CS-233(b) Introduction to Machine Learning (BA4) MS1

Kanghyeon Kim, Eunae Lee, Sangwon Lee

# Introduction

The recognition of handwritten mathematical symbols has numerous applications in education, scientific research, and human-computer interaction, being used to improve learning experiences, promote collaboration, and advance artificial intelligence. This project aims to implement and optimize the machine learning models for the classification of the hand-drawn mathematical symbols dataset, HASYv2 with K-Means, Logistic Regression, and Support Vector Machine.

## Method

2.1 Dataset

The HASYv2 dataset contains 168,233 images (32x32 pixels) of LaTeX symbols drawn by hand and 369 unique symbol classes, including numbers, Latin and Greek letters, and various mathematical operators. Only a sub-sample of the 10 most frequent classes (**Fig.1**) (*\int: 350 sampels, \sum: 339, \infty: 291, \alpha: 259, \xi: 256, \equiv: 251, \partial: 240, \mathds{R}: 237, \in: 229, \square: 217*) were used here. The dataset was split into 2138 train images and 531 test images and the train images were randomly split into 80% for training and 20% for validation by the implemented function, *train\_split()* in main.py.

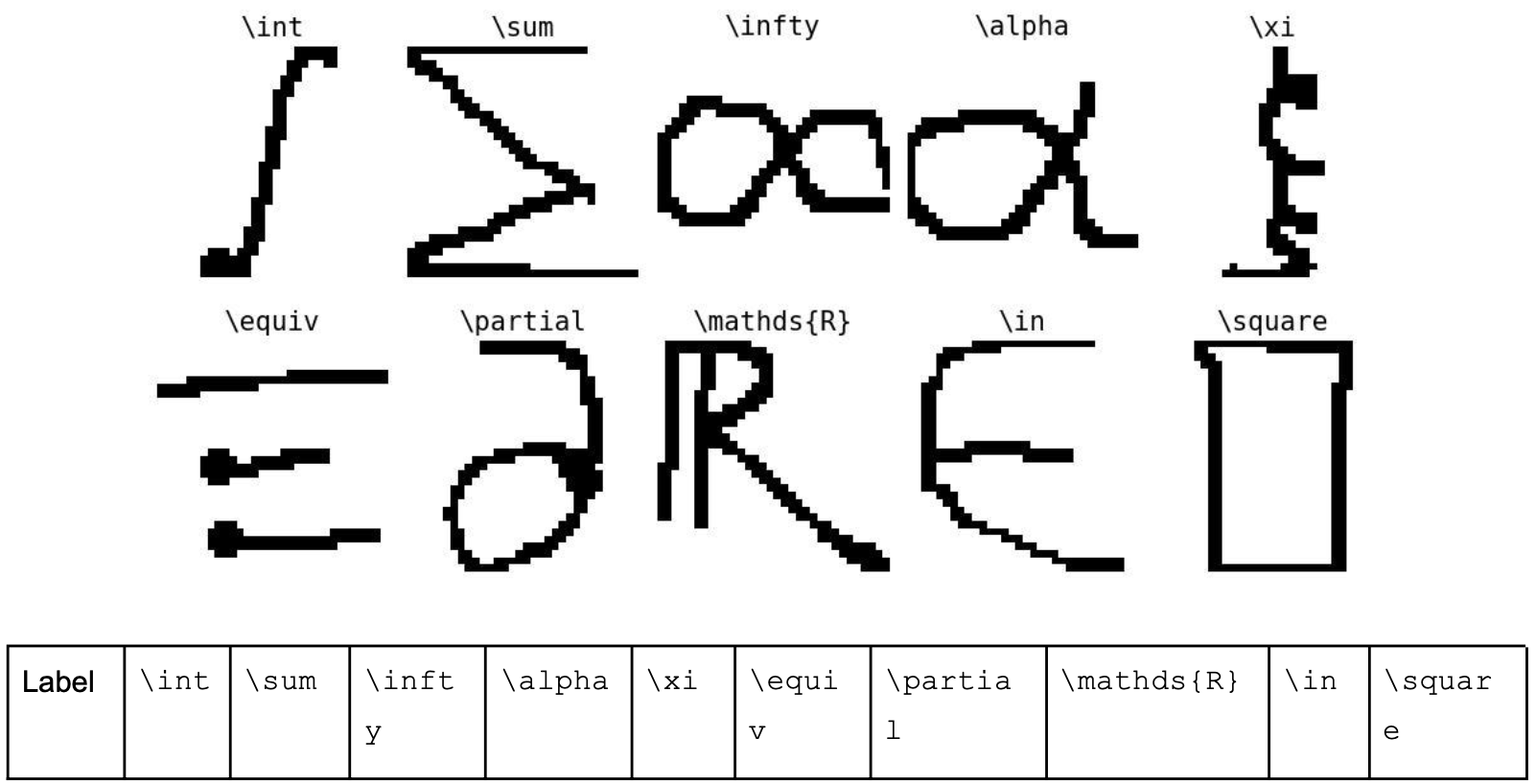


Figure 1. The sub-sample of the HASYv2 dataset

2.2 Preprocessing

Before applying the classification algorithms, the images were preprocessed to improve their quality and reduce the computational complexity through normalization.

2.3 Algorithms

We implemented K-means, Logistic Regression, and Support Vector Machine models with Python. For K-means, we made K (given number) clusters of data points based on their distance in the feature space and calculate the centers of these clusters. This process was repeated until the centers converged. The label which has the largest number of samples in the cluster was assigned to the cluster. For logistic regression, we fit a model that estimates a set of parameters that relate the predictors to the probability of each category using maximum likelihood estimation. In order to fit a model, the independent variables are assumed to have a linear relationship with the log-odds of the dependent variable. For support vector machine, we used the scikit-learn implementation, sklearn.svm.SVC.

2.3 Validation

We used cross-validation with 5 groups to evaluate the performance of a model with different values of the hyperparameters and choose the optimal values.

Logistic Regression: A supervised learning algorithm that models the probability of a given input belonging to a specific class using a logistic function.

Support Vector Machine (SVM): A supervised learning algorithm that constructs a hyperplane to separate the data points into different classes while maximizing the margin between the classes.

## Results

2.1 Dataset

The HASYv2

Despite the superior performance of

## Discussion

The results indicate that the Support Vector Machine outperforms both K-means and Logistic Regression in classifying hand-drawn mathematical symbols from the HASYv2 dataset. This can be attributed to the SVM's ability to handle complex relationships between the features, which is crucial in the case of handwritten symbols.

K-means, being an unsupervised learning algorithm, performed the worst among the three. This is expected as it does not utilize the labeled information available in the dataset. On the other hand, both Logistic Regression and SVM, which are supervised learning algorithms, performed better due to their ability to learn from labeled examples.

"인용문"

|  |  |  |
| --- | --- | --- |
|  | 열 머리글 | 열 머리글 |
| 행 머리글 | 텍스트 | 123.45 |
| 행 머리글 | 텍스트 | 123.45 |
|  |  |  |

* 컴퓨터, 태블릿 또는 휴대폰의 Word에서 이 문서를 보고 편집하세요.
* 텍스트를 편집하고, 그림, 도형, 표 같은 콘텐츠를 손쉽게 삽입하고, Windows, Mac, Android 또는 iOS 장치의 Word에서 클라우드에 문서를 원활하게 저장할 수 있습니다.