Bernoulli Classification Model Report

https://github.com/eunhong0925/gks-4008.git

20221009 GKS Kim Eunhong

01. Theory and Method

The goal of this project is to create a classification model for the MNIST handwritten dataset. The MNIST dataset is a collection of 28x28 pixel grayscale images of handwritten digits ranging from 0 to 9.

This model takes Bernoulli Naive Bayes approach to simplify the MNIST problem, treating it as a binary classification task. In other words, it supposes that each pixel is independent. It can be useful for the data that can be easily binarized and for simplification to problem solving offering striaghtfoward solution to the problem. Furthermore, this model is based on Bayes' theorem, utilizing prior probabilities and conditional probabilities to compute posterior probabilities. Additionally, log probabilities is used in the calculation process to reduce errors. Specifically, it follows a structure where it calculates the log probabilities (prior probabilities) for each of the ten classes and computes the feature probabilities (conditional probabilities) for the data in each class image. These calculations are based on the 768-dimensional data, where each pixel shoud be flattened for probability Through this process, it obtains the log posterior probabilities using this approach.

02. Implementation

1. Import necessary libraries
2. Load the MNIST dataset and Store
3. Convert the dataset from pandas DataFrames to NumPy arrays
5. Binarize the pixel value
6. Split the dataset into training and testing sets using a test size of 20% and a random seed.
7. Create a custom Bernoulli Naive Bayes model class
8. Initialize the model and train it using the training data.
9. Evaluate the model's performance
10. Calculate and print the accuracy of the model on the test data.

03. Experimental Results

The experimental results indicate that the implemented Bernoulli Naive Bayes model achieved an accuracy of approximately 70.77% on the MNIST test dataset. This accuracy represents the proportion of correctly classified handwritten digits. Additionally, the code includes a visualization of the model's predictions for a sample of 10 test images compared to their actual labels. This visual comparison helps understand how the model is performing on individual instances.

04. Problem and Limtation

One of the challenges in this project is that this model does not have high accuarcy. It may be the inherent limtation that MNIST images are grayscale and not binary, and the pixel values' independence assumption may not hold. Consequently, there is a possibility that the information in the dataset may not be fully reflected.