**MACHINE LEARNING WORKSHEET-1 SOLUTIONS**

**ANS-1)** C (between -1 and 1)

**ANS-2)** D (Ridge Regularisation)

**ANS-3)** C (hyperplane)

**ANS-4)** A (Logistic Regression)

**ANS-5)** A

**ANS-6)** B (increases)

**ANS-7)** C (Random Forests are easy to interpret)

**ANS-8)** B (Principal Components are calculated using unsupervised learning techniques),

C (Principal Components are linear combinations of Linear Variables)

**ANS-9**

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.

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels

**ANS-10)** A (max\_depth), B (max\_features), D (min\_samples\_leaf)

**ANS-11)** **Outlier:** An outlier, in mathematics, statistics is a specific data point that falls outside the range of probability for a data set. In other words, the outlier is distinct from other surrounding data points in a particular way. During Data pre-processing and Data Cleaning outliers should be removed from the data in order to get good results, **IQR (Inter Quartile Range)** method proved to be a good method for this outlier removal task.

**IQR (Inter Quartile Range):** In [descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics), the **IQR** also called the **mid-spread**, **middle 50%**, or **H‑spread**, is a measure of [statistical dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion), being equal to the difference between 75th and 25th [percentiles](https://en.wikipedia.org/wiki/Percentiles), or between upper and lower [quartiles](https://en.wikipedia.org/wiki/Quartile).

**Formula:** IQR = *Q*3 − *Q*1

**Example:** Suppose you have the following set of data: 1, 3, 4, 6, 7, 7, 8, 8, 10, 12, and 17.

Divide sample into 2 so lower half is 1, 3, 4, 6, 7 and upper half is 8, 8, 10, 12, 17 and now finding median for both half we will get Q1 and Q3 respectively.

**Statistical information**: median = 7, first quartile (Q1) = 4[third quartile](https://www.thoughtco.com/what-are-first-and-third-quartiles-3126235) (Q3) = 10.

**IQR***= Q*3 – *Q*1 = 10 – 4 = 6

Now,

lower= *Q*1 -(1.5 \* IQR)=4-(1.5\*6)= -4 ; upper= *Q*3 +(1.5 \* IQR)=10+(1.5\*6)=19

Condition for a number (say n) in a dataset to be outlier: any number **“n < lower or n > upper”** is an outlier.

Hence there is no outlier in our data as every number lies between -4 and 19.

**ANS-12)** **Bagging:** Bagging is used when the goal is to reduce the variance of a decision tree classifier**. Here the objective is to create several subsets of data from training sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees.** As a result, we get an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree classifier.

**Boosting**: Boosting is used to create a collection of predictors. **In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree.**When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly. This process converts weak learners into better performing model.

**Basic difference between Bagging and Boosting:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Bagging** | **Boosting** |  |
| **Partitioning of data** | Random | Higher vote to misclassified samples |  |
| **Goal to achieve** | Minimum variance | Increase accuracy |  |
| **Methods used** | Random subspace | Gradient descent |  |
| **Functions to combine single model** | Weighted average | Weighted majority vote |  |
| **Example** | Random Forest | Ada Boost |  |
| **Overfitting** | Reduces over-fitting of  the model. | Prone to over-fitting. |  |

**ANS-13)** In order to understand adjusted R-squared first understand R-squared:

**R-square** test is used to determine the goodness of fit in regression analysis. Goodness of fit implies how better regression model is fitted to the data points. More is the value of r-square near to 1, better is the model. But the problem lies in the fact that the value of r-square always increases as new variables (attributes) are added to the model, no matter that the newly added attributes have a positive impact on the model or not. Also, it can lead to over fitting of the model if there are large no. of variables.

**Adjusted r-square:**

* Adjusted r-squareon the other hand is a **modified** form of r-square whose value increases if new predictors tend to improve model’s performance and decreases if new predictors does not improve performance as expected.
* R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model.
* Adjusted R-squared adjusts the statistic based on the number of independent variables in the model.
* R2 shows how well terms (data points) fit a curve or line. Adjusted R2R2 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 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

**Adjusted r-square calculation**:

**R­­2adj =** **1- [(1-R2 )\*(n-1) / (n-k-1)]** Where, k is the no. of regressors and n is the sample size.

**ANS-14) 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.**
* Minimum and maximum value of features are used for scaling.
* It is used when features are of different scales.
* It is really affected by outliers
* Scales values between [0, 1] or [-1, 1].
* It is useful when we don’t know about the distribution
* **Formula**: Xnew =(X – Xmin )/(Xmax –Xmin )
* **Example: MinMaxScaler**

**Standardization:**

* **Standardization is another scaling technique where the values are centred 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.**
* Mean and standard deviation is used for scaling.
* It is used when we want to ensure zero mean and unit standard deviation.
* It is much less affected by outliers.
* It is not bounded to a certain range.
* It is useful when the feature distribution is Normal or Gaussian.
* **Formula**: Xnew =(X-mean)/(Standard deviation)
* **Example: StandardScaler**

**ANS-15)** **Cross Validation** is a very useful technique for assessing the performance of machine learning models. It helps in knowing how the machine learning model would generalize to an independent data set.

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

**Techniques**:

1. Leave one out CV
2. K-Fold CV
3. Stratifies K-Fold
4. Time series CV

**Advantages:**   
1. Reduces Overfitting.

2.Cross Validation helps in finding the optimal value of hyper-parameters.

**Disadvantages:**   
1. Increases Training Time

2. Needs Expensive Computation