MACHINE LEARNING-4

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

1. The value of correlation coefficient will always be:

Ans=C) between -1 and 1

1. Which of the following cannot be used for dimensionality reduction?

Ans=A) Lasso Regularisation

1. Which of the following is not a kernel in Support Vector Machines?

Ans=C) hyperplane

1. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Ans=A) Logistic Regression

1. 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)

Ans=D) Cannot be determined

1. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

Ans=B) increases

1. Which of the following is not an advantage of using random forest instead of decision trees?

Ans=C) Random Forests are easy to interpret

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

1. Which of the following are correct about Principal Components?

Ans=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?

Ans=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.

10. Which of the following is(are) hyper parameters of a decision tree?

Ans=B) max\_features

Q11 to Q15 are subjective answer type questions, Answer them briefly.

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

Ans=IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

An outlie**r** is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

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

Ans=In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.

13. What is adjusted R2 in linear regression. How is it calculated?

Ans=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.

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?

Ans=Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Ans=Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

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

 Disadvantage of k-fold cross validation relative to LOOCV: If the main purpose bias reduction, LOOCV should be preffered to k-fold CV since it tends to has less bias.