Mini Project

Xin Huang(xhuang2), Qi Pang(qpang)

1. Introduction:

In this project, we compare the accuracy of prediction between Linear Regression, Logistic Regression and Neural Network on the ‘Skin Segmentation Data Set’ (classification) where we get from

<https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation#>. The goal for our project is to get the best prediction algorithm for our dataset

Data:

The Skin Segmentation dataset is constructed over B, G, R color space. Skin and Nonskin dataset is generated using skin textures from face images of diversity of age, gender, and race people. The dataset has 3 features (Blue, Green, Red) over 245057 data points. The binary target Skin represents by 1 with 50859 samples and Nonskin represents by 0 with 194198 samples.

1. Design of Experiments:
2. Data splits:

In order to get better a performance on accuracy of prediction, we rebuild the dataset with 50000 Skin samples and 50000 Nonskin samples and shuffle the dataset by call python rebuildData.py.

For feature set, we also add a column of 1 as bias for classification.

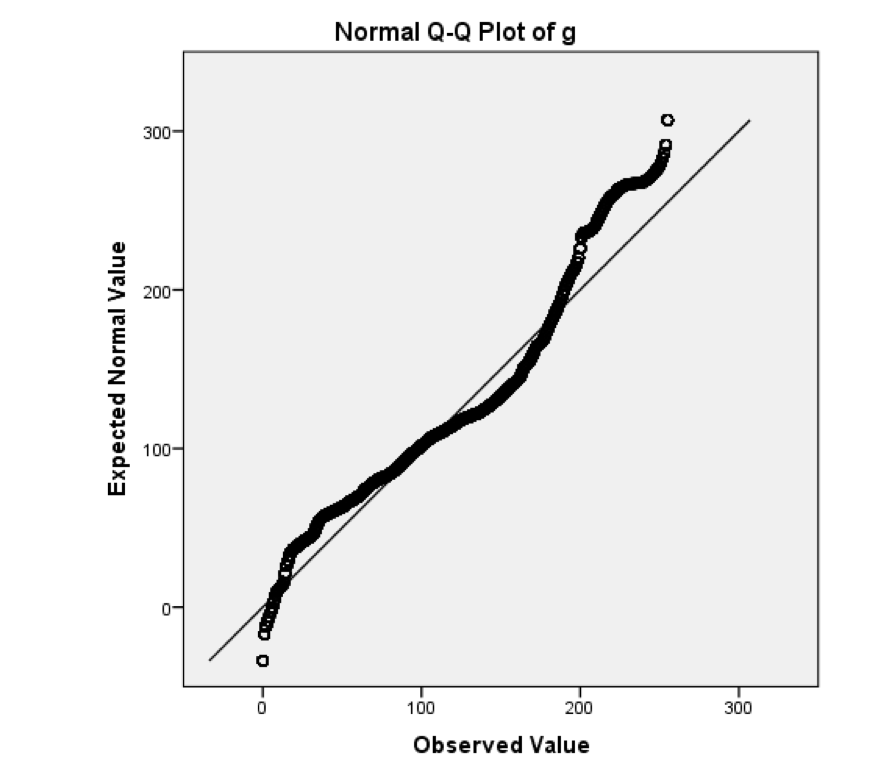
For training and testing, we use K-fold cross validation. We Splitting the dataset to 10 disjoint sets, each set has 10000 samples.

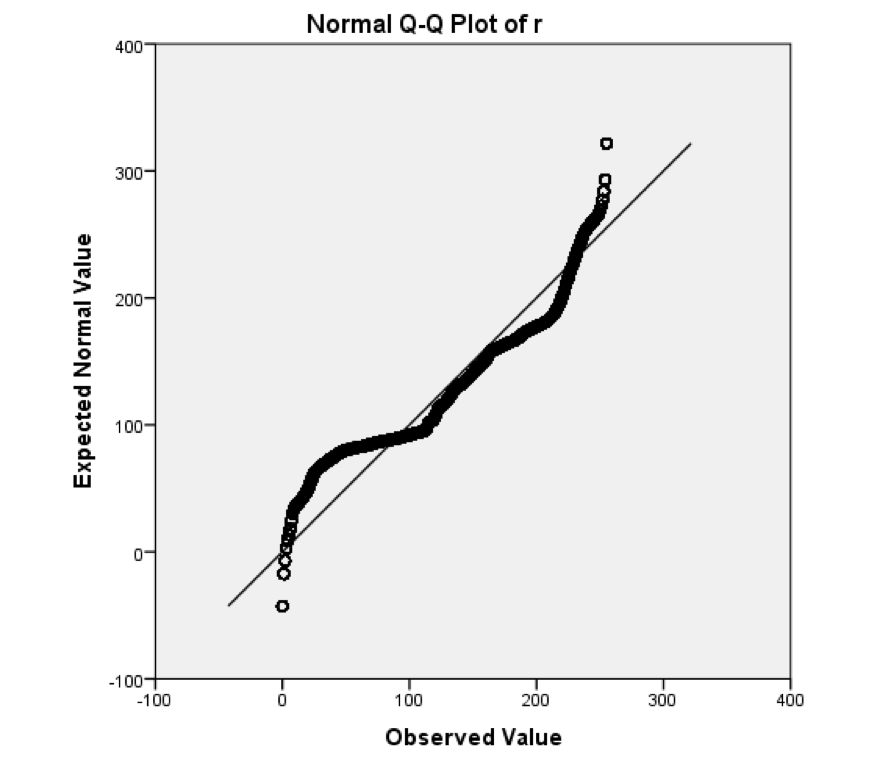
1. Selection of metric:

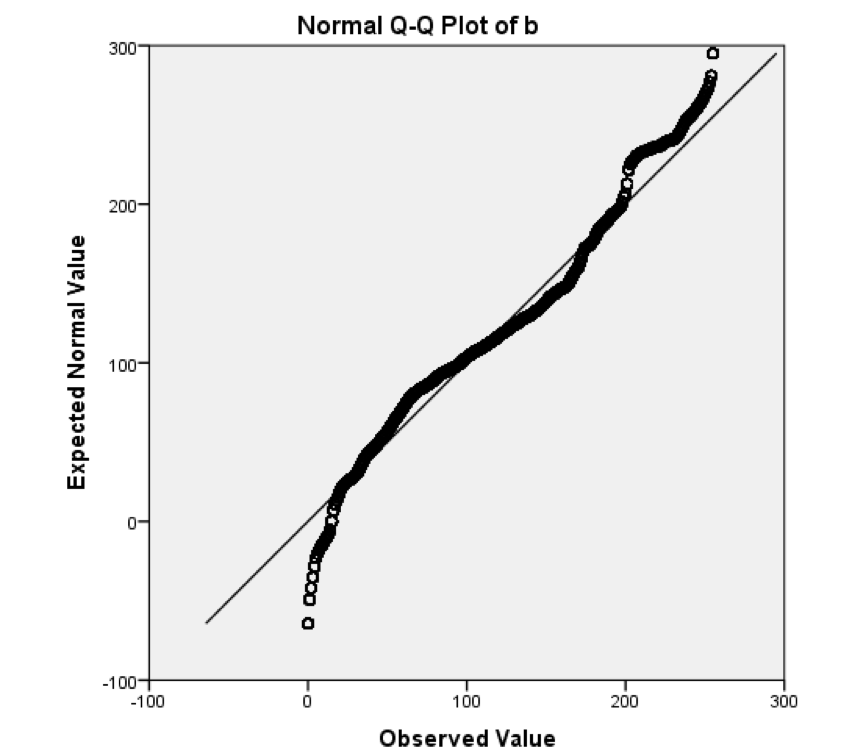
we are using accuracy as our metric, we have two methods, basically we calculate how many correct predictions do we have, then divided by number of test units, then times 100% this is the accuracy rate. our get error gives 100 - accuracy rate. For detail please see get accuracy (ytest, predictions) and get error (ytest, predictions) in our code.

1. Selection of algorithm:

We obtain QQ plots for our three features by SPSS (showing in blow). Those plots give us a hit that our features are not best fit of Gaussian distribution, because they are not best fit the line. So we are not choosing any algorithm related Gaussian (ex: Naive Bayes of Gaussian distribution). Also our targets are 0 and 1 which is binary. So we choose some generalized linear models like logistic and neural network.







1. Algorithm Test:

For each algorithm, we use 9 sets as training set and test on the other set. Running 10 times, make sure every set be tested as testing set. Then we obtain 10 models with 10 error of accuracy.

1. Statistical Significance Test of different Algorithms:

We used t-test for statistical significance test, we import python spicy for t-test

We have three groups of t-test, and alpha is 0.05

H0: Errors for Neural Network and Linear regression are the same

Ha: Neural Network has lower error

P-value is 0.000000, which leads us to reject h0

H0: Errors for Neural Network and Logistic regression are the same

Ha: Neural Network has lower error

P-value is0.000000, which leads us to reject h0

H0: errors for Linear and Logistic regression are the same

Ha: Linear regression has lower error

P-value is 0.045381 which leads us to reject h0

Three groups of t-test give tells us that Neural Network has a better performance than linear regression and logistic regression, and linear regression has a better performance than logistic regression for our dataset

1. Statistical Significance Test of Algorithm parameters:
2. Data visualization:

We run our best parameters (waiting for final result), we test and plot 10000 points. Then obtain the polts below. From the plot we have three assumptions

1. Neural Network has a better performance than Linear regression
2. Neural Network has a better performance than Logistic regression
3. Linear regression has a better performance than Logistic regression

Which leads us to have three groups of Statistical Significance Test below

