MACHINE LEARNING

WORKSHEET – 1

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

1. The value of correlation coefficient will always be:

C) between -1 and 1

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

D) Ridge Regularisation

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

C) hyperplane

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

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)

C) 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?

B) increases

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

C) Random Forests are easy to interpret

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

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

D) min_samples_leaf

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

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 drastically changes the mean, median and mode values for a dataset, so to remove these outliers we have different methods like box plot, Interquartile range method etc.

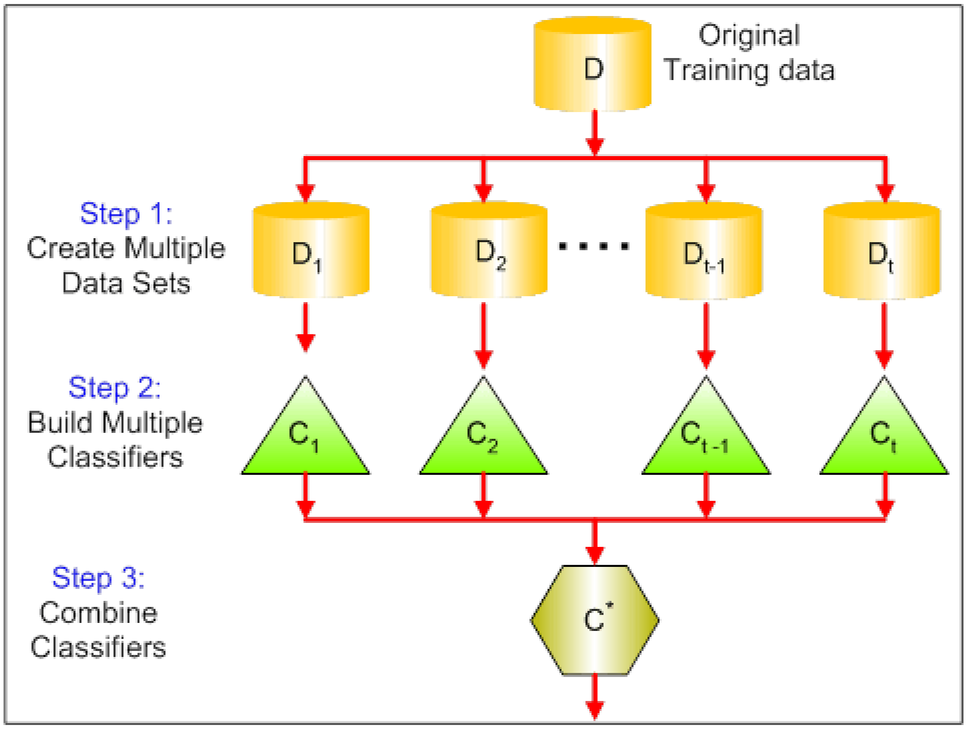
Inter Quartile Range(IQR) is used to measure variability by dividing a data set into quartiles/percentile We refer to the percentiles as quartiles (“quart”) because the data is divided into four groups via the 25th, 50th and 75th percentile values. Q1 represents the 25th percentile of the data,Q2 represents 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?

Bagging based Ensemble learning:

Bagging is one of the Ensemble construction techniques which is also known as ***Bootstrap Aggregation***. Bootstrap establishes the foundation of Bagging technique. Bootstrap is a sampling technique in which we select “n” observations out of a population of “n” observations. But the selection is entirely random, i.e., each observation can be chosen from the original population so that each observation is equally likely to be selected in each iteration of the bootstrapping process. After the bootstrapped samples are formed, separate models are trained with the bootstrapped samples. In real experiments, the bootstrapped samples are drawn from the training set, and the sub-models are tested using the testing set. The final output prediction is combined across the projections of all the sub-models.

The following infographic gives a brief idea of Bagging:



Boosting-based Ensemble learning:

Boosting is a form of *sequential learning* technique. The algorithm works by training a model with the entire training set, and subsequent models are constructed by fitting the residual error values of the initial model. In this way, Boosting attempts to give higher weight to those observations that were poorly estimated by the previous model. Once the sequence of the models are created the predictions made by models are weighted by their accuracy scores and the results are combined to create a final estimation. Models that are typically used in Boosting technique are XGBoost (Extreme Gradient Boosting), GBM (Gradient Boosting Machine), ADABoost (Adaptive Boosting), etc.

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

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 adjusted 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 probabilities 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 standardization and normalization?

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here’s the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1
- If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here’s the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Now, the big question in your mind must be when should we use normalization and when should we use standardization?

Normalization vs. standardization is an eternal question among machine learning newcomers. Let me elaborate on the answer in this section.

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.

It is a good practice to fit the scalar on the training data and then use it to transform the testing data. This would avoid any data leakage during the model testing process. Also, the scaling of target values is generally not required.

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

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.ehigher 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 leave 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 these 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.