ans.

Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.

R^2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects are included in the model. Adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might decrease if a specific effect does not improve the model.

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1.

Adjusted R^2 is always less than or equal to R^2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0 indicates a model that has no predictive value. In the real world, adjusted R^2 lies between these values.

Adjusted R squared is calculated by **dividing the residual mean square error by the total mean square error** (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

$$Adj.R^2 = 1 - \frac{SS_{res}/df_{error}}{SS_{total}/df_{total}}$$

Q. 14. What is the difference between standardization and normalization?

Ans.

1.Standardisation: Data standardization is a process in which the data is restructured in a uniform format. In statistics, standardization compares the variables by putting all the variables on the same scale. It is done by transforming the features by subtracting from the mean and dividing by the standard deviation. This process is also known as the Z-score. Mathematically, Standardisation is denoted as:

$$X_{stand} = \frac{X - mean(X)}{SD(X)}$$

2. Normalization: The process of arranging the data in a database is known as Normalization. It is a scaling technique used to reduce redundancy in which the values are shifted and scaled in a range of 0 and 1. Normalization is used to remove the unwanted characteristics from the dataset, and it is useful when there are no outliers as it cannot handle them. Mathematically, Normalisation is denoted as:

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The basic difference points between Standardization and Normalization

Standardization	Normalization
Scaling is done by mean and standard deviation.	Scaling is done by the highest and the lowest values.
It is applied when we verify zero mean and unit standard deviation.	It is applied when the features are of separate scales.
Not bounded	Scales range from 0 to 1
Less affected by outliers	Affected by outliers
It is used when the data is Gaussian or normally distributed	It is applied when we are not sure about the data distribution
It is also known as Z-Score	It is also known as Scaling Normalization

Q. 15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Ans. Cross-validation is a statistical method used to estimate the performance or accuracy of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds or partitions of the data, run the analysis on each fold, and then average the overall error estimate.

Advantage:

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.

Disadvantage:

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.