Project Part 1

Density Estimation and Classification

A**rijit Panda**

16th Feb 2020

NAÏVE BAYES CLASSIFICATION

Introduction

Naïve Bayes classification is a supervised learning algorithm used to classify discrete random output variable using Bayes theorem with an assumption of independent input variables.

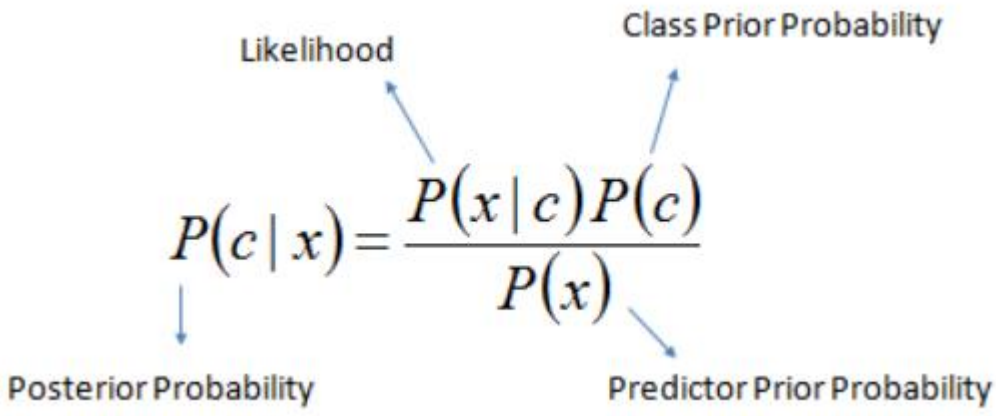
Objective

To create a model to detect digit “7” and “8” using MNIST data subset which has 6265 and 5851 training images for “7” and “8” digits respectively and to test the accuracy using 1028 images of “7” digit and 974 images of “8” digit.
Below two features for each image need to be extracted:

- The average of all pixel values in the image
- The standard deviation of all pixel values in the image

Calculations Required

1) Mean (μ)
2) Standard Deviation (**σ)**
3) Prior probability P(A) = Number of training set for A/Total number of training set
4) Probability Density Function:
   $pdf = (1/[\sigma * sqrt(2\pi)]) * e^{-(x - \mu)^2/2\sigma^2}$
5) Naïve Bayes



Pseudo Code (Algorithm)

1. Feature extraction of training input set: average (feature 1) and standard deviation (feature 2) of all the pixels.

2. Calculation mean, standard deviation and Prior Probability for training input set

3. For each input in testing set – features are extracted followed by calculation of the probability for digit 7 and digit 8 using the product of pdf and prior probability. Predictor prior probability is ignored, as it's same in both the classification.

4. The label with highest probability is the predicted output for the input set.

Accuracy Result:

DIGIT 7:
Accuracy: 0.9659533073929961

DIGIT 8:
Accuracy: 0.8459958932238193

TOTAL ACCURACY: 0.9075924075924076

RUNTIME DURATION: 20.886903 seconds

LOGISTIC REGRESSION

Introduction

Logistic regression is statistical model that predicts discrete values using sigmoid logistic function and a cut off threshold value.

Objective

To create a model to detect digit "7" and "8" using MNIST data subset which has 6265 and 5851 training images for "7" and "8" digits respectively and to test the accuracy using 1028 images of "7" digit and 974 images of "8" digit.

For each image, all 784-pixel values are provided to the logistic regression model for training.

Calculations Required

1) Mean (μ)
2) Standard Deviation (σ)
3) Sigmoid Function

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}.$$

4) Gradient ascent:
   Error function = (Expected value – Predicted Value) = (Y – sigmoidal function)
   Gradient = Product of training input vector and error function
   Parameter = parameter + α*gradient (where α is the learning parameter

Pseudo Code (Algorithm)

1. Feature extraction of training input set: all 784 pixel values are considered as features

2. Set initial weights/parameters to some arbitrary and calculate the final value of weights using gradient ascent with $\boldsymbol{\alpha}$=0.001

3. For each input in testing set – features are extracted followed by calculation of the probability using

New parameter in sigmoidal function.

4. If the resultant probability is greater than 0.5 (threshold) – it is digit 8 (label 1) Otherwise it is digit 7 (label 0)

Accuracy Result:

DIGIT 7:
Accuracy: 0.9659533073929961

DIGIT 8:
Accuracy: 0.9784394250513347

TOTAL ACCURACY: 0.972027972027972

RUNTIME DURATION: 0.959579 second