**MACHINE LEARNING**

**WORKSHEET – 1**

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- D) Ridge 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- A) 2.205 × old coefficient of ‘X’

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

Ans- A) remains same

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

Ans- A) Random Forests reduce overfitting

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.

1. Which of the following are applications of clustering?

Ans-

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

C) Identifying spam or ham emails

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

Ans- A) max\_depth B) max\_features D) min\_samples\_leaf

**subjective answer type questions**

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

Ans- An outlier in a data set is a value that is far away from the rest of the values in the data set.

Quartile Range(IQR):

The set of data can be described by its five number summary. These five numbers which given the information we need to find the patterns and outliers.

Data sets statistical Order

* The minimum or lowest value of the dataset
* The first quartile Q1, which represents a quarter of the way through the list of all data
* The [median](https://www.thoughtco.com/what-is-the-median-3126370) of the data set, which represents the midpoint of the whole list of data
* The third quartile Q3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

Range would be difficult to extrapolate otherwise. Similar to the range but less sensitive to outliers is the interquartile range. The [interquartile range](https://www.thoughtco.com/what-is-the-interquartile-range-3126245) is calculated in much the same way as the range. All you do to find it is subtract the first quartile from the third quartile.

IQR = *Q*3 – *Q*1

The interquartile range shows how the data is spread about the median. It is less susceptible than the range to outliers.

 the interquartile range can be used to detect outliers.

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.
5. What is the primary difference between bagging and boosting algorithms?

Ans-

**The Bagging** Also known as Bootstrap Aggregation is an ensemble method. First, we create random samples of the training data set (sub sets of training data set). Then, we build a classifier for each sample. Finally, results of these multiple classifiers are combined using average or majority voting. Bagging helps to reduce the variance error.

Bagging: Random Forest

**The Boosting** provides sequential learning of the predictors. The first predictor is learned on the whole data set, while the following are learnt on the training set based on the performance of the previous one. It starts by classifying original data set and giving equal weights to each observation. If classes are predicted incorrectly using the first learner, then it gives higher weight to the missed classified observation. Being an iterative process, it continues to add classifier learner until a limit is reached in the number of models or accuracy. Boosting has shown better predictive accuracy than bagging, but it also tends to over-fit the training data as well.

Boosting : Ada Boost, Gradient Boosting, XGBoosting, etc

1. What is adjusted R2 in logistic regression. How is it calculated?

Ans –

It is a measure of explained variation, instead of goodness of fit.

All the outcomes/dependent variables are often reasonably modeled using rectilinear regression . Perhaps the second commonest sort of regression model is logistic regression, which is acceptable for binary outcome data.

Logistic regression models are fitted using the tactic of maximum likelihood - i.e. the parameter estimates are those values which maximize the likelihood of the info which are observed. R squared measure is defined as

R^{2}\_{\text{null}} = 1- \frac{log(L\_c)}{log(L\_{\text{null}})}

where L\_c denotes the (maximized) likelihood value from the present fitted model, and L\_{\text{null}} denotes the corresponding value except for the null model - the model with only an intercept and no covariates.

1. What is the difference between standardisation and normalisation?

Ans –

|  |  |
| --- | --- |
| Normalization | Standardization |
| Normalization usually means to scale a variable to have a values between 0 and 1. | Standardization is the process of transforming a variable to one with a mean of 0 and a standard deviation of 1. |
|  |  |

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

Ans- Cross validation is a technique for assessing how the statistical analysis generalizes to an independent data set. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Using cross-validation, there are high chances that we can detect over-fitting with ease.

* The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.
* The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. The variance of the resulting estimate is reduced as k is increased.