TDT4173

INDIVIDUAL ASSIGNMENT

LOGISTIC REGRESSION

This algorithm is a binary classification, supervised algorithm. It requires features which are inputs to the model and labels – output of the model for prediction.

Logistic regression is based on probability. It measures the probability of an event occurring or otherwise. It is used in predicting dependent variables from independent variables. It introduces a transformation of linear functions into logarithmic terms. This is known as Sigmoid functions.

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

It fits the parameters by maximizing the log likelihood using gradient descent.

$$\theta_j = \theta_j - \alpha(h_\theta(x) - y)x$$

Alpha is the learning rate and is tuned to reach convergence.

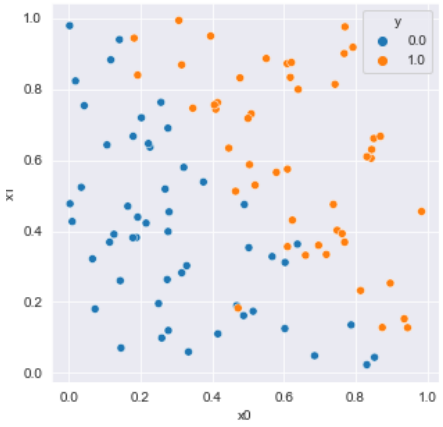
INDUCTIVE BIAS

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

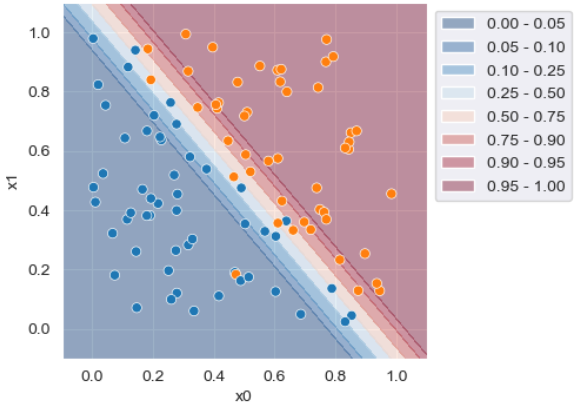
It assumes that the error term (bias - epsilon) from the linear equation is Independently and Identically distributed (IID).

It also assumes that there is linear (decision) boundary that separates the groups of data in binary form. Os and 1s. Confidence is higher the further away from the decision boundary.

PLOTS



Before processing



After processing

RESULTS/PREPROCESSING

For dataset 1.

An accuracy of 95.0% and cross entropy of 0.175 was achieved.

For the more complex dataset 2.

Training

An accuracy of 99.2% and cross entropy of 0.057 was achieved.

Test

An accuracy of 98.6% and cross entropy of 0.060 was achieved.

Feature engineering had to be applied here by transforming into a different space – quadratic. Squaring the inputs to the model gave the optimum metrics.

K-MEANS

This algorithm is an unsupervised learning algorithm that enables machine learning models to group data based on similarities.

Usually, the number of clusters(k) expected is a hyperparameter that aids in the grouping of data(clusters). Below are the steps involved in this algorithm.

1.Select k centroids randomly

2.Make k clusters by assigning each data point to closest centroid

3.Compute new centroids for each cluster

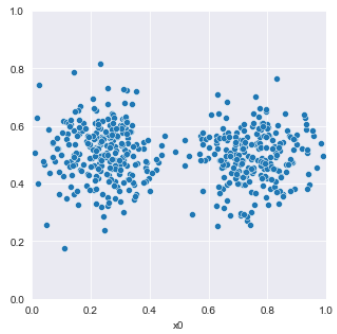
4. Repeat steps 2 and 3 until convergence

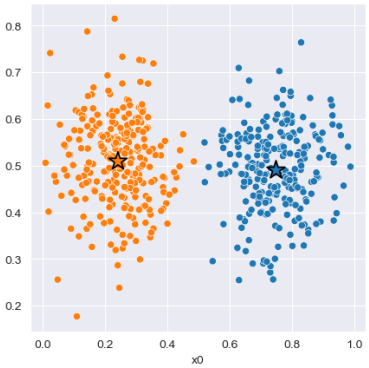
INDUCTIVE BIAS

The number of clusters is assumed to be known.

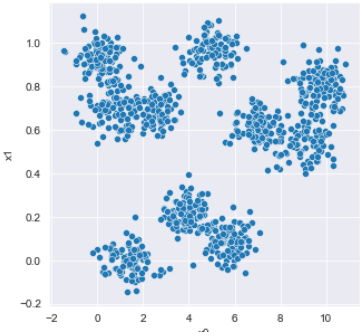
The centroids, which are randomly placed, results in different position of clusters for each iteration. This is observed for each run.

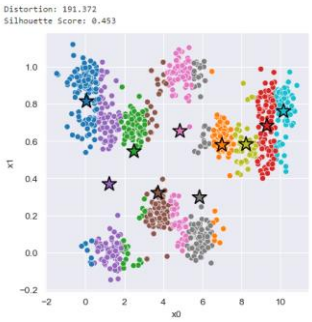
PLOTS



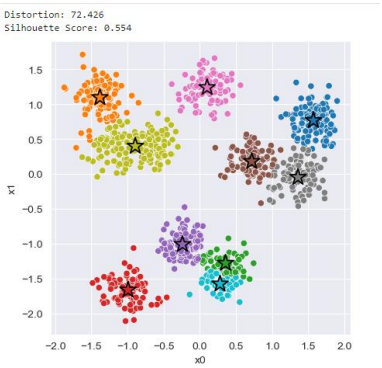


Clustering of dataset 1 after processing data





Clustering of dataset 2 before feature scaling



Clustering of dataset 2 after feature scaling

RESULTS/PREPROCESSING

For dataset1.

Silhouette Score of 0.672 and distortion of 8.837 was achieved.

10 clusters are identified using randomized initializations. The results differ for every iteration but shows pretty good clusters and silhouette scores.

Feature scaling using standardization was the solution to the complex data. This makes the mean of the dataset 0 and the standard deviation, 1.