Density Estimation and Classification Project Report

Shachi Shah
ID#: 1217121828
School of Computing and Augmented Intelligence
Arizona State University Online
Azusa, California, United States of America
spshah22@asu.edu

I. Introduction

The project aims to implement a binary classification system for images of digits "0" and "1" from a subset of the MNIST dataset. The tasks include feature extraction, parameter estimation for a Naive Bayes classifier, classification of test data, and evaluation of model performance. The dataset contains 5,000 samples each for digits "0" and "1" in the training set, and 980 and 1,135 samples, respectively, in the test set.

II. Implementation

A. Feature Extraction

Each image is represented as a 28x28 grayscale matrix. To simplify the classification problem, two features were extracted:

a. Feature 1: The average brightness of all pixels.
b. Feature 2: The standard deviation of brightness.

B. Parameter Estimation

To implement the Naive Bayes classifier, we assume the features are independent and follow a Gaussian distribution. For each feature and each class ("0" and "1"), the following parameters were computed:

a. Mean of Feature 1 and 2
b. Variance of Feature 1 and Feature 2

The calculate_parameters function was used to calculate these parameters for the training data.

C. Classification

The Naive Bayes classifier was implemented using the Gaussian probability density function, which computes the likelihood of a feature given the class parameters. The classifier outputs the class (0 or 1) with the highest posterior probability for each test sample. This was achieved using the naive_bayes function.

D. Accuracy Calculation

The accuracy of the classifier was computed as the proportion of correctly classified samples in the test set for both digit "0" and digit "1". This was done using the calculate_accuracy function.

III. Results

The project output is a list containing:

a. ASU ID
b. Means and Variances of both features for digits “0” and “1”
c. Classification accuracies for the test sets of digits “0” and “1”

Final Output:

```
['0900', 44.214642857142856,
115.55025890254062,
87.43059533674749,
101.46880329712313,
19.41101964285714,
32.70034380610885,
61.38655612666858,
84.80371151566892,
0.9163265306122449,
0.9233480176211454]
```

d. Mean and Variance: Feature 1 is higher for digit “0” indicating that it is generally brighter than digit “1”.
e. Variance is also higher for digit “0”

Accuracy:

a. Digit “0” Test Set: 91.63%
b. Digit: “1” Test Set: 92.33%

The accuracy values indicated that the naive Bayes classifier performs well on the sets given.

A.

IV. Observations

A. The extracted features (brightness and standard deviation) provide a simple yet

effective way to differentiate between digits "0" and "1."

B. The assumption of Gaussian-distributed features is reasonable for this dataset.
C. Despite differences in the number of samples in the test sets, the classifier achieved balanced accuracy for both classes.

V. Challenges and Solutions

A. *Feature scaling:* The brightness values were not normalized, but this did not impact accuracy
B. Dataset Understanding: Initial exploration of the dataset was crucial to ensure correct feature extraction and parameter estimation.
C. *Edge Cases:* The classifier occasionally misclassifies images with non-standard handwriting or significant noise.

VI. Conclusion

This project demonstrated the implementation of a Naive Bayes classifier on a simplified classification problem. The model achieved high accuracy using basic statistical features, highlighting the effectiveness of probabilistic methods for image classification. Future work could explore adding more features (e.g., edge detection) or experimenting with more advanced classifiers like logistic regression or neural networks.

References

[1] Yiran, Luo. “Density Estimation and Classification Project” *CSE 575 - Statistical Machine Learning*, Ira A. Fulton Schools of Engineering, 13 January 2025.