MACHINE LEARNING WORKSHEET

1. The value of correlation coefficient will always be:

A) between 0 and 1

B) greater than -1

C) between -1 and 1

D) between 0 and -1

Answer : between -1 and 1

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

A) Lasso Regularisation

B) PCA

C) Recursive feature elimination

D) Ridge Regularisation

Answer : Ridge Regularisation

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

A) linear

B) Radial Basis Function

C) hyperplane

D) polynomial

Answer : linear

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

A) Logistic Regression

B) Naïve Bayes Classifier

C) Decision Tree Classifier

D) Support Vector Classifier

Answer : 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'

B) same as old coefficient of 'X'

C) old coefficient of 'X' ÷ 2.205

D) Cannot be determined

Answer : old coefficient of 'X' ÷ 2.205

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

A) remains same

B) increases

C) decreases

D) none of the above

Answer: increases

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

A) Random Forests reduce overfitting

B) Random Forests explains more variance in data then decision trees

C) Random Forests are easy to interpret

D) Random Forests provide a reliable feature importance estimate

Answer : Random Forests explains more variance in data then decision trees

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

A) Principal Components are calculated using supervised learning techniques

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

D) All of the above

Answer : 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

B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.

C) Identifying spam or ham emails

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Answer : 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?

A) max_depth

B) max_features

C) n_estimators

D) min_samples_leaf

Answer : max_depth , max_features ,min_samples_leaf

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

Answer : Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

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

Inter Quartile Range(IQR): IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

- Q1 represents the 25th percentile of the data.
- Q2 represents the 50th percentile of the data.
- Q3 represents the 75th percentile of the data.

If a dataset has 2n / 2n+1 data points, then
Q1 = median of the dataset.
Q2 = median of n smallest data points.
Q3 = median of n highest data points.

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.

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

Answer :

Bagging is a method of merging the same type of predictions.

Boosting is a method of merging different types of predictions.

Bagging decreases variance, not bias, and solves over-fitting issues in a model.

Boosting decreases bias, not variance.

Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data.

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

Bagging gives equal weight to each model, whereas in Boosting technique, the new models are weighted based on their results.

In boosting, new subsets of data used for training contain observations that the previous model misclassified. Bagging uses randomly generated training data subsets.

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

Answer : Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression.

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.

Let’s have a look at the formula for adjusted R-squared to better understand its working.

$$Adjusted\ R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$

Here,

n represents the number of data points in our dataset

k represents the number of independent variables, and

R represents the R-squared values determined by the model.

So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

Adjusted R squared is calculated by **dividing the residual mean square error by the total mean square error** (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

14. What is the difference between standardisation and normalisation?

Answer : In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values.

Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

Normalization is the process of organizing data in a database. This includes creating tables and establishing relationships between those tables according to rules designed both to protect the data and to make the database more flexible by eliminating redundancy and inconsistent dependency.

Standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables.

Typically, to standardize variables, you calculate the mean and standard deviation for a variable.

Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

Standardization does not get affected by outliers because there is no predefined range of transformed features.

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

Answer : Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

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.

Cross-validation is a machine learning technique where the training data is split into two parts: A training set and a test set. The training set is used to build the model, and the test set is used to evaluate how well the model performs when in production.

Advantage :

An advantage of using this method is that we make use of all data points and hence it is low bias.

Disadvantage:

The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point.