Machine Learning
Classification Models

Submitted by: Noor Ayesha

Date: 08/08/2022

IESEG School of Management

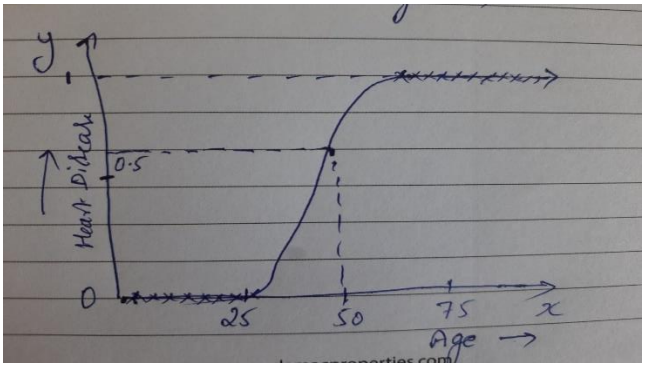
Logistic Regression:

The logistic Regression is a machine learning model used for binary classification in which the output predictions can only take one of the two possible values it can be either true or false.

Some of the real-world applications where the logistic regression can be used to classify data for example are 1) if the email is spam or not 2) if the transaction is fraud or not 3) it can be also used to find out if the person is having a disease or not 4) to find out if the tumour is a malignant or not.

How it Works:

This can be very well understood with the help of the following example, let's say our goal is to predict if the person is having heart disease or not based on our input feature which is age (for simplicity I have taken a very simple data set where we are only considering only one input feature that is the age)



Hence in order to make new predictions based on this given dataset, we have to draw a curve that fits appropriately on the dataset points. This curve is called the sigmoid curve.

The sigmoid curves take 0.5 as the value when our x is 0 and whenever x is very large it takes roughly 1 and when the x is very low it takes almost close to zero as the value. And the formula for this curve for one input feature is shown in the image below.

In our example if the probability is greater than 0.5 then we will classify it as 1 means the person has a heart disease if this probability is less than 0.5 then we will classify it as 0. The dotted line in the image above at the age of 50 indicates that the person at this has a higher than 50% of chance of getting heart disease.

The logistic regression formula is:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Where: x is the input value, y is the predicted output, b0 is the bias or intercept term, b1 is the coefficient for the single input value (x)

In the equation, each column from input data has an associated b coefficient (a constant real value).

Advantages:

1) highly comprehensible
2) We will get well-calibrated predicted probabilities in the output.
3) Model development and predictions are quick.
4) There is no need to scale features.

5) can perform effectively even with few observations

Dis-Advantages:

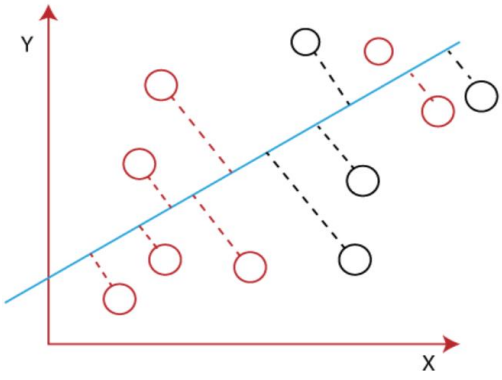
1) Given that it is extremely susceptible to noise, feature engineering is necessary.
2) Performance may be impacted by the correlation of the independent features.
3) Logistic regression should not be considered if there is less data than features because this could result in overfitting.
4) Logistic regression's linear decision surface makes it unable to handle non-linear issues.

Linear Discriminant Analysis:

This is used as a pre-processing step in Machine Learning algorithms, and it is a supervised dimensionality reduction technique used to solve more than 2 classes of classification problems. Logistic Regression fails when dealing with many classification issues involving clearly defined classes. but, LDA manages these fairly effectively.

How it Works: The main goal of LDA is to project variables from higher dimension space to lower dimension space. We can quickly convert a 2 and 3-dimensional graph into a 1-dimensional plane using LDA.

If we take the situation where we need to efficiently categorize two classes in a 2-D plane with an X-Y axis. The two groups of data points may be completely separated using a straight line in LDA. However, LDA enables us to split an X-Y axis into two separate axes with a straight line, then project data into the new axis. As a result, we can shrink the 2-D plane to 1-D and optimize the gap between these classes.



Linear Discriminant Analysis use the following standards to develop a new axis:

- It increases the gap between the means of the two classes.
- It reduces variance within the specific class.

LDA creates a new axis using the two principles outlined above to maximize the distance between the means of the two classes and minimize the variation within each.

Advantages:

1. Fast classification
2. Easy to implement
3. Linear decision boundary

Disadvantages:

1. It's a complex algorithm
2. if the distribution of our data is significantly non-Gaussian, the LDA might not perform very well.

Decision Tree:

This is a type of supervised ML algorithm in which data is continuously split based on certain parameters. Every branch of the tree symbolizes a potential choice, event, or response.

The approach to building a decision tree is as follows:

- The tree is built using the top-down divide-and-conquer method.
- To build a tree, we need a root/parent node (a variable that best classifies the training data) and then split the data based on the selected attributes.
- Test the selected attributes based on some algorithms (e.g., entropy)
- Repeat this for each branch

Important terms in Decision tree:

Entropy: It is a metric for the dataset's randomness or unpredictability.
Information gain: It is the measurement of the entropy's decline following the dataset's division.
Leaf Node: The classification or decision is carried by the leaf node.
Decision Node: There will be two or more branches from this decision node.
Root Node: The topmost choice is referred to as the root node.

Stopping conditions for the split:

- A particular node's data are all members of the same class.
- There are no more variables available for splitting.
- The number of observations per nod decreases

Advantage of Decision Tree:

Simple to Understand: Even those without an analytical background can easily understand the decision tree output.

Useful in Data Exploration: One of the quickest methods for determining the most important factors and the relationship between two or more variables is to use a decision tree.

Managing outliers: It is not significantly impacted by outliers or missing values. It supports both Numeric and categorical data types.

Non-Parametric Approach: The decision tree is regarded as a non-parametric approach. Decision trees, therefore, don't make any generalizations about the space distribution or the design of the classifier.

Dis-Advantages of Decision Trees:

Overfitting: One of the most prevalent practical challenges for decision tree models is overfitting.

Continuous variable restriction: While utilizing a continuous range of numerical variables, the Decision tree sheds when several categories of discrete variables are used.

Random Forest:

The supervised ML algorithm Random Forest, also known as Random Decision Forest, is more frequently applied to classification and regression applications. It is a technique that builds several Decision Trees during the training stage. The final choice made by the random forest is by using the majority vote of the trees.

How it Works: Each decision tree in the ensemble that makes up the random forest method is built of a data sample taken from a training set with a replacement known as the bootstrap sample. Test data are reserved for one-third of the training sample, often known as the out-of-bag (oob) sample. The dataset is subsequently given a second randomization injection by feature bagging, increasing dataset diversity and decreasing decision tree correlation. The prediction will be determined differently depending on the type of issue. The individual decision trees will be averaged for the regression job, and for the classification task, the predicted class will be determined by a majority vote, or the most common categorical variable. The prediction is then finalized by cross-validation using the oob sample.

Advantages:

1) No overfitting: The use of multiple trees reduces the risk of overfitting
2) Training time is less
3) High accuracy: Runs efficiently on a large database
4) For large data, it produces highly accurate predictions
5) Estimates missing data
6) Random Forest can continue to be accurate even when a significant amount of data is missing.

Dis-Advantages:

1) More time is required to train the model
2) Interpretation is complex
3) Computationally expensive
4) More utilization of memory

Boosting Tree:

This algorithm overcomes the drawbacks of decision tree and random forest algorithm. Boosting corrects mistakes made by earlier decision trees. In boosting, new trees are created by taking into account the mistakes made by trees in earlier rounds. As a result, new trees are grown one after the other. Each tree depends on the one before it. Sequential learning is the name for this method of learning.

Key Points:

- This method has strong predictive power, but even less interpretability than forests.
- Each successive tree uses the residuals of the previous tree.

- Boosted trees are the trees that have undergone the boosting process.
- It has even more hyperparameters than forests to control model building Learning rate.
- The process of boosting is iterative. Each tree depends on the one before it. As a result, it is challenging to parallelize the boosting algorithm training process. There will be more training time. This is boosting algorithms' primary flaw.

Advantages:

- As an ensemble model, boosting has a simple-to-read and understanding algorithm, making it simple to interpret its predictions.
- Through the application of its clone methods, including bagging, random forests, and decision trees, the prediction capability is effective.
- Boosting is a strong technique that easily reduces overfitting.

Disadvantages:

- Boosting has the drawback of being sensitive to outliers because every classifier is required to correct the mistakes made by the predecessors. As a result, the technique is overly reliant on outliers.
- The method's near impossibility to scale up is another drawback. This is because it is challenging to streamline the process because every estimator rests its accuracy on the prior predictions.
- If working with a large dataset then computation might be expensive.

Support Vector Machine:

A supervised learning technique called SVM sorts data into one of two groups after looking at it. Additionally, it is among the most reliable Machine Learning algorithms.

SVM's major goal is to establish the optimum decision boundary (also known as a hyperplane) that can divide n-dimensional space into classes so that we may quickly classify fresh data points in the future. SVM selects the extreme points (or vectors) that aid in constructing the hyperplane. Support vectors are used to describe these extreme circumstances, which is how the method got its name, Support Vector Machine.

Important terms in SVM:

Support Vector: The closest data point to the hyperplane is referred to as a support vector. The use of these support vectors increases the classifier's margin.

Hyperplane: The optimal decision boundary that aids in classifying the data points is known as a hyperplane.

Margin: The Margin is the separation of the vectors from the hyperplane.

Advantages:

1) Regularization parameters i.e. lambda helps to figure out if there is bias or over fitting of the data and naturally avoids bias and overfitting.
2) It functions admirably and has a distinct margin of separation.
3) In high-dimensional spaces, it works well.

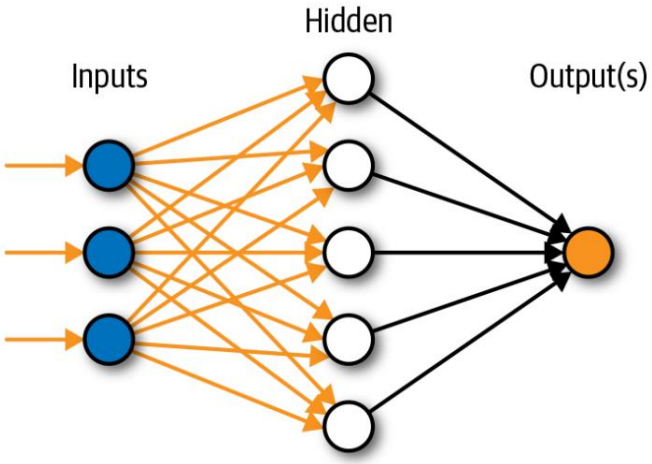
4) In situations where there are more dimensions than samples, it is effective.
5) It is also memory efficient because it only needs a small fraction of training points for the decision function (known as support vectors).

Disadvantages:

1) Since more training time is needed when we have a huge data set, it doesn't perform well.
2) When the data set includes additional noise, such as when the target classes are overlapping, it also doesn't perform very well.
3) Probability estimates are not directly provided by SVM; instead, they are computed via a costly five-fold cross-validation.

Neural Network :

Typically, artificial neural networks are referred to as human brain neural networks (NN). In reality, neural networks are multi-layer Perceptron. The basic building block of multi-layered neural networks is the perceptron (Shown in below figure).



The input data layer (Blue) is processed against a hidden layer (White) to produce the final output (Orange). Each layer is made of multiple nodes or neurons. A weighted sum of the blue input node values is represented by each white node in the hidden layer. The output node(orange) is a weighted sum of the hidden layer nodes(white). Hence the final output is the linear combination of its inputs. Multiple hidden layers can exist in a single neural network.

How an Individual Neuron works:

- A neuron reads inputs and does some mathematical operations and produces one output. The weight of each input is multiplied as: x1 = x1*w1, x2 = x2*w2.
- The weighted inputs are then added collectively with bias b: (x1*w1)(x2*w2)
- At the end, an activation function is applied to the sum: y=f(x1*w1+x2*w2+b).

Activation function:

We can immediately introduce nonlinearity into a model of a nonlinear problem. Each hidden layer node can be passed via a nonlinear function. Each node's value in the Hidden Layer will undergo a

nonlinear function transformation before being transferred to the weighted sums of the following layer. This is called the activation function.

Key Points about ANN and its Application:

1. Neural networks are a set of algorithms that simulate how a human brain operates in order to find connections among enormous volumes of data.
2. As a result, they frequently mimic the synapses and connections between neurons seen in the brain.
3. They are used in a range of financial services applications, including forecasting, market research, fraud detection, and risk assessment.
4. Deep learning algorithms use neural networks that have multiple process layers, or "deep" networks

Advantages of ANN:

1. Efficiency: The training data noise is relatively well-tolerated by ANN learning techniques. Errors in the training examples are possible, but they won't affect the results.
2. Following ANN training, the data may still produce output with missing information. The degree to which the performance is lost here depends on how critical the missing data is.
3. A neural network is made to continuously learn and produce better results. Once trained, the system may generate output without requiring complete inputs. The program or applications become more user-friendly as they are used.
4. Instead of being kept in a database, information is kept on the entire network, much like in conventional programming. The network continues to operate even if a few pieces of information in one location disappear.

Dis-Advantages of ANN:

1. Dependency on Hardware: Due to the topology of artificial neural networks, parallel processing-capable processors are required. The equipment's actualization is therefore dependent on this.
2. The primary issue with ANNs is the network's mysterious operation: When ANN offers a perplexing solution, it doesn't explain why or how. This makes the network less trustworthy.
3. Guarantee of appropriate network architecture: For choosing the structure of artificial neural networks, there is no set rule. Experience, experimentation, and trial-and-error are required to determine the ideal network structure.
4. Its difficult to estimate how long the network will last.
5. It also heavily depends on data because it acts according to the data fed to the machine The more the data is used, the more accurate results are generated.

References:

https://corporatefinanceinstitute.com/resources/knowledge/other/boosting/

https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning

https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b

https://uaps2015.princeton.edu/papers/150738

http://www.nhn.ou.edu/~abbott/REU/Doctor_finaltalk.pdf

https://www.javatpoint.com/linear-discriminant-analysis-in-machine-learning

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6583308/

https://www.insightbig.com/post/machine-learning-logistic-regression-with-python

https://web.archive.org/web/20220209211226/https://www.stoodnt.com/blog/common-machine-learning-algorithms-must-know/

https://link.springer.com/article/10.1007/s13202-022-01492-3

https://github.com/kalperen/MachineLearningGuide

https://www.ibm.com/cloud/learn/random-forest#toc-how-it-wor-nIcKpuXp

https://towardsdatascience.com/introduction-to-boosted-trees-2692b6653b53#:~:text=Boosting%20transforms%20weak%20decision%20trees,known%20as%20ensemble%20meta%2Dalgorithms.

https://corporatefinanceinstitute.com/resources/knowledge/other/boosting/

https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/#

https://towardsdatascience.com/multi-layer-neural-networks-with-sigmoid-function-deep-learning-for-rookies-2-bf464f09eb7f