Assignment 1

GitHub link to code:

https://github.com/KatrineWik/Assignment1/blob/main/assaignment1_katrine_wik.ipynb

Content

Introduction 1
Working with the data 1
Logistic Regression Model..2
Model Evaluation 2
Results and Conclusion..2
Plots..3
Refrances..5

Introduction

The goal of this assignment is to build a machine learning model that can automatically classify songs into two genres: Pop and Classical. This is done by using logistic regression with stochastic gradient descent (SGD) on a dataset of songs from Spotify.

Working with the data

In the first step, we load the dataset 'SpotifyFeatures.csv' using the Pandas library. Second step we filter the data to only include songs classified as either Pop or Classical. Pop songs are labeled as 1, and Classical songs are labeled as 0. The code will then count amount off zeros and ones to give us the amount of numbers pop and classical songs. Next, we select the 'liveness' and 'loudness' features to be used for classification.

The dataset is split into a training set (80%) and a test set (20%), while maintaining a balanced class distribution. Shuffling is also applied to ensure that the order of songs during training does not affect the model's performance, so the songs appear randomly.

Logistic Regression Model

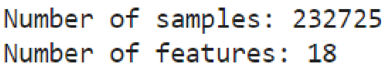
We implemented a logistic regression model from scratch using stochastic gradient descent. The model is trained over multiple epochs, with the weights and bias updated at each step based on the prediction error. The sigmoid function is used to convert the linear combination of the input features into a probability, which is then used to predict the class. And in this case the classes are divided into two, pop and classic.

Different learning rates were tested, and a learning rate of 0.05 provided the best train accuracy 92.56%. The training error decreased as the number of epochs increased, demonstrating the model's learning process. But at rate 5 we got an lower accuracy on 49,74%. My guess is that it reached the point of overfeeding.

Model Evaluation

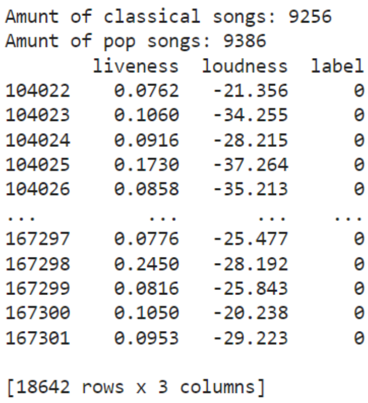
After training the model, we evaluated its performance on the test set. The accuracy of the classifier on the test set was calculated, and a confusion matrix was generated to provide a detailed analysis of the classification results.

Results and Conclusion



```
Number of samples: 232725
Number of features: 18
```

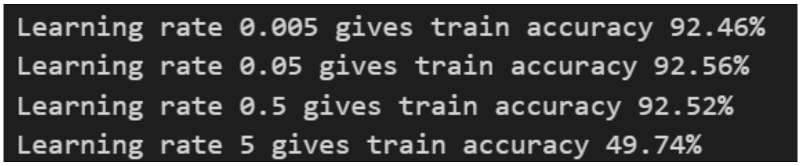
Figure 1 amount of samples and feautures



```
Amunt of classical songs: 9256
Amunt of pop songs: 9386
        liveness  loudness  label
104022    0.0762   -21.356      0
104023    0.1060   -34.255      0
104024    0.0916   -28.215      0
104025    0.1730   -37.264      0
104026    0.0858   -35.213      0
...          ...       ...    ...
167297    0.0776   -25.477      0
167298    0.2450   -28.192      0
167299    0.0816   -25.843      0
167300    0.1050   -20.238      0
167301    0.0953   -29.223      0

[18642 rows x 3 columns]
```

Figure 2 Amount of songs in each category, and dividing the songs into liveness and loudness



```
Learning rate 0.005 gives train accuracy 92.46%
Learning rate 0.05 gives train accuracy 92.56%
Learning rate 0.5 gives train accuracy 92.52%
Learning rate 5 gives train accuracy 49.74%
```

Figure 3 Diffrent learning rate results

The logistic regression model was able to classify songs into Pop and Classical genres with accuracy 92.20%. The confusion matrix provided additional insights into the model's performance, showing how many songs were correctly and incorrectly classified for each genre.

In conclusion, while the model performed well overall, there is still room for improvement. Exploring additional features or using more advanced models could potentially lead to even better classification.

Plots

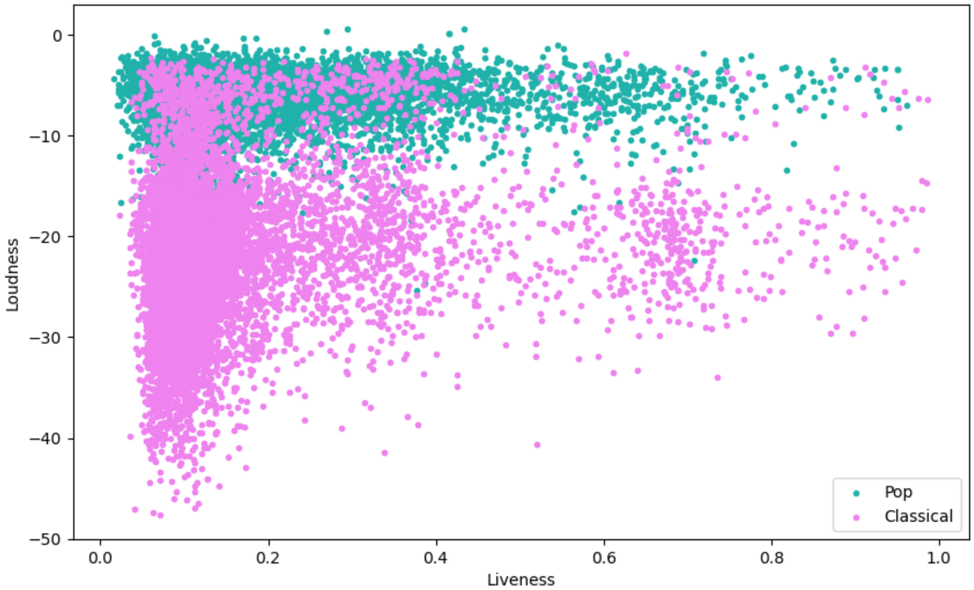


Figure 4 samples on the liveness vs loudness plane, with a different color for each class

Below is the plot of training error as a function of epochs. This shows how the model improves over time as it learns from the training data. The decrease in error over time is a good indication that the model is learning.

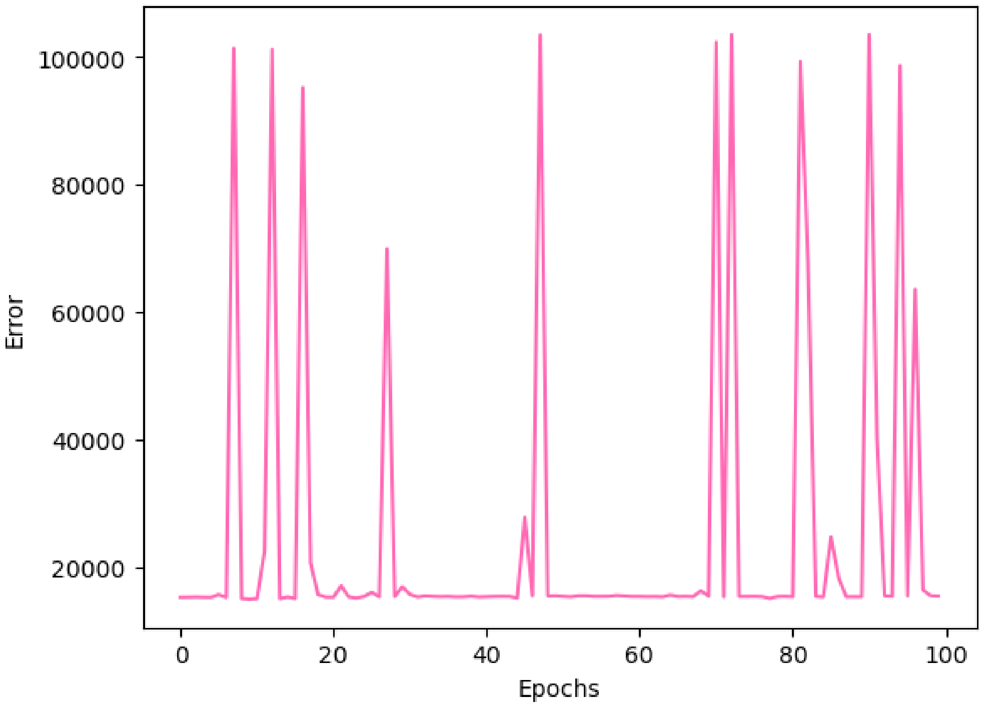


Figure 5 training error as a function of epochs

The following plot illustrates how different learning rates affect the model's training process. We can observe how a slower learning rate might give a more stable convergence, while larger rates could lead to quicker but less stable learning.

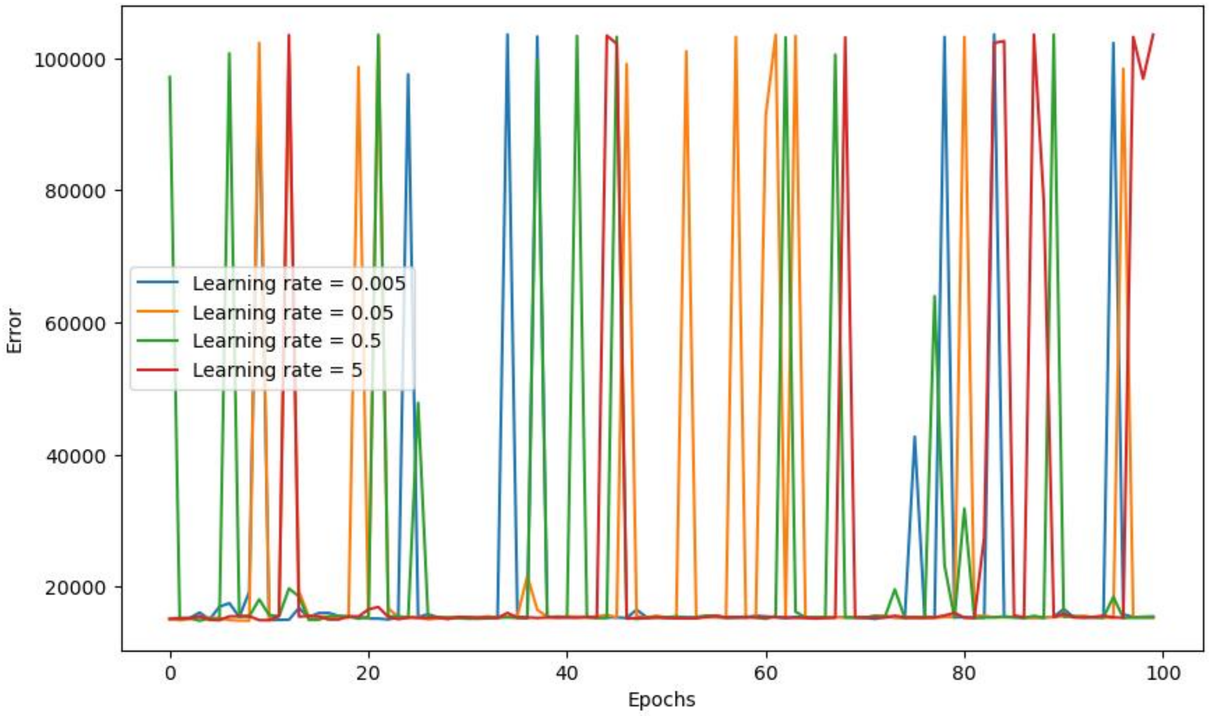


Figure 6 Different learning rates affect the model's training process

The confusion matrix below shows the performance of the logistic regression model on the test set. It provides insights into how well the classifier distinguishes between Pop and Classical songs. The matrix shows true positives, false positives, true negatives, and false negatives.

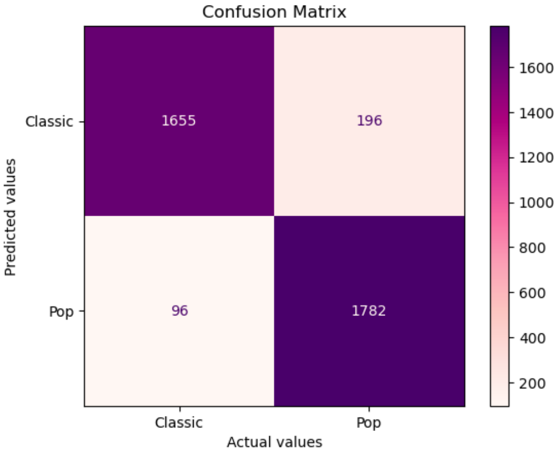


Figure 7 Condusion Matrix

Refrances

Rahman, M. S., & Islam, M. R. (2022). *Effect of feature selection on the accuracy of music popularity classification using machine learning algorithms*. Retrieved from ResearchGate: https://www.researchgate.net/publication/364970799