MACHINE LEARNING SHEET 4

1. C) between -1 and 1

2. All are used for dimensionality reduction

3. C) hyperplane

4. Naïve Bayes Classifier

5. B) same as old coefficient of ‘X’

6. D) none of the above

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. B) Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts, C) Identifying spam or ham emails

10. A) max_depth, D) min_samples_leaf

11. Outliers are data points that differ significantly from other observations. Find Interquartile range, then find upside outlier by adding median of upper half and (1.5 * interquartile range) then find downside outlier by subtracting median of lower half and (1.5 * interquartile range)

12. In Bagging Training data subsets are drawn randomly from entire training dataset while in Boosting each new subset contains components that were misclassified by previous methods.
    Bagging is to tackle overfitting, while Boosting is to reduce bias.

13. Adjusted R^2 statistics is used when number of independent variables increase. As we increase independent variables in our equation, the R^2 also increases. This doesn’t mean results would improve. To rectify this , we use adjusted R^2 VALUE which penalises excessive use of such features which do not correlate with output data.

    $R^2_{adj} = 1 – [(1 - R^2)(N – 1)] / (N - p – 1)$

    N = Total sample size, p = Number of predictors

14. Standardization (or **Z-score normalization**) is the process of rescaling the features so that they’ll have the properties of a Gaussian distribution, while Normalization basically shrinks the range of the data such that the range is fixed between 0 and 1 (or -1 to 1 if there are negative values)

15. In Model building, if testing data is repeated from training data, then the prediction model will overfit. To avoid that we use Cross validation. Cross validation forms different data subsets from the entire data set and tests with different combinations.

    Disadvantage is that it is very time consuming hence very costly.