Machine Learning

Corporate Office Address

iASYS Technology Solutions Pvt. Ltd.

25/5 Rajiv Infotech Park

Hinjewadi Phase-III Pune 411057 (India)

Ph: +91 20 2552 0602

www.iasys.co.in

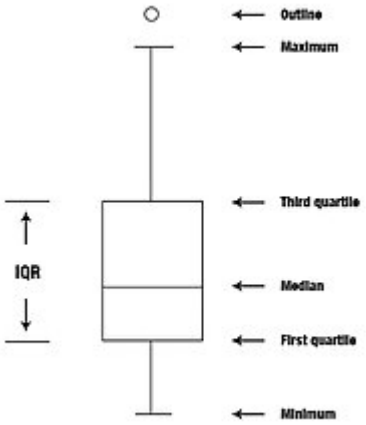
Date: *June 08th 2023*

Documentation by: *Adharsh S*

Table of Contents

1. Basic Understanding Required

- Int
- Float
- String
- List: Cluster of values with different or the same data type
- Tuple: Same as list but the values cannot be changed after the creation.
- Set: Cannot have duplicate values
- Dictionary: Key -Value pair
- Boolean: True / False
- If Elif Else
- For Loop/ While Loop - Indexing
- Functions
- Classes and Objects
  - Classes are the blueprints and objects are the instants of the class.
- Arrays (SIMD – Simple Input Multiple Data)
- Central Measure - Mean / Median / Mode
  - Mode: Categorical data
  - Mean: Numeric data with a smaller number of outliers
  - Median: Numeric data with a greater number of outliers
- Spread – Range / IQR / Standard Deviation
  - Range: Max – Min
  - IQR: 75^{th} Percentile – 25^{th} Percentile
  - SD: Sqrt (Mean of X^2 – (Mean of X) ^2)
- Visualization (EDA – Exploratory data analysis)
  - Scatter Plot
  - Histogram: For continuous data
  - Bar Plot: (X – Feature and Y – Frequency) – For discrete data
  - Pie Plot: For Discrete Data
  - Box Plot: For finding outliers.



- Data Cleaning
  - Missing / Null Values
    # Drop the data
    # Impute with the existing data
  - Data not in the right format
  - Duplicates
  - Outliers
  - Structured Data
  - Textual Data related issues
- Data Science Solutioning Process
  - Define the problem statement.
  - Collect the data for the solution.
  - Clean the data.
  - EDA (Exploratory data analysis)
  - Construct a model building (Multiple Models)
  - Validation of Models.
  - Interpret the model (Knowing how the model comes up with the particular conclusion)
  - Deployment

2. Machine Algorithm

Any algorithm where the rules are automatically learned by the algorithm is called the ML algorithm.

ETA Formal Definition: An algorithm is called an ML algorithm (measured using a metric A) if the performance of the algorithm increases in each task T with more experience E.

Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

3. 6 JARS OF ML

3.1 Data JAR

- How to prepare the data before applying with ML
- Cleaning of the data
- Encoding of the data
  - Ordinal data (with some order) – Label Encoding (Binary category)
  - Nominal data (without some order) – One hot Encoding
- Split the data.
  - Train data – Develop the ML model.
  - Test data – Test the ML model.
- Scaling of the dataset (not mandatory)
  - Normalization / Standard Scaling
    For X data = (X – mean(X))/ S.D (X)

3.2 Task JAR

- Supervised Learning – Predict a target variable.
  - Regression Predictions – Predict a Continuous Value
  - Classification Predictions – Predict a Categorical Value
- Unsupervised Learning
  - No Target Value.
  - Analyze the pattern.
  - Clustering - group the data
  - Dimensionality reduction
- Reinforcement Learning
  - Feedback will be delayed when we do the predictions.

3.3 Model JAR

- A mathematical formula/ representation of an ML algorithm
- Y = mx +c (straight line equation), where m and c are the parameters of the model which we are the unknown values.
- Algorithm finds the best value of m and c to find a relationship between y and x.
- All the ML models can be explained using a mathematical model and all models will have parameters.
- Logistic regression 1/ 1+ e^ (-mx – c)

3.4 Loss JAR

- Formula to measure how far my predictions are from the target variable.
- Lower the loss, better the predictions, and vice versa.
- Eg: 1/n * summation (i = 1 to n) (yi – yi^)^2 : Mean Square error
- From the historical data, the best parameters will be found with the least loss.

3.5 Learning JAR

- Hit and Trial Method
  - Will try out all possible parameters and choose the least loss parameters.
- Gradient Descent Approach
  - Without knowing how the parameter vs loss graph is, we will use gradient descent to find the least loss.
  - Choose a random value for "m".
  - Compute slope
  - Walk along the slope in a downward direction.
  - Continue the same until you find the bottommost point.
  - We can find only the local best value.
  - GD algorithm doesn't work in all cases. The GD will work if there is only one minimum (ConveX). If there are many minima (non-convex), the GD will not work.
  - For a non-convex loss function, we calculate the least loss starting from multiple points of m.
  - For both convex and non-convex loss, we use GD.

3.6 Evaluation JAR

4. Supervised Learning

The Algorithm where the models are getting trained with a set of target values is called Supervised Learning.

It is of two types:

- Regression Algorithm
- Classification Algorithm

4.1 Regression Algorithm

The algorithm, which is used to find a true value, is called the regression algorithm.

4.1.1 Linear Regression Algorithm

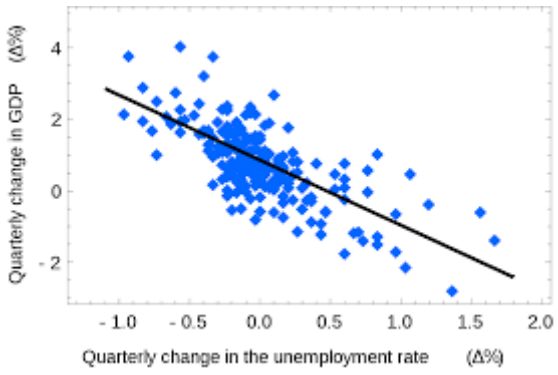
A linear regression algorithm is a model where the target and parameters are related by power 1.

Y =mx + c or Y = mx^2 + c follows linear expression.

Y = (m^2)*x + c or Y = mx +c^2 does not follow linear expression

It finds a linear relationship or straight-line relationship between the feature and target.

- Data
- Task – Regression - Supervised Learning – Predict a continuous value.
- Model – Y = mx + c, where m and c are the parameters



- Can be found using plotting.
- Correlation: A formula that can measure the strength of a linear relationship
- Correlation can take a value between -1 to +1
- Correlation = Summation (i= to n) [(Xi – mean(X))(Yi – mean(Y)]/[V(x)*V(y)]
- Value greater than 0.2 to 1 or -0.2 to -1, shows a good linear relationship either in positive or negative direction.
- Low correlation (-0.2 to 0.2) means no linear relationship exists. But it doesn't mean that there is no relationship.
- If there is no correlation for a linear relationship, I can do transforming the feature.
- X will be converted as X^2 or X^3 or sqrt(X) or e^X or log(X)
- If it results in a linear relationship, I can apply linear regression.

- Loss
  - Mean Squared error – when there is less outlier as the square is more sensitive for outliers. (Convex loss function)
  - Mean absolute error – when there are a lot of outliers. (We will consider a convex loss function)
- Learning
  - Gradient Descent
- Evaluation JAR
  - R^2 will be based on the problem statement, which is linear regression here.
  - R^2 = 1 – ((summation i=1 to n)[(Yi – Yi^)^2)/((summation i=1 to n)(Yi – Y^)^2]
  - Yi – true value, Yi^ - predicted value, Y^ - Mean of true target value)
  - R2 will take a value between – infinity to 1.
  - R2 having a higher value is better than the model.
  - If R^2 comes less than the baseline value 0 is a useless model.

Simple Linear Regression ML Model

https://github.com/Adharsh0001/Machine-Learning/blob/main/Linear_Regression.ipynb

Multiple Linear Regression ML Model

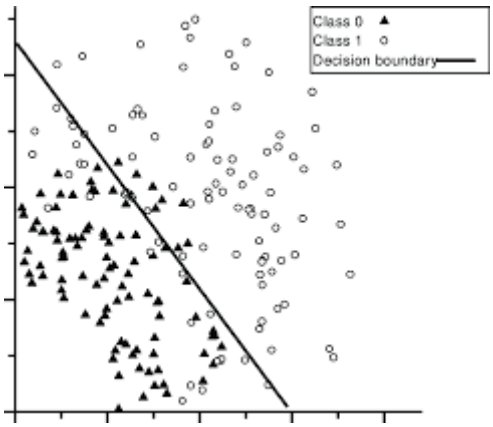
https://github.com/Adharsh0001/Machine-Learning/blob/main/Multiple_Linear_Regression.ipynb

4.2 Classification Algorithm

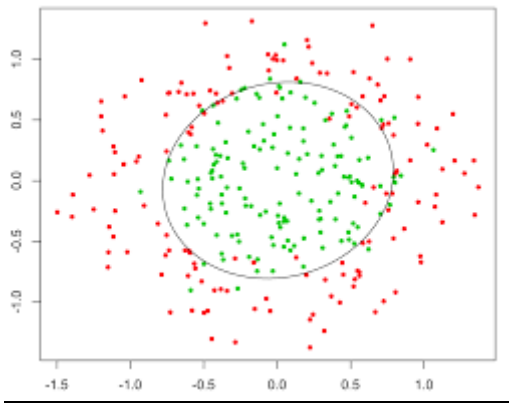
As an ML expert, I will try Logistic Regression, KNN and Decision Tree and will use the model with the best result (cross-validation score)

4.2.1 Logistic Regression

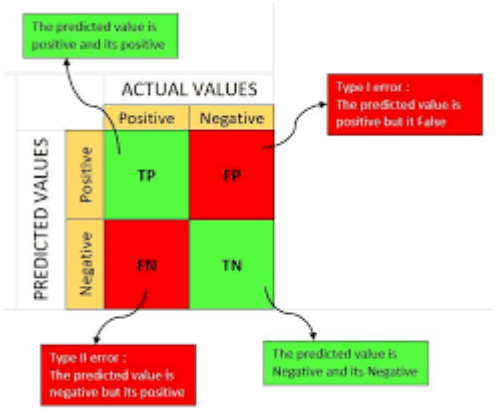
- Scaling is not mandatory for logistic regression.
- Data JAR – Clean, Encode, Split, Scale
- Task: Supervised Learning _ Logistic for classification
- Model JAR
  - Y = 1/ 1 + e^ (-mx-c) – Sigmoid Function gives graph like S
  - Y will take a value between 0 to 1
  - We will consider the threshold value as 0.5.
  - If the value comes above 0.5, we accept or if the value comes below 0.5, we reject.
- Logistic Regression model can find only linear decision boundaries.



- Eg for Non-Linear Decision Boundary

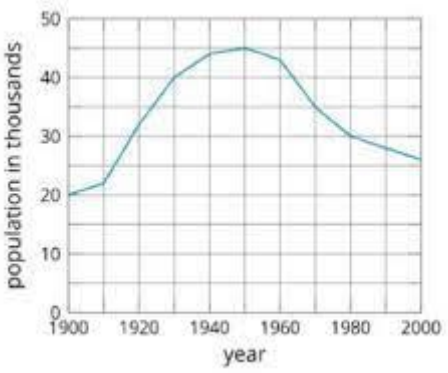


- Loss JAR
  - Log Loss = summation of i =1 to n (-Yi log(Yi^))-((1-Yi)*log(1-Yi^))
  - Yi – true category, Yi^ - predicted probability
  - Lower the loss, the better the model.
- Learning JAR
  - Hit and Trial – Time-consuming
  - Gradient Descent – It will work 100% for the Logistic Regression model.
- Evaluation Metric Jar
  - Accuracy
    = (number of correct predictions/ total number of predictions) *100
  - F1-Score Confusion Matrix



- F1 -Score = True Positive / (TP + ½ (FP +FN))
- F1 score can take values between 0 to 1.
- Higher the value, the higher the accuracy.
  - AUROC - Area under the receiver operating characteristic curve
- ROC plot
- From the confusion matrix
  True positive rate (TPR) = TP / (TP +FN)
  False positive rate (FPR)= FP / (TN+FP)

- From the TPR and FPR, we need to draw a graph which is called ROC. Where FPR will be on X-axis and TPR will be on Y-axis.
- We will get one TPR and FPR for one threshold value. So multiple threshold values are used to find more TPR and FPR.
- We will change the thresholds from 0.1 to 1 at an increment rate of 0.1.



- The area under the ROC is AVROC
- AUROC can take a value between 0 to 1.
- Higher the value, the better the model.
- AUROC value of more than 0.5 is a good model.
- AUROC value less than 0.5 is a useless model. (Added advantage over F1 score)

Logistics Regression ML Model

https://github.com/Adharsh0001/Machine-Learning/blob/main/Logistic_Regression.ipynb

4.2.2 K-Nearest Neighbour (KNN)

- For the models where we don't have linear decision boundaries, we can go with KNN's nearest neighbor.
- For classification problems, we can use KNN.
- KNN works using distance metrics.
  - Euclidian metrics
    Sqrt [(X2- X1) ^2 + (Y2 – Y1) ^2]
- It will get the K value from the datasets.
- It will compute the distance between all data points.
- Choose the K nearest neighbors.
- Predict the majority class.
- Whenever we give a value in ML, those are called hyperparameters as the user will load the K value.
- Hyperparameters are found by Hit and Trial method followed by Cross-Validation Score.
- Data – Clean / Encode/Split /Scale
  - Scaling is a must for the KNN algorithm.
- Task JAR – Supervised Learning – Classification (Can be used for regression as well)
- Model - It will get the K value from the datasets.
- It will compute the distance between all data points.
- Choose the K nearest neighbors.
- Predict the majority class.
- Loss
  - There is no loss function as there is no parameter.
  - It is called a Lazy algorithm.
- Learning
  - No Learning
- Evaluation JAR
  - Auroc / F1 score / Accuracy.

Cross Validation

- Since we can't use test data for evaluating the model, we come up with the idea of testing the model with a cross-validation method.

Data – 70% to train and 30% to test

Data is again split into 5 or 10 bins.

Let's assume 70% of data is around 1000datas.

Then each bin will have 100datas if we are using 10 bins.

Test 1 - Bin 2 to Bin 10 will be used to create the model – Bin 1 will be used to find the AUROC value.

Test 2 - Bin 3 to Bin 1 will be used to create the model – Bin 2 will be used to find the AUROC value.

.

.

.

Test m - the test continues till the number of bins is divided.

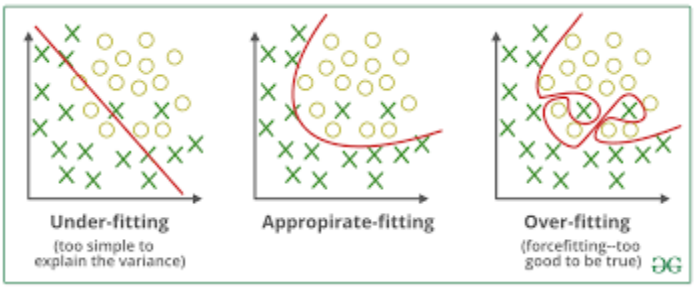
The average of all the AUROC values is considered and this process is called cross-validation.

Overfitting

- If the model starts to memorize data instead of pattern, it is called Overfitting.
- If the model is performing well with the training data than another model but performs poorer than another model over test data or validation data is called an Overfit model.

Underfit Model

- If a model is lost in the training data set and as well as in test data with another model, it is called an Underfit model.



- In a KNN model, if the K value increases, the plot decision becomes smoother i.e., the model gets smoother.
- Trade-off happens by comparing training data accuracy and cross-validation score.
- The model with the best values with both train accuracy and Val accuracy is the best model.

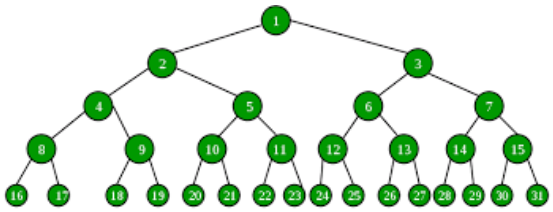
KNN ML Model

https://github.com/Adharsh0001/Machine-Learning/blob/main/KNN_Algorithm.ipynb

4.2.3 Decision Trees

4.2.3.1 Decision Trees for Classification

- Decision trees can be used for both regression and classification.
- It follows the binary tree.



- The place where we ask questions is called a node.
- The place where we make the decision is called a leaf.
- The levels and leaves are connected by a formula (2^(n-1))
- To build a decision tree, the ML must ask the correct questions.
- A question must have a clear separation. (Right Question)
- To check the measure of the split of data, we use entropy.
  - Entropy – Measures the Randomness
  - Entropy = summation i=1 to k (-Pk * log(K))
  - K – no. of classes in the data
  - Pk – Proportion of data in class k
  - Entropy ranges between 0 to 1
  - The lower the entropy, the better the split.
  - Reduction of Entropy = entropy of the parent – the proportion from parent * entropy of child 1 - the proportion from parent * entropy of child 2
  - Reduction in entropy ranges between 0 to 1.
  - The higher the reduction, the question get selected.
- Data - Clean / Encode (Can go with Label Encode. It is sufficient) / Split / Scaling (Scaling is not mandatory for the decision tree)
- Task – Supervised Learning – Classification (Can be used for regression as well)
- Model – Follows binary tree.
- Loss – Reduction in Entropy.
- Learning – Hit and Trial
- Evaluation JAR – Accuracy/ F1-Score/ AUROC
- If the decision tree grows more, it becomes more complex; hence, we can expect to overfit the issue.
- Hence, we pass a hyperparameter that defines the depth of the decision tree.

- The hyperparameter will be finalized using the hit and trial method and cross-validation.
- The feature which has the greatest absolute value will be the best feature.

Decision Tree for Classification ML Model

https://github.com/Adharsh0001/Machine-Learning/blob/main/Decision_Tree.ipynb

4.2.3.2 Decision Trees for Regression

- Decision tree used for Regression problems.

Decision tree for Regression ML Model

https://github.com/Adharsh0001/Machine-Learning/blob/main/Decision_Tree_for_Regression.ipynb

4.3 Imbalanced Learning

- It's Supervised Learning.

- Imbalanced data:

Any dataset where 70% of data belongs to one class or is skewed to one class is called an Imbalanced dataset.

- Most of the classification algorithms will be based on an Imbalanced dataset.

Problems in Imbalanced datasets

Problem 1

- Accuracy is a bad metric for an Imbalanced dataset. Since even 99% of accuracy will not be sufficient for the evaluation of the dataset.
- E.g.: Fraud transaction – 1% and Proper transaction – 99%
- F1 score and AUROC can be used.

Problem 2

- All the traditional ML algorithms will give biased results for imbalanced datasets.
- We are interested in predicting the minority class.
- We need to balance the data.
- For balancing the data, we will remove most of the value from the majority class (Under sampling) or increase more value in the minority class (Oversampling)
- Under sampling leads to Under sampling and Oversampling leads to Overfit issues.

Smart Way to Under Sampling

- Group the majority class into sub-class by using K-Means class.
- Remove the data around some % from all the sub-class.
- This method is called Cluster Centroid Under Sampling.

Smart Way to Over Sampling

- Choose a data point randomly.
- Compute the 2 nearest neighbors.
- Calculate the center of the three different points and add new data.

- SMOTE – Synthetic Minority Oversampling Technique

Combined Way of Balancing the Data

- SMOTEENN – Reduces the majority class near the Minority class and also creates more minority class data to balance the dataset.
- Cluster Centroid for removing data but the data that are only close to the minority class.

Imbalanced Learning ML Model

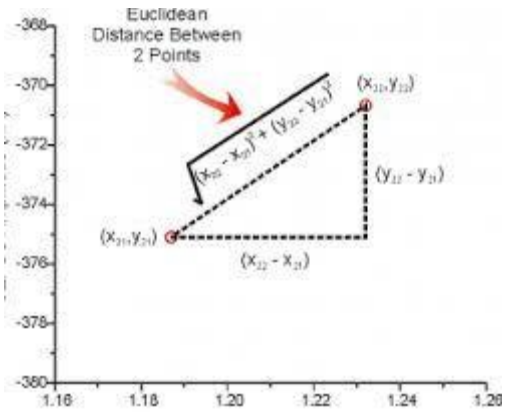
https://github.com/Adharsh0001/Machine-Learning/blob/main/Imbalanced_Learning.ipynb

5. Unsupervised Learning

- There are no targets.
- We are interested in finding the patterns.

5.1 K – means Clustering.

- It follows Euclidian distance.



- Scaling is mandatory.
- Get the value from the data scientist.
- The K which is passed is called a hyperparameter.
- Splitting is not required as we are not trying to predict.
- Randomly chosen points (k) are called Centroids.
- Will measure the distance between each data to each centroid and will assign it to the closest to the centroids.
- Then the data will be clustered into a few groups as per the closest distance from the centroids.
- Then we will find the average value for each group and call them the centroid.
- Will repeat the process again, measuring the distance and forming a new group and finding the avg value as the new centroid.
- Will repeat this process, until the group doesn't get changed.
- To find the best value of K, we go for ELBOW PLOT
- We will have inertia for it.
- Inertia = Within the sum of squared distance/distance between two clusters.
- Inertia is the opposite of how tight a cluster is and we need a low value of inertia.

- Will plot the inertia values in a graph. The graph where it takes a sharp value, the value of k at that portion is called the best value of K.
- Randomness will affect my K-means clustering algorithm. As per the randomness, we will get different clusters every time.

Drawback of the K-Means Algorithm

- It clusters only in a spherical structure.

K-Means Algorithm ML Model

https://github.com/Adharsh0001/Machine-Learning/blob/main/K_means_clustering.ipynb

6. Reference Books

INTRODUCTION TO STATISTICAL LEARNING

ELEMENTS OF STATISTICAL LEARNING

PATTERN RECOGNITION