**MACHINE LEARNING WORKSHEET – 1**

In Q1 to Q7, only one option is correct, Choose the correct option: (correct answers are marked in yellow)

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

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

A) Lasso Regularisation B) PCA

C) Recursive feature elimination D) 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

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

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

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

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

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?

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

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.

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

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: **Outliers** can be defined as the value that lies outside, the values that are either much smaller or larger than most of the other values in a data set. For example in the scores 25,29,3,32,850,33,27,28 both 3 and 850 are "outliers". The outliers may suggest experimental errors, variability in a measurement, or an anomaly. For example: The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset as it drasitcally change the mean,mode and median values for a dataset , so to remove these outliers we have different methods like box plot, Interquartile range method etc.

**Inter Quartile Range(IQR)** is used to measure variability by dividing a data set into quartiles/percentile We refer to the percentiles as quartiles (“quart”) because the data is divided into four groups via the 25th, 50th and 75th percentile values.( Q1 represents the 25th percentile of the data,Q2 represents the 50th percentile of the data,Q3 represents the 75th percentile of the data.) The IQR defines the middle 50% of the data(Q3-Q1).The IQR can be used to identify outliers by defining limits on the sample values that are a factor k of the IQR below the 25th percentile or above the 75th percentile. The common value for the factor k is the value 1.5. 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 – 1.5IQR or above Q3 + 1.5 IQR are outliers.  
**By setting these ranges in our dataset we successfully remove the outliers from the data set**

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

Ans: Bagging and Boosting are both ensemble techniques, where a set of weak learners are combined to create a strong learner that obtains better performance than a single one.

Bagging tries to implement similar learners on small sample population and then takes a mean of all the predictions. In generalised bagging, we can use different learner on different population as this can help in reducing variance error.

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it adjust the weight of this observation. Boosting in general decreases the bias error and build strong predictive models, however they may overfit on training data.

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

Ans: R-squared(R2) is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.R2 and adjsuted R2 are the method used in evaluating the performance of linear regression model using ordinary least squares (OLS) method.

Since Logistic regression is used to predict probablities and rely on “maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply

The analogous metric of adjusted R2 in logistic regression is AIC (Akaike Information Criteria). AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value and it can be calculated as:

AIC=−2log(L)+2K where L = maximum likelihood from the MLE estimator, K is number of parameters

14. What is the difference between standardisation and normalisation?

Ans: In the field of data science, normalization is a transformation of data which allows easy comparison of the data downstream. There are many types of normalizations. Scaling being one of them. You can also log the data, or do anything else you want. The type of normalisation we use would depend on the outcome we want, since all normalisations transform the data into something else.

Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

Standardization typically means that the range of values are "standardized" to measure how many standard deviations the value is from its mean.

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

Ans: Generally we perform training on the 70% of the given data-set and rest 30% is used for the testing purpose. The major drawback of this method is that we perform training on the 50% of the dataset, it may possible that the remaining 50% of the data contains some important information which we are leaving while training our model i.e higher bias.

This drawback can be overcome by cross validation technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set, by doing this we train our model with all the data in subsets, the types of cross-validation used are:

LOOCV (Leave One Out Cross Validation)

In this method, we perform training on the whole data-set but leaves only one data-point of the available data-set and then iterates for each data-point.

K-Fold Cross Validation

In this method, we split the data-set into k number of subsets(known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

Major Advantage and Disadvantage of this techniques are:

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.