Logistic Regression

- In statistics, the logistic model is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model.
- In Logistic Regression, the dependent variable is binary (a categorical variable that has two values such as “yes” and “no”) rather than continuous and it can also be applied to ordered categories (ordinal data), that is, variables with more than two ordered categories.

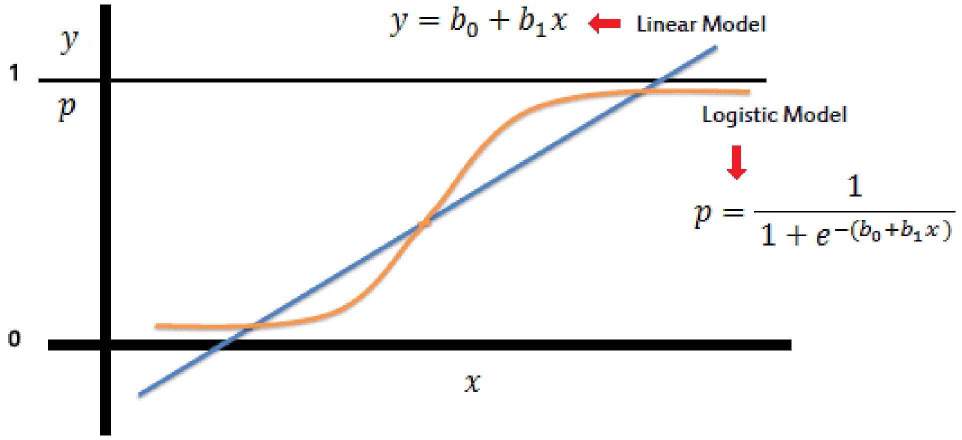
Why use logistic regression rather than ordinary linear regression?

Logistic Regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical).

A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



The Logistic Curve

The logistic curve relates the independent variable, X, to the rolling mean of the DV (dependent variable), P (). The formula to do so may be written either.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Or

$$P = \frac{1}{1+e^{-(a+bX)}}$$

where P is the probability of a 1 (the proportion of 1s, the mean of Y), e is the base of the natural logarithm (about 2.718) and ‘a’ and ‘b’ are the parameters of the model.

Derivation

In logistic regression, the dependent variable is a logit, which is the natural log of the odds, that is,

$$odds = \frac{P}{1-P}$$

$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find
logit(P) = a + bX,

$$\ln(\frac{P}{1-P}) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

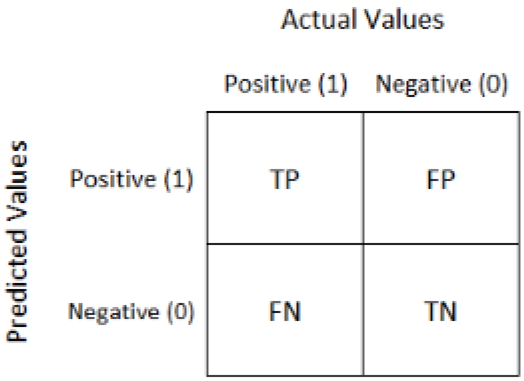
Model Evaluation Matrix for Classification

The different ways are as follows:

- Confusion matrix
- Accuracy
- Precision
- Recall
- Specificity
- F1 score
- Precision-Recall or PR curve
- ROC (Receiver Operating Characteristics) curve
- PR vs ROC curve

Confusion Matrix:

It is a performance measurement for machine learning classification problems where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

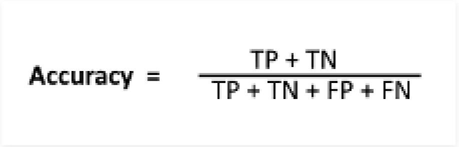
Definition of the terms:

Positive (P) : Observation is positive (for example: is an apple).
Negative (N) : Observation is not positive (for example: is not an apple).
True Positive (TP) : Observation is positive, and is predicted to be positive.
False Negative (FN) : Observation is positive, but is predicted negative.
True Negative (TN) : Observation is negative, and is predicted to be negative.
False Positive (FP) : Observation is negative, but is predicted positive.

Classification Rate/Accuracy:

Classification Rate or Accuracy is given by the relation:

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.



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

Recall/Sensitivity/True Positive Rate:

Recall can be defined as the ratio of the total number of correctly classified positive examples divided to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN).

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

Precision:

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP).

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

F-measure/F-stats/F1 Score:

Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.
The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

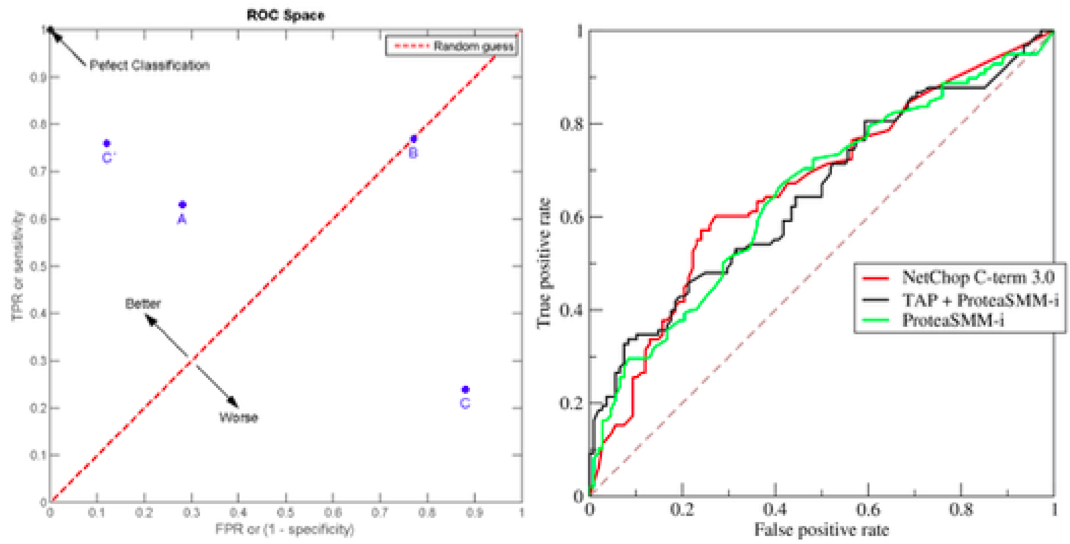
Specificity:

Percentage of negative instances out of the total actual negative instances. Therefore, the denominator (TN + FP) here is the actual number of negative instances present in the dataset. It is similar to recall but the shift is on the negative instances. Like finding out how many healthy patients were not having cancer and were told they don't have cancer. Kind of a measure to see how separate the classes are.

$$\frac{TN}{TN + FP}$$

ROC Curve:

ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases. ROC AUC is just the area under the curve, the higher its numerical value the better.



Some Theoretical Concepts:

Assumptions of Logistics Regression:

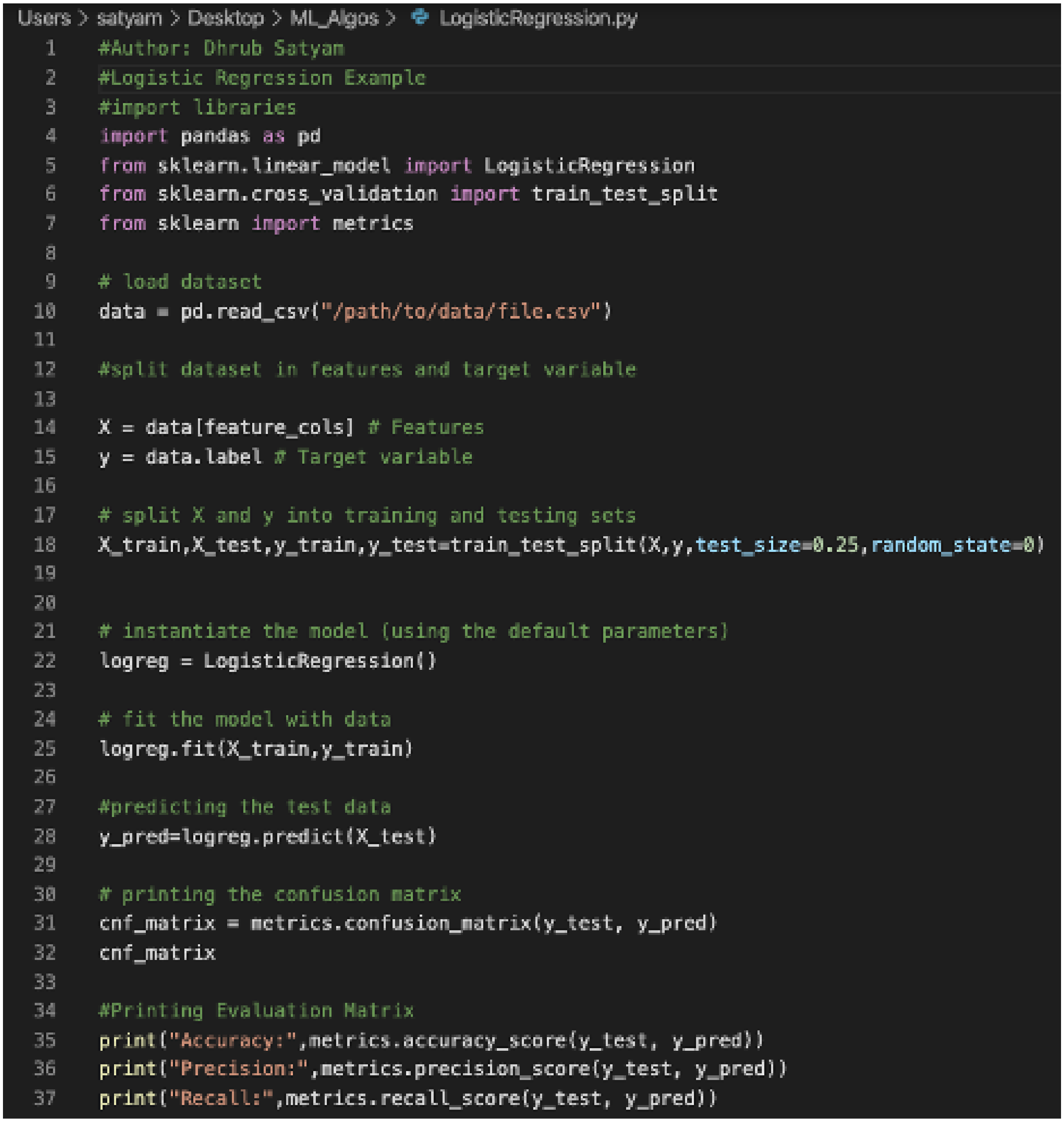
1. The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.
2. There is a linear relationship between the logit of the outcome and each predictor variable. Recall that the logit function is logit(p) = log(p/(1-p)), where p is the probabilities of the outcome.
3. There are no influential values (extreme values or outliers) in the continuous predictors
4. There are no high intercorrelations (i.e. multicollinearity) among the predictors.

How Logistics Regression deals with Multiclass classification?

The most famous method of dealing with multiclass classification using logistic regression is using the one-vs-all approach. Under this approach, a number of models are trained, which is equal to the number of classes. The models work in a specific way.

For example, the first model classifies the datapoint depending on whether it belongs to class 1 or some other class; the second model classifies the datapoint into class 2 or some other class. This way, each data point can be checked over all the classes.

Code Snippet:



```
#Author: Dhrub Satyam
#Logistic Regression Example
#import libraries
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import train_test_split
from sklearn import metrics

# load dataset
data = pd.read_csv("/path/to/data/file.csv")

#split dataset in features and target variable

X = data[feature_cols] # Features
y = data.label # Target variable

# split X and y into training and testing sets
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=0)


# instantiate the model (using the default parameters)
logreg = LogisticRegression()

# fit the model with data
logreg.fit(X_train,y_train)

#predicting the test data
y_pred=logreg.predict(X_test)

# printing the confusion matrix
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix

#Printing Evaluation Matrix
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```