Machine Learning assignment 4

1. The value of correlation coefficient will always be:

Ans : C) between -1 and 1

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

Ans : A) Lasso Regularisation and D) Ridge Regularisation

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

Ans : C) hyperplane

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

Ans : A) Logistic Regression

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)

Ans : 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?

Ans : D) none of the above

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

Ans : C) Random Forests are easy to interpret

8. 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

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?

Ans : A) max_depth B) max_features D) min_samples_leaf

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

Ans : outliers are values within a dataset that vary greatly from the others—they're either much larger, or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty**.**

In IQR method of identifying outliers we set up a "fence" outside of Q1 (first quartile) and Q3 (third quartile). Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers.

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

Sl No	Bagging	Boosting
1.	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2.	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3.	Each model receives equal weight.	Models are weighted according to their performance.
4.	Each model is built independently.	New models are influenced by the performance of previously built models.
5.	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.
6.	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias
7.	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8.	In this base classifiers are trained parallelly	In this base classifiers are trained sequentially.
9.	Example: The Random forest model uses Bagging.	Example: The AdaBoost uses Boosting techniques

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

Adjusted R2 measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/\text{df}_e}{SS_{\text{tot}}/\text{df}_t}$$

In the above equation, dft is the degrees of freedom n– 1 of the estimate of the population variance of the dependent variable, and dfe is the degrees of freedom n – p – 1 of the estimate of the underlying population error variance.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. The equation is as below

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

14. What is the difference between standardisation and normalisation?

Sl No.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

Advantage : Reduces Overfitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

Disadvantage : Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.