

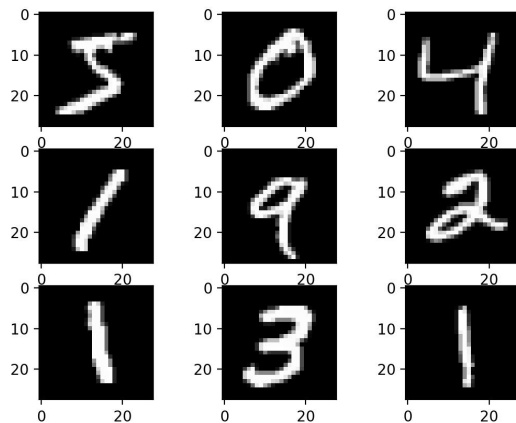
Handwritten digits recognition

Ana Anđelić SW14/2019

In this project we will learn how to build a model to detect handwritten digits using MNIST dataset. We will build linear and non-linear SVM classifiers, compare them and try to tune hyperparameters to get the most effective model.

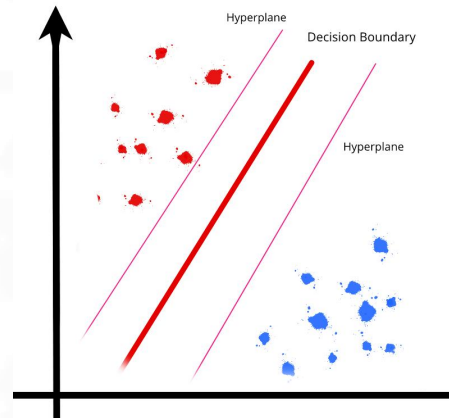
Dataset

MNIST dataset contains 60000 training images and 10000 images for testing. Each image is 28x28, which means it has 784 pixels/features. Each image is labeled with a number it represents. Dataset was constructed from the original ubyte files.



Model

Since this dataset is too large to train all at once, SGD classifier was trained in batches each containing 1000 instances. After that linear and non-linear SVC models were trained using only a portion of the original dataset (10000 instances). Non-linear RBF kernel outperformed linear by 0.05 when comparing F1 score.



Hyperparameters

First we scaled the dataframe using `scale()` method. After that we set different variations of gamma and C hyperparameters. Dataset was separated into training, test and validation set in ratio 6:2:2. After tuning the hyperparameters results were C=10 and gamma=0.001 which gives us 0.95 for F1 score.

