Answers for Machine Learning Worksheet-4

1. C) between -1 and 1
2. A) Lasso Regularisation
3. C) hyperplane
4. A) Logistic Regression
5. A) 2.205 × old coefficient of 'X'
6. B) D) none of the above (it may increase or decrease as it is a hyper-parameter to tune depending on the data)
7. C) Random Forests are easy to interpret
8. B) Principal Components are calculated using unsupervised learning techniques, C) Principal Components are linear combinations of Linear Variables.
9. A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index, D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.
10. A) max_depth , B) max_features, D) min_samples_leaf

11. Outliers are observations that lie at an anomalous distance from other data points in a random sample from a population.
    Inter Quartile Range (IQR) is the range between 25^{th} percentile(Q1) and the 75^{th} percentile (Q3), i.e, IQR = Q3-Q1.
    IQR method of outlier detection simply implies that, the data points that are either below IQR*1.5 or above IQR*1.5 are considered as outliers.

12. The primary difference between bagging and boosting algorithm is that, bagging algorithm works by combining the results of multiple weak models/learners (for instance, all decision trees) in parallel to get a generalized result, whereas, Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous

model, in which the succeeding models are dependent on the previous model.

13. ***Adjusted R^2*** value tells us how much variance in the dependent variable would be accounted for if the model had been derived from the population from which the sample was taken, also it tells us that, whether adding new variables to the model actually increases the model fit, and penalizes the model score if the R^2 score does not increase sufficiently.
    Mathematically, it is calculated by:

    Adjusted R^2 $= 1 - \frac{(1-R^2)(N-1)}{(N-M-1)}$

    Here,
    R^2 = R-squared value determined by the model
    N = Number of independent variables
    M = Number of data points

    From above formula, we see that, by increasing the number of features if the R^2 score does not increase significantly than the overall **Adjusted R^2** score would decrease and vice-versa.

14.

S. No.	Normalization	Standardization
1.	It uses maximum and minimum value to scale the feature in range of [0,1] or [-1,1].	It uses mean and standard deviation of the feature to scale it, in order to have zero mean and 1 standard deviation.
2.	It is sensitive to outliers.	It comparatively less affected by outliers.

3.	It is usually used when the distribution of the data is not known or is not Gaussian.	It is used when the distribution is almost like a Gaussian or bell-shaped.

15. Cross-validation in Machine Learning is a model evaluation technique, where the input data is divided into training and validation set internally, and the model is trained on the training set and evaluated on the validation set. It is used to detect overfitting where the model fails to generalize a pattern.

    Advantage: It can be used in hyper-parameter tuning of the model to select best parameters based on the validation score.

    Disadvantage: Since, a subset of the input data (or training data) is used as a validation data, therefore, is it recommended to have sufficient data to perform this technique in order to prevent bias in the dataset.