Fundamentals of Statistical Learning and Pattern Recognition

CSE 569

Project Report on Feature Extraction, Normalization, Density Estimation and Bayesian Classification

Submitted To

Prof. Dr. Baoxin Li

Submitted By

Sai Vikhyath Kudhroli

ASU Id: 1225432689

Date: 03 October 2022

Goal: To classify the digits “3” and “7” using Bayesian Decision Theory. Dataset used is a subset of the MNIST dataset, only the images of digits “3” and “7” are used.

Procedure: The project is comprised of four sub-tasks:

- Feature Extraction.
- Normalization.
- Estimation of Maximum Likelihood Estimates.
- Classification using Bayesian Decision Theory.

Feature Extraction: The following features are extracted from the raw-image data:

- Skewness (k_i) – It is the measure of lack of symmetry which is computed using the given formula:

$$k_i = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^3}{Ns^3}$$

$where\ \bar{y} = Mean, s = Standard\ Deviation, N = Number\ of\ values$

- Bright Dark Ratio (r_i) – It is the ratio of all brighter pixels to all darker pixels. The decision whether a pixel is bright or dark is taken based on a threshold T.

$$r_i = \frac{No\ of\ bright\ pixels}{No\ of\ dark\ pixels}$$

$where\ a\ pixel\ is\ bright\ if\ pixel\ value\ (v) > threshold\ (T)$

Normalization: The features extracted are then normalized for the mathematical operations to be performed efficiently and optimally. The following formula is used for normalization:

$$\left[N_i^1, N_i^2\right]^T = \left[\frac{(k_i - M_1)}{s_1}, \frac{(r_i - M_2)}{s_2}\right]^T$$

Estimation of Maximum Likelihood Estimates: For estimating maximum likelihood parameters, it is assumed that the distributions of all the features follow a normal or gaussian distribution. Then the maximum likelihood estimates (μ, σ) are given as follows:

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\sigma = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu)^2$$

These parameters are used to estimate the likelihood function which is used in Bayesian classification.

Classification using Bayesian Decision Theory: The estimated likelihood parameters and the prior probabilities are used to make predictions about the unseen data using Bayesian Decision Theory. Decision rule to predict the data is given as follows:

$$If$$

$$\left\{p\left(\frac{3}{k_i}\right) \times p\left(\frac{3}{r_i}\right) \times p(3)\right\} > \left\{p\left(\frac{7}{k_i}\right) \times p\left(\frac{7}{r_i}\right) \times p(7)\right\}$$

$$Then\ Decide\ 3$$

$$Else\ Decide\ 7$$

Results: The results of the algorithm are segregated based on the four types of configurations:

- $T = 150, p(3) = 0.5, p(7) = 0.5$
- $T = 150, p(3) = 0.3, p(7) = 0.7$
- $T = 200, p(3) = 0.5, p(7) = 0.5$
- $T = 200, p(3) = 0.3, p(7) = 0.7$

Graph depicting the distribution of the features, Maximum Likelihood Estimates, Confusion matrix, Accuracy and Error-Rate are reported in each case for both train and test datasets.

Graphs depicting the distributions of features for each class: Frequency graph is plotted against the feature value. Since, the features vary widely at extremely high precisions, the feature values are rounded to a single decimal to plot the distributions.

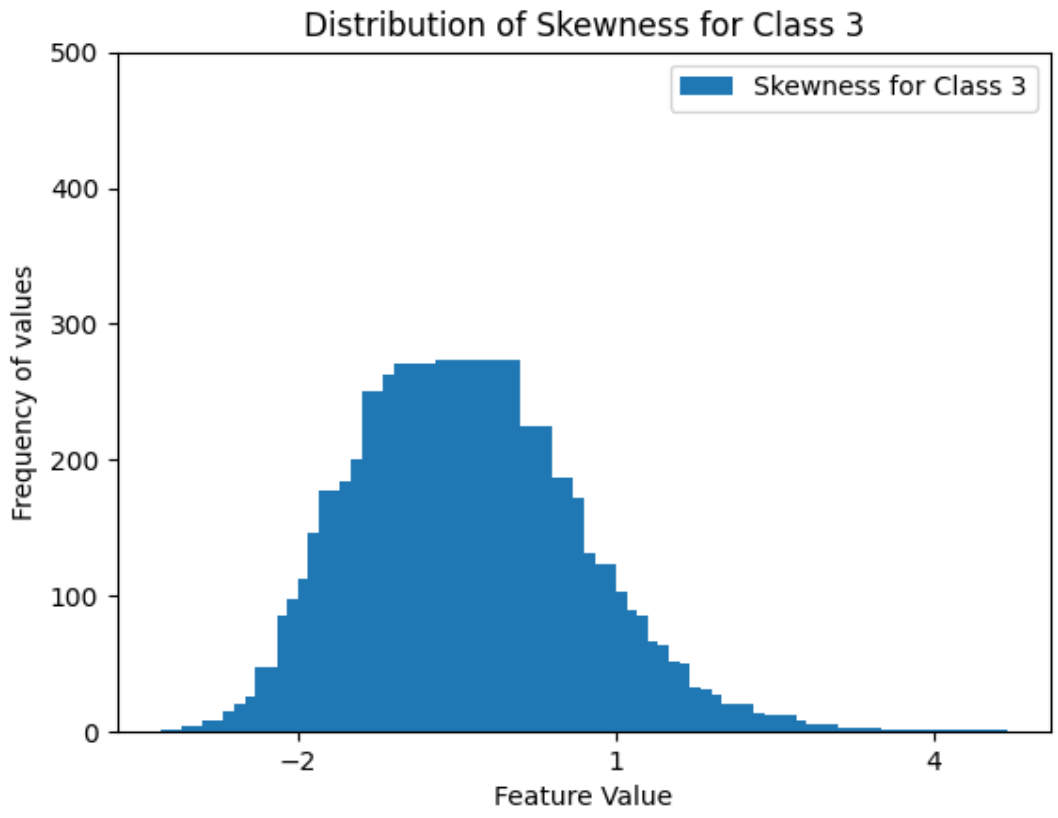

Fig 1. Distribution of Skewness for class 3 in train data.

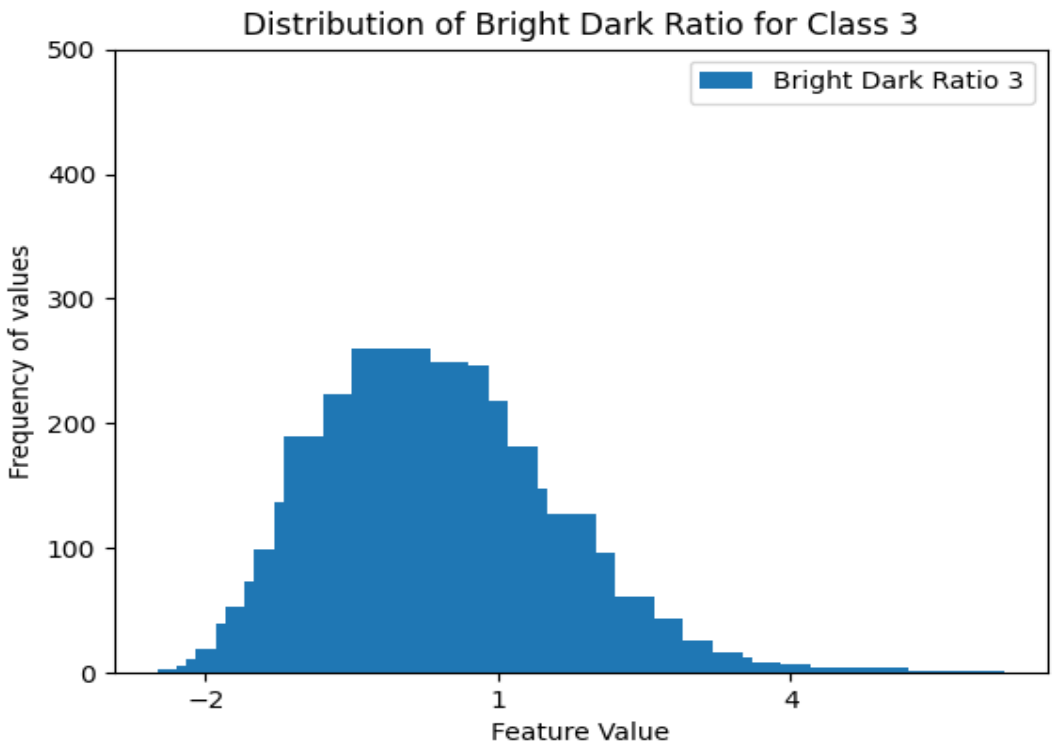

Fig 2. Distribution of Bright to Dark Ratio for class 3 in train data.

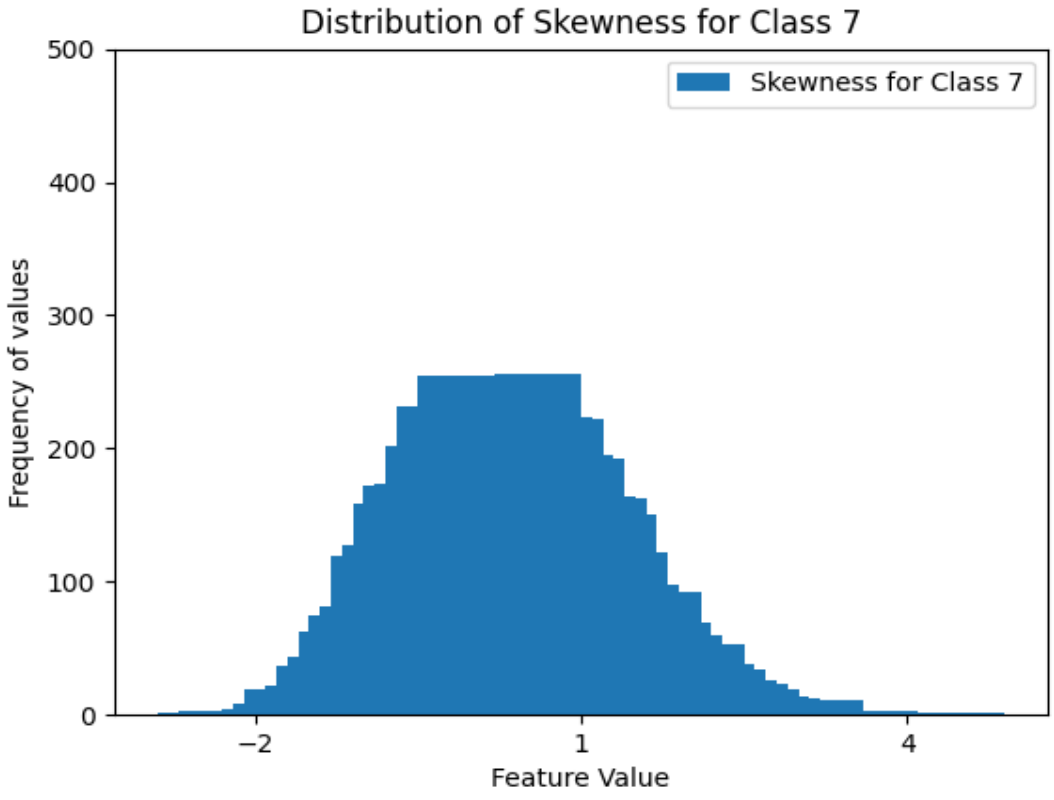

Fig 3. Distribution of Skewness for class 7 in train data

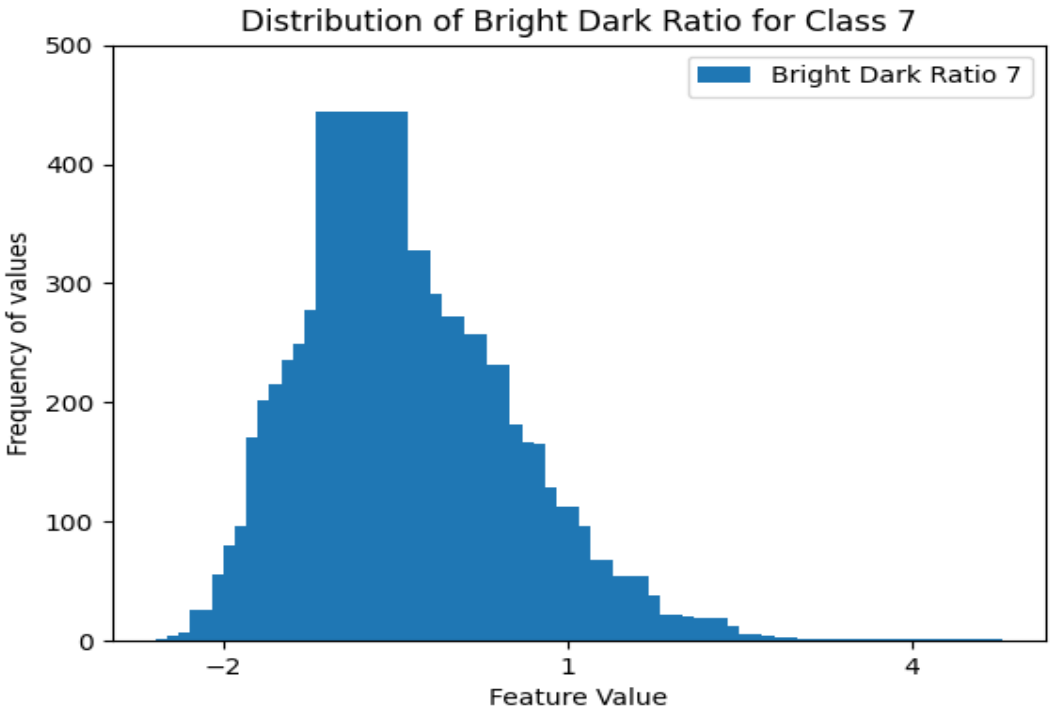

Fig 4. Distribution of Bright to Dark Ratio for class 7 in train data.

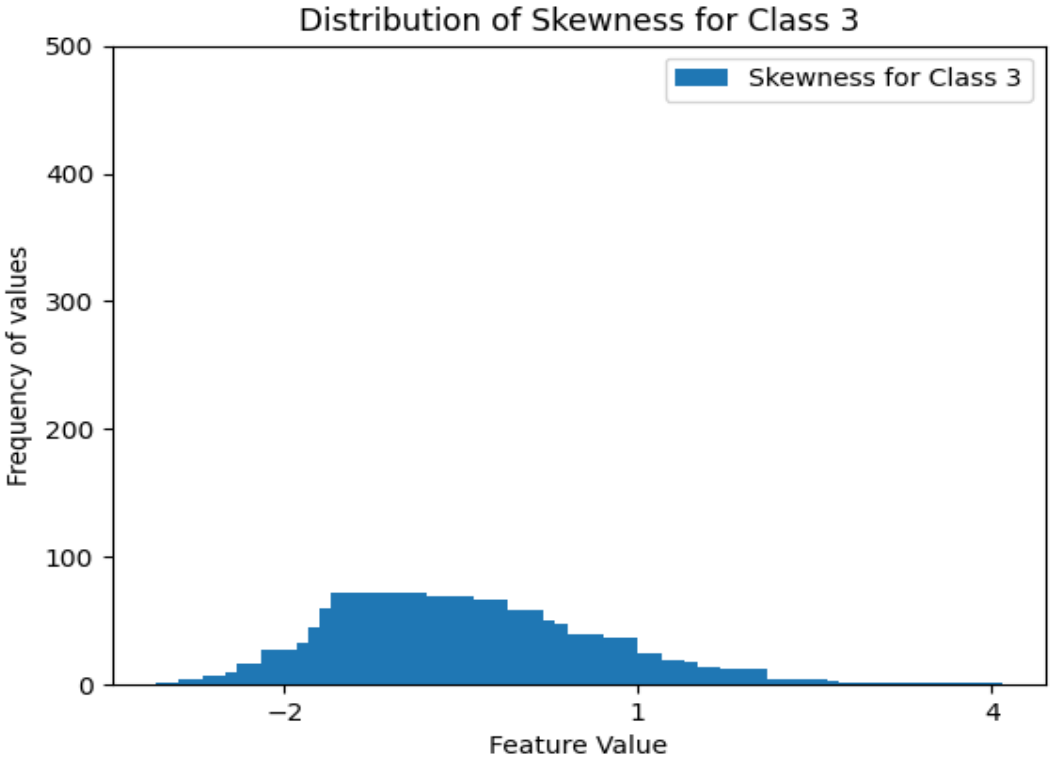

Fig 5. Distribution of Skewness for class 3 in test data

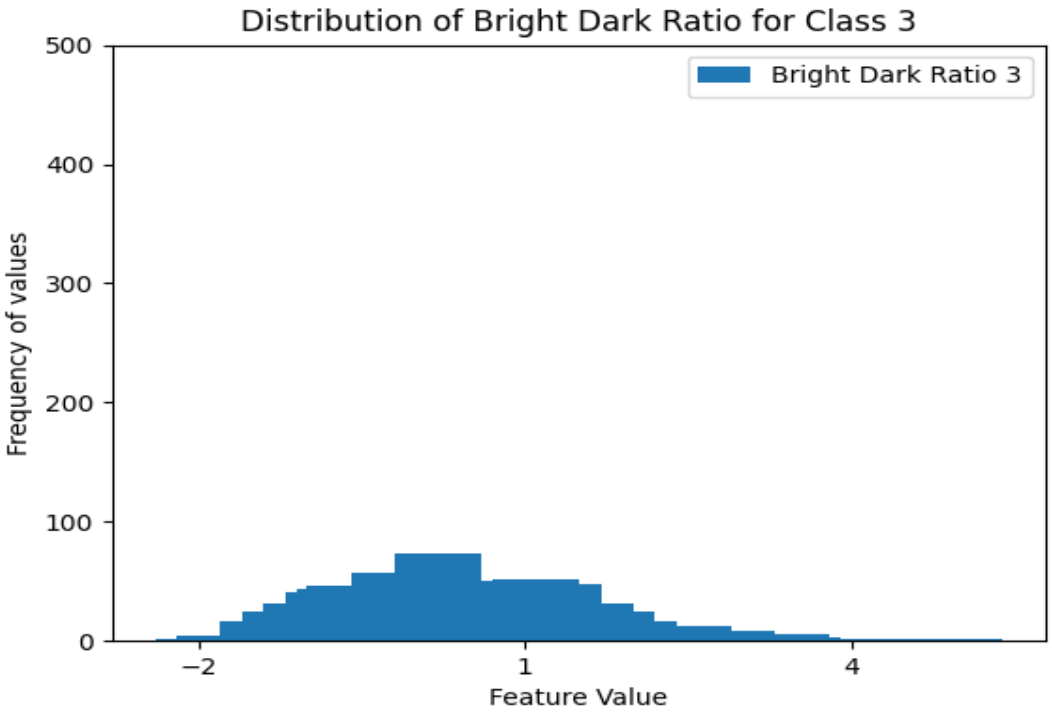

Fig 6. Distribution of Bright to Dark Ratio for class 3 in test data.

Fig 7. Distribution of Skewness for class 7 in test data.

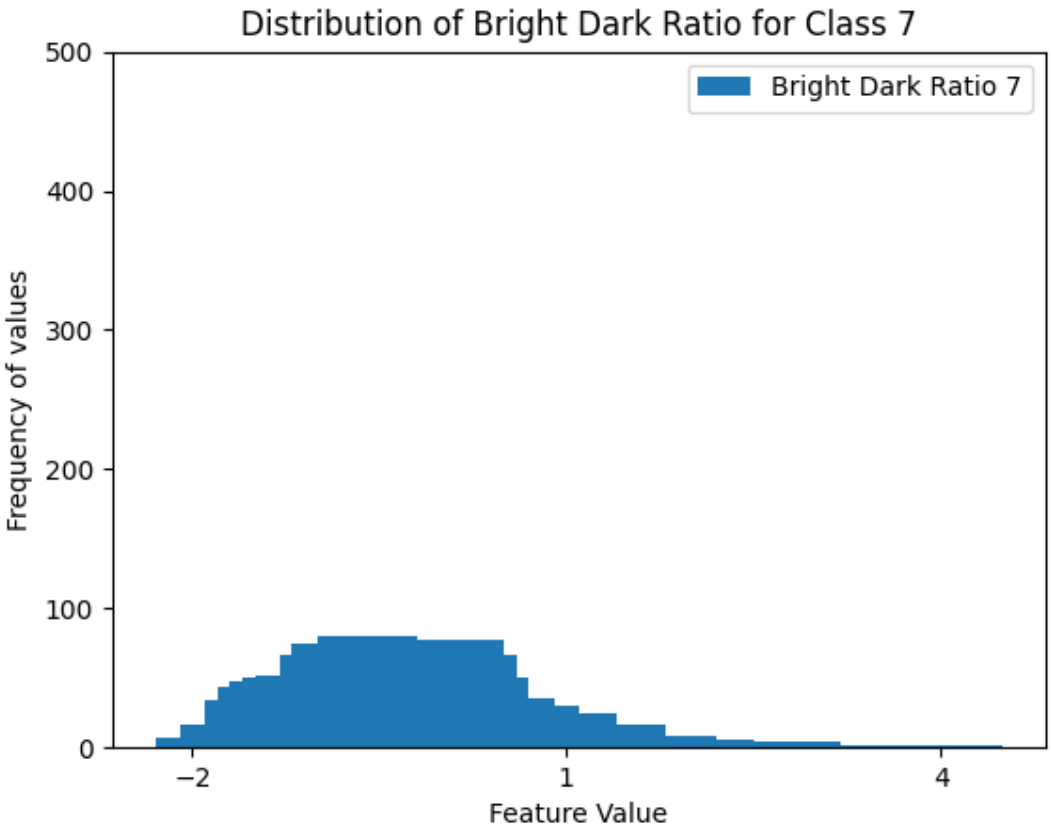

Fig 8. Distribution of Bright to Dark Ratio for class 7 in test data.

Case 1 ($T = 150, p(3) = 0.5, p(7) = 0.5$):

```
C:\Users\vikhy\OneDrive\Desktop\ASU Coursework\Fundamentals of Statistical Learning and Pattern Recognition\Project 1>python BayesianDecisionClassifier.py

**********************************************************************************************************************************************************

================================================================================
              MAXIMUM LIKELYHOOD ESTIMATES FOR PARAMETERS: {T = 150, P(3) = 0.5, P(7) = 0.5}
================================================================================
MLE for Skewness when class is 3 :  -0.37951982606124907 0.8436088113449371
MLE for Skewness when class is 7 :  0.3715847071615968 0.8740223548771973
MLE for Bright Dark Ratio when class is 3 :  0.3585463484384893 1.103363098001918
MLE for Bright Dark Ratio when class is 7 :  -0.3510497495508368 0.6496945174098441
================================================================================


================================================================================
        TRAIN DATA CONFUSION MATRIX FOR PARAMETERS: {T = 150, P(3) = 0.5, P(7) = 0.5}
================================================================================
+------------------+---------------+---------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+---------------+---------------+
| Predicted Class 3 |      3219      |      1491      |
+------------------+---------------+---------------+
| Predicted Class 7 |      2494      |      4344      |
+------------------+---------------+---------------+
================================================================================


================================================================================
        TRAIN DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 150, P(3) = 0.5, P(7) = 0.5}
================================================================================
Accuracy : 65.49186 %
Error-Rate : 0.34508
================================================================================


================================================================================
        TEST DATA CONFUSION MATRIX FOR PARAMETERS: {T = 150, P(3) = 0.5, P(7) = 0.5}
================================================================================
+------------------+---------------+---------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+---------------+---------------+
| Predicted Class 3 |      837       |      424       |
+------------------+---------------+---------------+
| Predicted Class 7 |      591       |      1034      |
+------------------+---------------+---------------+
================================================================================


================================================================================
        TEST DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 150, P(3) = 0.5, P(7) = 0.5}
================================================================================
Accuracy : 64.83021 %
Error-Rate : 0.35170
================================================================================

**********************************************************************************************************************************************************
```

The value of accuracy and error-rate are rounded-off to 5 decimal places.

Case 2 ($T = 150, p(3) = 0.3, p(7) = 0.7$):

```
**********************************************************************************************************************************************************

================================================================================
               MAXIMUM LIKELYHOOD ESTIMATES FOR PARAMETERS: {T = 150, P(3) = 0.3, P(7) = 0.7}
================================================================================
MLE for Skewness when class is 3 :  -0.37951982606124907 0.8436088113449371
MLE for Skewness when class is 7 :  0.3715847071615968 0.8740223548771973
MLE for Bright Dark Ratio when class is 3 :  0.3585463484384893 1.103363098001918
MLE for Bright Dark Ratio when class is 7 :  -0.3510497495508368 0.6496945174098441
================================================================================


================================================================================
        TRAIN DATA CONFUSION MATRIX FOR PARAMETERS: {T = 150, P(3) = 0.3, P(7) = 0.7}
================================================================================
+------------------+----------------+----------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+----------------+----------------+
| Predicted Class 3 |      2450      |      889       |
+------------------+----------------+----------------+
| Predicted Class 7 |      3263      |      4946      |
+------------------+----------------+----------------+
================================================================================


================================================================================
        TRAIN DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 150, P(3) = 0.3, P(7) = 0.7}
================================================================================
Accuracy : 64.04572 %
Error-Rate : 0.35954
================================================================================


================================================================================
================================================================================
        TEST DATA CONFUSION MATRIX FOR PARAMETERS: {T = 150, P(3) = 0.3, P(7) = 0.7}
================================================================================
+------------------+----------------+----------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+----------------+----------------+
| Predicted Class 3 |      652       |      246       |
+------------------+----------------+----------------+
| Predicted Class 7 |      776       |      1212      |
+------------------+----------------+----------------+
================================================================================


================================================================================
        TEST DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 150, P(3) = 0.3, P(7) = 0.7}
================================================================================
Accuracy : 64.58766 %
Error-Rate : 0.35412
================================================================================

**********************************************************************************************************************************************************
```

The value of accuracy and error-rate are rounded-off to 5 decimal places.

Case 3 ($T = 200, p(3) = 0.5, p(7) = 0.5$):

```
**********************************************************************************************************************************************************

================================================================================================
              MAXIMUM LIKELYHOOD ESTIMATES FOR PARAMETERS: {T = 200, P(3) = 0.5, P(7) = 0.5}
================================================================================================
MLE for Skewness when class is 3 :  -0.37951982606124907 0.8436088113449371
MLE for Skewness when class is 7 :  0.3715847071615968 0.8740223548771973
MLE for Bright Dark Ratio when class is 3 :  0.30674366711084583 1.1364800330406106
MLE for Bright Dark Ratio when class is 7 :  -0.30033017484225155 0.6840509409638845
================================================================================================


================================================================================================
================================================================================================
       TRAIN DATA CONFUSION MATRIX FOR PARAMETERS: {T = 200, P(3) = 0.5, P(7) = 0.5}
================================================================================================
+------------------+---------------+---------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+---------------+---------------+
| Predicted Class 3 |      3118      |      1500      |
+------------------+---------------+---------------+
| Predicted Class 7 |      2595      |      4335      |
+------------------+---------------+---------------+
================================================================================================


================================================================================================
       TRAIN DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 200, P(3) = 0.5, P(7) = 0.5}
================================================================================================
Accuracy : 64.53931 %
Error-Rate : 0.35461
================================================================================================


================================================================================================
       TEST DATA CONFUSION MATRIX FOR PARAMETERS: {T = 200, P(3) = 0.5, P(7) = 0.5}
================================================================================================
+------------------+---------------+---------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+---------------+---------------+
| Predicted Class 3 |      808       |      434       |
+------------------+---------------+---------------+
| Predicted Class 7 |      620       |      1024      |
+------------------+---------------+---------------+
================================================================================================


================================================================================================
       TEST DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 200, P(3) = 0.5, P(7) = 0.5}
================================================================================================
Accuracy : 63.47886 %
Error-Rate : 0.36521
================================================================================================
================================================================================================

**********************************************************************************************************************************************************
```

The value of accuracy and error-rate are rounded-off to 5 decimal places.

Case 4 ($T = 200, p(3) = 0.3, p(7) = 0.7$):

```
**********************************************************************************************************************************************************

==========================================================================================
               MAXIMUM LIKELYHOOD ESTIMATES FOR PARAMETERS: {T = 200, P(3) = 0.3, P(7) = 0.7}
==========================================================================================
MLE for Skewness when class is 3 :  -0.37951982606124907 0.8436088113449371
MLE for Skewness when class is 7 :  0.3715847071615968 0.8740223548771973
MLE for Bright Dark Ratio when class is 3 :  0.30674366711084583 1.1364800330406106
MLE for Bright Dark Ratio when class is 7 :  -0.30033017484225155 0.6840509409638845
==========================================================================================


==========================================================================================
        TRAIN DATA CONFUSION MATRIX FOR PARAMETERS: {T = 200, P(3) = 0.3, P(7) = 0.7}
==========================================================================================
+------------------+---------------+---------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+---------------+---------------+
| Predicted Class 3 |      2218      |      867       |
+------------------+---------------+---------------+
| Predicted Class 7 |      3495      |      4968      |
+------------------+---------------+---------------+
==========================================================================================


==========================================================================================
        TRAIN DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 200, P(3) = 0.3, P(7) = 0.7}
==========================================================================================
Accuracy : 62.22723 %
Error-Rate : 0.37773
==========================================================================================
==========================================================================================


==========================================================================================
        TEST DATA CONFUSION MATRIX FOR PARAMETERS: {T = 200, P(3) = 0.3, P(7) = 0.7}
==========================================================================================
+------------------+---------------+---------------+
|                  | Actual Class 3 | Actual Class 7 |
+------------------+---------------+---------------+
| Predicted Class 3 |      606       |      241       |
+------------------+---------------+---------------+
| Predicted Class 7 |      822       |      1217      |
+------------------+---------------+---------------+
==========================================================================================


==========================================================================================
        TEST DATA ACCURACY AND ERROR-RATE FOR PARAMETERS: {T = 200, P(3) = 0.3, P(7) = 0.7}
==========================================================================================
Accuracy : 63.16701 %
Error-Rate : 0.36833
==========================================================================================

**********************************************************************************************************************************************************
```

The value of accuracy and error-rate are rounded-off to 5 decimal places.

References:

- MNIST database
  - http://yann.lecun.com/exdb/mnist/
- NIST statistics handbook
  - https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
- Probability concepts explained: Maximum likelihood estimation – Jonny Brooks Bartlett
  - https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
- Maximum Likelihood Estimation – Alexander Katz and Eli Ross
  - https://brilliant.org/wiki/maximum-likelihood-estimation-mle/