Multinomial Logistic Regression

In statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. It is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.

In a normal logistic problems there are typically two discrete values of the target which are labled as 0 or 1. In a multinomial or multiclass logistic resgression problem there are more than two classes are present.

Following is an example:flower species is a target variable and The independent variables are sepal length , sepal width, petal length and petal width - based on which the target class is predicted.

Algorithmic Approach to solving a multi-class Problem-One vs All Classification

Instead of y=0,1 we will expand our definition so that y=0,1...n Basically we re-run binary classification multiple times, once for each class.

The Theoretical procedure is simple.

1. Divide the problem into n+1 binary classification problems (+1 because the index starts at 0)i.e 1 For each class.
2. Predict the probability the observations are in that single class.
3. Prediction = max(probability of the classes) i.e class with the highest probaibility is selected

For each sub-problem, we select one class (YES) and lump all the others into a second class (NO). Then we take the class with the highest predicted value.

The diagram below explains the process

Multinomial Logistic Regression Implementation with Code

Dataset

The dataset contains details about irish flowers.

- Target variable is Species which has 3 distinct classes- setosa, virginica,versicolor.
- The independent variables are sepal length , sepal width, petal length and petal width, based on these the target class shall be predicted

In [23]:
```
import pandas as pd
import numpy as np
from sklearn import datasets
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn import linear_model
from sklearn import metrics
data = pd.read_csv('iris.csv', index_col=0)
data.head()
```

Out[23]:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa

Multi class Logistic code Implementation :

In [73]:

```
Features=['Sepal.Length','Sepal.Width','Petal.Length','Petal.Width']#Segr
Target=['Species']#Segregating Target
X_train, X_test, y_train, y_test = train_test_split(data[Features],data[T
mul_lr = linear_model.LogisticRegression(multi_class='multinomial', solve
print("Multinomial Logistic regression Train Accuracy: " ,metrics.accurac
print("Multinomial Logistic regression Test Accuracy: " ,metrics.accuracy
df = pd.DataFrame()
df['Predicted']= mul_lr.predict(X_test)
df['Actual']=y_test.dropna().values
print(df) #Predictions(predicted classes) on Test dataset
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\utils\validation.py:52
  y = column_or_1d(y, warn=True)
```

Explanation

In the above, we have created a multiclass logistic regression classifier which have predicted the target values of test dataset with an accuracy of 95.55% . Above, we can see the distribution of actual target values and predicted target. In solving the problem we have used 70% of the dataset for training and rest 30% for testing.

Questionarrie

1. Why multi class regression is considered a combination of multiple logistic regression problems?
2. Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.
    A. We need to fit n models in n-class classification problem
    B. We need to fit n-1 models to classify into n classes
    C. We need to fit only 1 model to classify into n classes
    D. None of these

3.What is the advantage of a stratified multiclass logistic regression model over a one vs all logistic model?

In [ ]: