Performance Evaluation of Various Classification Algorithms

Shafali Deora
Amritsar College of Engineering & Technology, Punjab Technical University

---***---

Abstract - *Classification is a technique in which the data is categorized into 2 or more classes. It can be performed on both structured/ linear as well as unstructured/ non-linear data. The main goal of the classification problem is to identify the category of the test data. The 'Heart Disease' dataset has been chosen for study purpose in order to infer and understand the results from different classification models. An effort has been put forward through this paper to study and evaluate the prediction abilities of different classification models on the dataset using scikit-learn library. Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbor are some of the classification algorithms chosen for evaluation on the dataset. After analysis of the results on the considered dataset, the evaluation parameters like confusion matrix, precision, recall, f1-score and accuracy have been calculated and compared for each of the models.*

Key Words: Pandas, confusion matrix, precision, logistic regression, decision tree, random forest, Naïve Bayes, SVM, sklearn.

1. INTRODUCTION

Machine learning became famous in 1990s as the intersection of computer science and statistics originated the probabilistic approaches in AI. Having large-scale data available, scientists started building intelligent systems capable of analysing and learning from large amounts of data.

Machine learning is a type of artificial intelligence which provides computer with the ability to learn without being explicitly programmed. Machine learning basically focuses on the development of the computer programs that can change accordingly when exposed to new data depending upon the past scenarios. It is closely related to the field of computational statistics as well as mathematical optimization which focuses on making predictions using computers. Machine learning uses supervised learning in a variety of practical scenarios. The algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. The model is first trained by feeding the actual datasets to help it build a mapping between the dependent and independent variables in order to predict the accurate output.
Classification belongs to the category of supervised learning where the computer program learns from the data input given to it and then uses this learning to classify new observations. This approach is used when there are fixed number of outputs. This technique categorizes the data into a given number of classes and the goal is to identify the category or class to which a new data will fall.

Below are the classification algorithms used for evaluation:

1.1 Logistic Regression

Logistic Regression is a classification algorithm which produces result in the binary format. This technique is used to predict the outcome of a dependent variable or the target wherein the algorithm uses a linear equation with independent variable called predictors to predict a value which can be anywhere between negative infinity to positive infinity. However, the output of the algorithm is required to be class variable, i.e. 0 or 1. Hence, the output of the linear equation is squashed into a range of [0, 1]. The sigmoid function is used to map the prediction to probabilities.

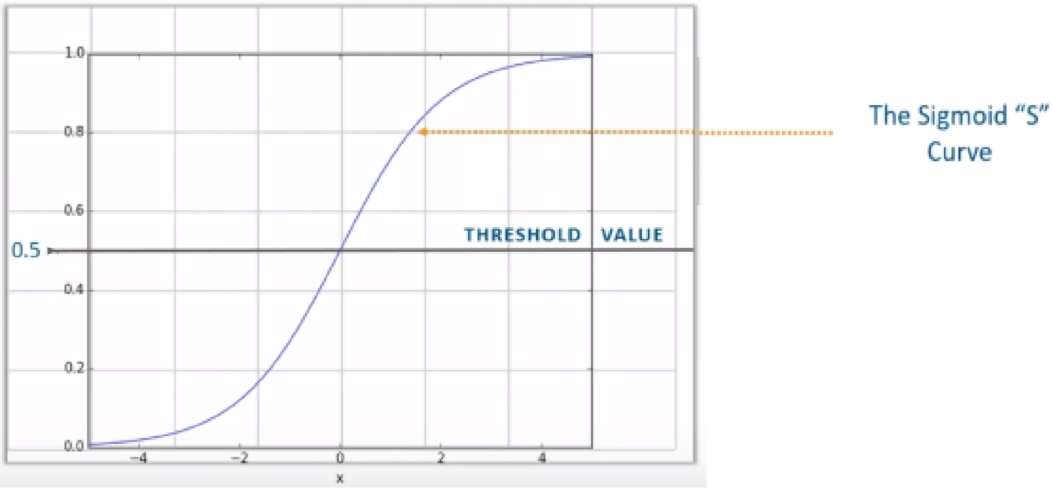
Mathematical representation of sigmoid function:

$$F(z) = \frac{1}{1+e^{-z}}$$

Where

F(z): output between 0 and 1
z: input to the function

Graphical representation of sigmoid function:

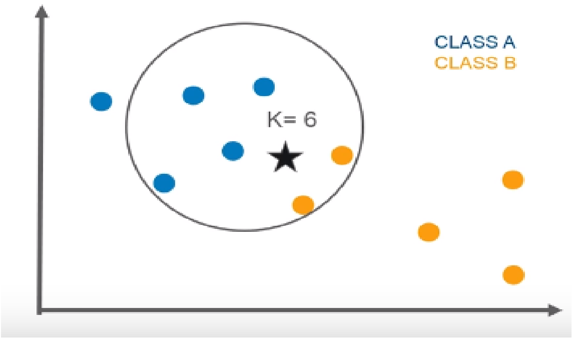


1.2 K- Nearest Neighbor:

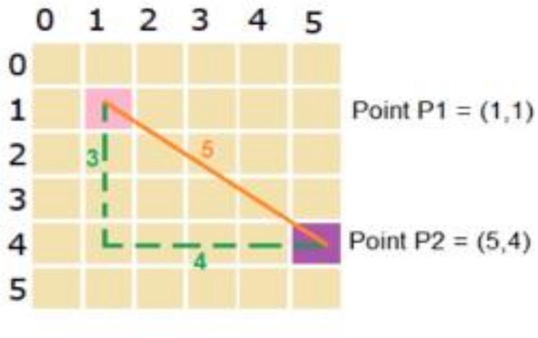
K nearest neighbour is a supervised learning technique that stores all the available cases and classifies the new data based on a similarity measure. It uses the least distance measure in order to find its nearest neighbours where it looks at the 'K' nearest training data points for each test data point and takes the most frequently occurring class and assign that class to the test data.

K= number of nearest neighbours

Let's say for K=6, the algorithm would look out for 6 nearest neighbours to the test data. As class A forms the majority over class B, hence class A would be assigned to the test data in the below example.



The distance between two points can be calculated using either Euclidean distance which is the least distance between two points or Manhattan distance which is the distance between the points measured along the axis at right angle.

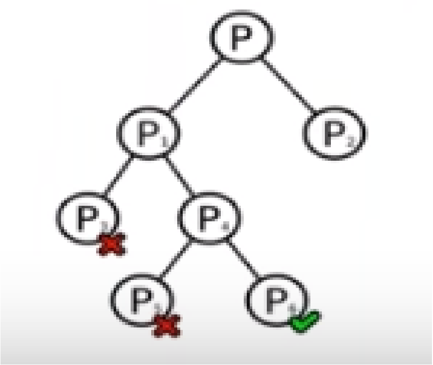


Euclidean Distance = $\sqrt{(5-1)^2 + (4-1)^2}$ = 5

Manhattan Distance = |5 – 1| + |4 – 1| = 7

1.3 Decision Tree

Decision Tree is the tree representation of all the possible solutions to a decision based on certain conditions. The tree can be explained by two entities- decision nodes and leaves where leaf nodes are the decisions or the final outcomes and the decision nodes are where the data is split. There is a concept of pruning in which the unwanted nodes are removed from the tree.



The CART (Classification and Regression Tree) algorithm is used to design the tree which helps choosing the best attribute and deciding on where to split the tree.
This splitting is decided on the basis of these factors:

Gini Index: Measure of impurity used to build decision tree in CART.

Information Gain: Decrease in the entropy after a dataset is split on basis of an attribute is information gain and the purpose is to find attribute that returns the highest information gain.

Reduction in Variance: Algorithm used for continuous target variables and split with lower variance is selected as the criteria for splitting.

Chi Square: Algorithm to find the statistical significance between sub nodes and parent nodes.

1.4 Random Forest

The primary weakness of Decision Tree is that it doesn't tend to have the best predictive accuracy, partially because of high variance, i.e. different splits in the training data can lead to different results. Random forests uses a technique called bagging in order to reduce the high variance.

Here we create an ensemble of decision trees using bootstrapped samples of the training set which is basically sampling of training set with replacement. These decision trees are then merged together to get a more accurate and stable prediction. Here a random subset of features is chosen by the algorithm for splitting the node. The larger number of trees will result in better accuracy.

1.5 Support Vector Machine (SVM):

This is a type of technique that is used for both classification as well as regression problems however, we will discuss its classification aspect. In order to classify data, it makes use of hyperplanes which act like decision boundaries between the various classes and make segments in such a way that each segment contains only one type of data. SVM can also classify non-linear data using the SVM kernel function which is basically transforming the data into another dimension in order to select the best decision boundary. The best or optimal line that can separate the classes is the line that has the largest distance between the closest data points and itself. This distance is calculated as the perpendicular distance from the line to the closest points.

1.6 Naïve Bayes:

Bayes theorem tells us the probability of an event, given prior knowledge of related events that occurred earlier. It is a statistical classifier. It assumes that the effect of a feature on a given class is independent of other features and due to this assumption is known as class conditional independence. Below is the equation:

P (A|B) = (P (B|A) · P (A))/P (B)
P (A|B): the probability of event A when event B has already happened (Posterior)
P (B|A): the probability of event B when event A has already happened (Likelihood)
P (A): probability of event A (Prior)
P (B): probability of event B (Marginal)

<u>Proof of Bayes Theorem:</u>

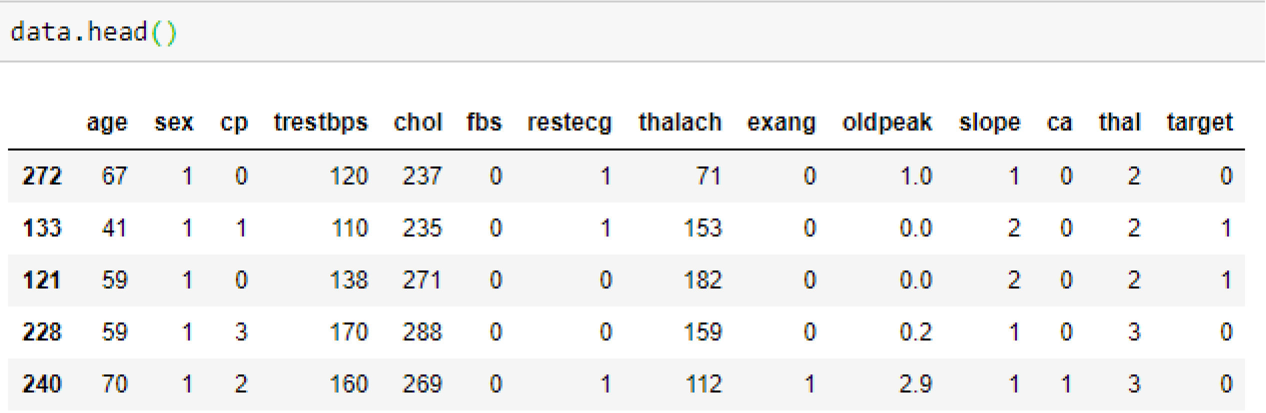
P (A|B) = P (A intersection B)/P (B)
P (B|A) = P (B intersection A)/P (A)
Since P (A intersection B) = P (B intersection A)
P (A intersection B) = P (A|B) * P (B) = P (B|A) * P (A)
P (A|B) = (P (B|A) · P (A))/P (B)

2. METHODOLOGY:

2.1 Dataset:

The data has one csv file from which the training and testing set will be formed. The training sample will be used to fit the machine learning models and their performance will be evaluated on the testing sample. The main objective for the models will be to predict whether a patient will have a heart disease or not.

Below is the head of the dataset:



```
data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
272	67	1	0	120	237	0	1	71	0	1.0	1	0	2	0
133	41	1	1	110	235	0	1	153	0	0.0	2	0	2	1
121	59	1	0	138	271	0	0	182	0	0.0	2	0	2	1
228	59	1	3	170	288	0	0	159	0	0.2	1	0	3	0
240	70	1	2	160	269	0	1	112	1	2.9	1	1	3	0

The target variable is in binary form which is the dependent variable that needs to be predicted by the models.

Data Dictionary:

- Age
- Sex
- Chest pain type (4 values)
- Resting blood pressure
- Serum cholesterol in mg/dl
- Fasting blood sugar > 120 mg/dl
- Resting electrocardiographic results (values 0,1,2)
- Maximum heart rate achieved
- Exercise induced angina
- Old peak = ST depression induced by exercise relative to rest
- Slope of the peak exercise ST segment
- Number of major vessels (0-3) coloured by fluoroscopy
- Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

2.2 Evaluation Parameters:

2.2.1 Confusion Matrix:

In order to evaluate how an algorithm performed on the testing data, a confusion matrix is used. The rows of this matrix correspond to what machine learning algorithm predicted and the columns correspond to the known truth.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

2.2.2 Precision:

It is calculated by dividing the true positives by total sum of the true positives and the false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

2.2.3 Recall:

Recall is also known as sensitivity. It is the total number of true positives upon total sum of the true positives and the false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

2.2.4 Accuracy:

It is the total sum of the true positives and the true negatives divided by overall number of cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2.3 Experimental Results:

Classifier	**Class**	**Classification Report**				**Confusion Matrix**
		Precision	Recall	f1-score	Accuracy	
Logistic Regression	0	0.91	0.65	0.76	78.02%	[[32 17] [3 39]]
	1	0.7	0.93	0.8		
KNN	0	0.76	0.51	0.61	64.83%	[[25 24] [8 34]]
	1	0.59	0.81	0.68		
Decision Tree	0	0.67	0.63	0.65	63.73%	[[31 18] [15 27]]
	1	0.6	0.64	0.62		
Random Forest	0	0.89	0.67	0.77	78.02%	[[33 16] [4 38]]
	1	0.7	0.9	0.79		
SVM	0	0	0	0	46.15%	[[0 49] [0 42]]
	1	0.46	1	0.63		
Naïve Bayes	0	0.9	0.71	0.8	80.21%	[[35 14] [4 38]]
	1	0.73	0.9	0.81		

3. CONCLUSION:

In this effort we tested 6 different types of supervised learning classification models on the Heart Disease dataset. For the train-test split, the test size of .3 was chosen. The above results show that the Naïve Bayes Classifier got the most correct predictions and classifying only 18 samples wrong, probably because the dataset favours the conditionally independent behaviour of the attributes assumed by this classifier whereas the SVM performed the worst with just over 46%.

4. REFERENCES:

[1]. Suzumura, S., Ogawa, K., Sugiyama, M., Karasuyama, M, & Takeuchi, I., Homotopy continuation approaches for robust SV classification and regression, 2017.

[2]. Sugiyama, M., Hachiya, H., Yamada, M., Simm, J., & Nam, H., Least-squares probabilistic classifier: A computationally efficient alternative to kernel logistic regression, 2012.

[3]. Kanu Patel, Jay Vala, Jaymit Pandya, Comparison of various classification algorithms on iris datasets using WEKA, 2014.

[4]. Y.S. Kim, "Comparision of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size", 2008.