Machine Learning

Worksheet4

1. C) between -1 and 1

2. D) Ridge Regularisation

3. C) hyperplane

4. D) Support Vector Classifier

5. B) same as old coefficient of 'X'

6. C) decrases

7. C) Random Forests are easy to interpret

8. B and C

9. A and D

10. A,B,C and D

11. Outliers are the data points in the dataset which significantly differ from other observations. These are also known as the reason for noise in the dataset. We can classify data into 3 quartiles.
1st (lower) quartile (Q1): median of the lower half of the data
2nd quartile (Q2): median of the entire data
3rd (upper) quartile (Q3): median of the upper half of the data
IQR is given by the difference of Q3 and Q1. This is the range where bulk of the data lies. The data points lower with values lower than 1.5*Q1 and higher than 1.5*Q3 are generally termed as outliers cause problems while preparing statistical analysis.

12. Bagging and boosting are the types of method used in ensemble learning techniques.

Bagging algorithm takes homogenous independent weak learning models and combines them parallel and learn from them.
while Boosting algorithm takes homogenous weak learners and learn them sequentially and adaptively improve the models prediction making it a strong learner.

13. R2 shows how well data fit a curve. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit. Adjusted R2 will be always lesser than R2.
In simple terms adjusted R2 penalizes if there is data that does not add values to the model. Adj R2 is calculated by the following formula:
Adj R2 = 1 -[(1-R^2)(n-1)/n-k-1]
Where n is the number of data points and K is the number of predictors

14. Normalization and Standardization are the methods used for scaling the data. 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. Standardization is another scaling technique where the values are centered 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.

15. Cross validation is a technique used to fit different data and test different data in every iteration. This technique which involves reserving a particular sample of a dataset on which you train the model. Then, you test your model on this sample before finalizing it. This makes sure that the sample used for training and testing does not bias the model. For eg, if the cv is set to 5, then there are 5 train test splits done of the data, each with different data and testing with the remaining i.e. test data of very split.