Linear Regression: Simple Technique for Inferential Statistics using Pandas

Swayanshu Shanti Pragnya[1] Shakti Chaturvedi[2]
[1]Department of Computer Science & Engineering
[1]CUTM, Hyderabad, India – 7500033 [2]CDAC Post Graduation Diploma, Pune, India

Abstract— Inferential statistics is otherwise known as predictive statistics that meant to do prediction using simple statistics. Statistics is the efficient and most used dogma for the world of data analysis. Any sort of analysis is incomplete without simple linear equations as, all the implementation needs a mathematical derivation for better understanding which leads to prior data visualization. Predictive analysis is not new but still requires human interface for facing more critical problems like every human does in their day to day life. Before prediction, data collection, identification, segregation are primarily concerned for data analysis. So for that, machine learning algorithms are useful to solve such problems. This paper is about the collation between linear and logistic regression by using Pandas, implementation of SVM (Support Vector Machine) and result analysis using confusion matrix.

Key words: Data Analysis, Linear Regression, Logistic Regression, SVM, Prediction, Confusion Matrix

I. Introduction

Analysis of data is required to explore the attributes more thoroughly. By the acknowledgement of the data sets we can decide how far it is required to predict for the future. As by the end we need the prediction and again how far that prediction is sccurate. But only analysing a data is not sufficient when it comes to analysis that too by using statistics only. So at this point predictive analysis comes which is nothing but a part of inferential statistics. Here we try to infer any outcome based on analysing patterns from previous data just to predict for the next dataset.

When it comes to prediction first buzzword came i.e. machine learning. So machine learning combine’s statistical analysis and computer science for the prediction purpose. Machine learning also introduced to self-learning process from particular data. This learning reduce the gap between computer and statistics. Large amount of data prediction can be possible by human interaction as a human brain can analyse the situation with various aspects. Here the partition in algorithms occur i.e. Supervised (used for labelled data) and unsupervised (no labelled data for learning) algorithm.

Here we have introduced supervised learning algorithms i.e. regression and classification algorithms.

In the first section we have explained about linear regression, logistic regression, their differences, SVM, hyper plane. In the second part we have shown the complete code to perform logistic regression and SVM classification. IN the third part we have calculated the result we got in terms of confusion matrix.

A. *Need of Regression*

Both classification and regression are most frequently used Data mining techniques. Regression comes into eye view when we need to predict dependant variable which has relation with other data.

Example- In our given titanic data set the number of survived passenger is somehow dependent upon which class the passenger is travelling as well as which cabin they were sitting. So for predicting which person survived is completely dependent upon all these attribute so here we will use regression technique to predict.

As the name itself defines Classification is all about the categorization of data based on condition.

Support Vector Machine algorithm can give high accuracy when the data set is small and as well as less missing values in the given dataset.

II. Languages

A. *Python*

Open source as well as easy to understand, syntax is easy for beginners and used for statistical data analysis.

B. *Pandas*

Highly used library for data analysis. Easy to understand. Open source as well as easy to use in data manipulation.

C. *Numpy*

Used for scientific computing with python.

D. *Matplotlib*

It is a mathematical extension from numpy as well as primarily used for plotting graphs.

III. Method

Linear and logistic regression both are used for prediction purpose. But what’s the difference is much more important to know. These are the following attributes to know the difference between these two regression algorithms.

A. *Outcome after Regression*

In linear regression the result we got is continuous whereas logistic regression has limited number of possible values.

B. *Dependent Variable*

Logistic regression is used for the instance of true/false, yes/no, 0/1 which are categorical in nature but linear regression is used in case of continuous variable like number, weight, height etc.

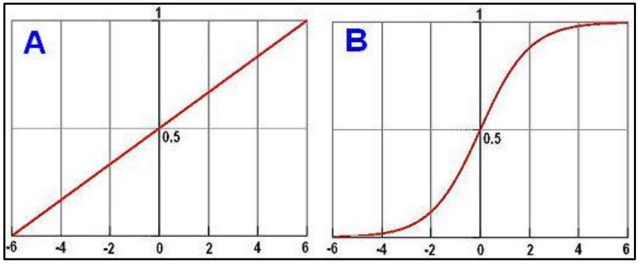


Fig. 1: Linear & Logistic Regression

1) Equation

Linear regression gives a linear equation in the form of Y = aX + B, means degree 1 equation But, logistic regression gives curved equation which is in form of Y = e^X/1 + e^-X

2) Minimization of error

Linear regression (LR) uses ordinary least squares method which minimizes the error and Logistic Regression use least square method which reduces the error quadratically[6].

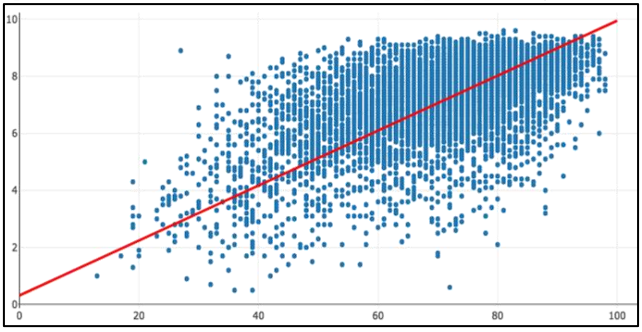


Fig. 2: Linear Regression with Nearest Data Set

IV. SUPPORT VECTORS

These are the vectors (magnitude and direction) which take support for classification purpose near to the hyper plane.

A. Hyper Plane

Generally plane can be formed in 2 dimensions but more than 2D it is called hyper plane. Though support vectors are drawn in more than two dimensions that's why it splits data through hyper plane s[2].

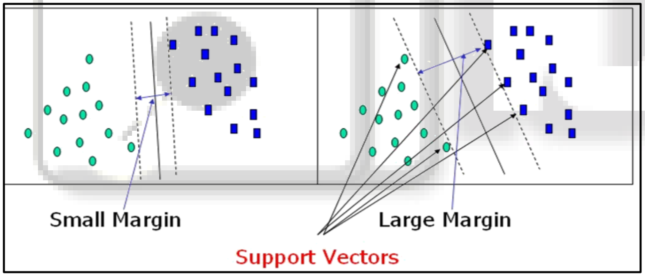


Fig. 3: SVM

In the above example we saw the set of blue and red dots separated but in the next picture the splitting is done via hyper plane to segregate data set in two different clusters.

B. Way to Find Right Hyper Plane

Nearest data point and hyper plane distance is known as margin. So when the margin is less the chance of correct segregation is more.

C. Support Vector Machines

1) Pros

Accuracy is more on smaller as well as cleaner data set.

2) Cons

Not appropriate for larger data set as lesser accuracy.

3) Uses

Primarily used for text classification, detecting spam, image recognition (colour based classification) [3], hand writing identification and sentiment analysis etc.

V. CODE & EXPLANATION

1) Step 1: Irrespective of any regression or classification algorithm initially need to import libraries like pandas, numpy, matplotlib, seaborn and from sklearn linear, logistic regression and svm module.
2) Step 2: Loading data in csv file format as the data set has been taken from Kaggle Tianic completion. Where train and test data set were taken for regression.
3) Step 3: Select required columns in X (mostly independent variable) and in Y take dependant column as per here number of passenger survived is dependant that's why taken in Y.
4) Step 4: Data cleaning and fill null values to prepare data set completely.[5]
5) Step 5: For knowing which column is dependant more to the output column we need to plot graphs by using regression type.

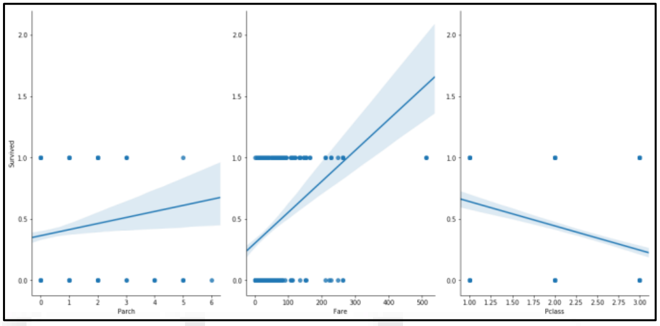


Fig. 4: Correlation in between Columns

6) Step 6: Split the data set into train and test by using sklearn.
7) Step 7: Finally call regression function whether it is linear, svm.

Complete code for Linear regression:

```
import matplotlib.pyplot as plt.
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.linear_model import LinearRegression
data = pd.read_csv('titanic_train.csv')// Reading data
required_cols = ['Fare','Sex','Age','Pclass']
X = data[required_cols]
X = data[['Fare','Sex','Age','Pclass']]
Y = data['Survived']
Y = data.Survived
X['Pclass'].isnull()
#fillna value
X['Age'] = X['Age'].fillna(X['Age'].median())
X.Sex[X.Sex == 'male'] = 0
X.Sex[X.Sex == 'female'] = 1
X.Sex[X.Sex == 'NaN'] = 2
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size
=0.33,random_state=42)
from sklearn import svm
clf=svm.LinearSVC()
clf.fit(X_train,Y_train)
o/p-
LinearSVC(C=1.0, class_weight=None, dual=True,
fit_intercept=True,  intercept_scaling=1,
loss='squared_hinge', max_iter=1000, multi_class='ovr',
penalty='l2', random_state=None, tol=0.0001, verbose=0)
clf.predict(X_test[0:1]
clf.predict(X_test[0:20])
o/p- array([0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1,
........................................1, 0, 1], dtype=int64)
```

Before calculating the accuracy we need to build a confusion matrix to know all the attributes like precision, recall, number of positive outcome and finally model accuracy.

For SVM

```
clf = SVC()
scoring = 'accuracy'
score = cross_val_score(clf, train_data, target, cv=k_fold,
n_jobs=1, scoring=scoring)
print(score)
O/P-
[0.83333333 0.80898876 0.83146067 0.82022472
0.84269663 0.82022472 0.84269663 0.85393258
0.83146067 0.86516854]
round(np.mean(score)*100,2)
O/P-
83.5
clf = SVC()
clf.fit(train_data, target)
test_data = test.drop("PassengerId", axis=1).copy()
prediction = clf.predict(test_data)
```

VI. Result Analysis

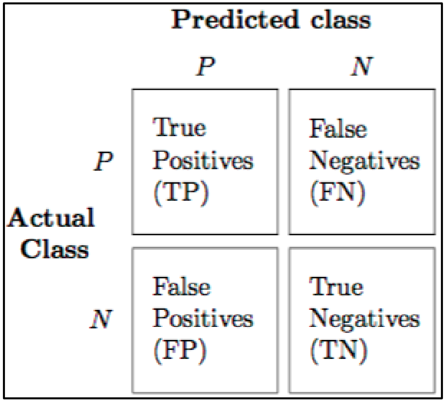
A. *Confusion Matrix*

It contains information about actual and predicted classifiers done by classification system. See Fig 7 for the example of confusion matrix. Here we calculate

Recall = TP / (TP+FN)

Precision = TP / (TP+FP)

Precision is about calculating how far our model prediction is accurate in terms of positivity. After execution we got the result that precision is 80% that means the positivity or accuracy of correct prediction is more.



		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Fig. 7: Confusion Matrix Example

```
#Higher value = better classifier
Recall = TP / float(FN + TP)
print(metrics.recall_score(Y_test, y_pred_class))
0/p-
0.408333333333
0.408333333333
Precision = TP / float(TP + FP)\
print(precision)
print(metrics.precision_score(Y_test, y_pred_class))
o/p-
0.809523809524
0.809523809524
```

VII. Conclusion

Here we have studied the difference in linear regression, logistic regression and SVM. We have executed the code by using Pandas language and got the output successfully. At the end we have calculated SVM model accuracy and got the precesion as 80% by using SVC module from sklearn. As the objective was for basic information gathering and code execution which is computed with identifiable accuracy. We have also performed confusion matrix for result analysis and got the positive prediction rate is more than the false predicted result.

References

[1] Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms Tryambak Chatterjee. International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-6)

[2] A Comparative Analysis on Linear Regression and Support Vector Regression Kavitha S Assistant Professor Computer Science and Engineering Bannari Amman Institute of Technolgy Sathyamangalamkvth.sgm@gmail.com

[3] An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain Park, Hyeoun-Ae College of Nursing and System Biomedical Informatics National Core Research Center, Seoul National University, Seoul, Korea

[4] Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology, 54(10), 979-985. Bewick, V., Cheek, L., & Ball, J. (2004).

[5] Statistics review 13: Receiver operating characteristic curves. Critical Care (London, England), 8(6), 508512. http://dx.doi.org/10.1186/cc3000