Worksheet4 MACHINE LEARNING

In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:
C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?
C) Recursive feature elimination

3. Which of the following is not a kernel in Support Vector Machines?
D) polynomial

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
B) Naïve Bayes Classifier

5. In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' will be?
(1 kilogram = 2.205 pounds)
A) 2.205 × old coefficient of 'X'

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
B) increases

7. Which of the following is not an advantage of using random forest instead of decision trees?
A) Random Forests reduce overfitting

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. Which of the following are correct about Principal Components?
B) Principal Components are calculated using unsupervised learning techniques
C) Principal Components are linear combinations of Linear Variables.

9. Which of the following are applications of clustering?
A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
C) Identifying spam or ham emails
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. Which of the following is(are) hyper parameters of a decision tree?
A) max_depth B) max_features
D) min_samples_leaf

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

We can use IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5times the IQR and then subtract this value from Q1 and add this value to Q3.

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

Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of original data. Boosting is a iterative strategy for adjusting an observation’s weight based on the previous classification.

13. What is adjusted R_2 in linear regression. How is it calculated?

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables(predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

14. What is the difference between standardisation and normalisation?

Normalized dataset will always have values that range between 0 and 1. A standardized dataset will have mean of 0 and standard deviation of 1, but there is no specific upper or lower bound for maximum and minimum values.

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

Cross-Validation is a very powerful tool. It helps us better use our data, and it gives much more information about algorithm performance. In complex machine learning models, it’s sometimes easy not pay enough attention and use the same data in different steps of pipeline.