Distributed Feature Selection and One class classification on NHRR data set

IIT2015032 Rohan MR
IIT2015039 Nishant Verma
IIT2015042 Raghav Saboo
IIT2015045 Harsh Vardhan

Problem Statement

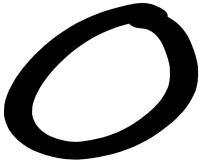
Generate data sets of hand written digits and perform one class classification. Further we will perform Distributed feature selection to obtain global feature set without reduction in model accuracy.

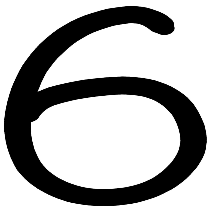
Why generating our own dataset?

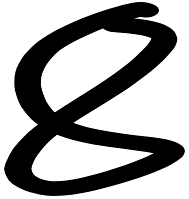
Steps used in Dataset Generation

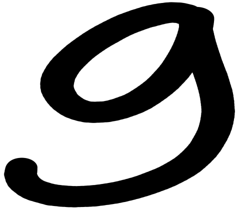
1 Create image data

We have generated data of 4 handwritten digits namely 0,6,8,9.

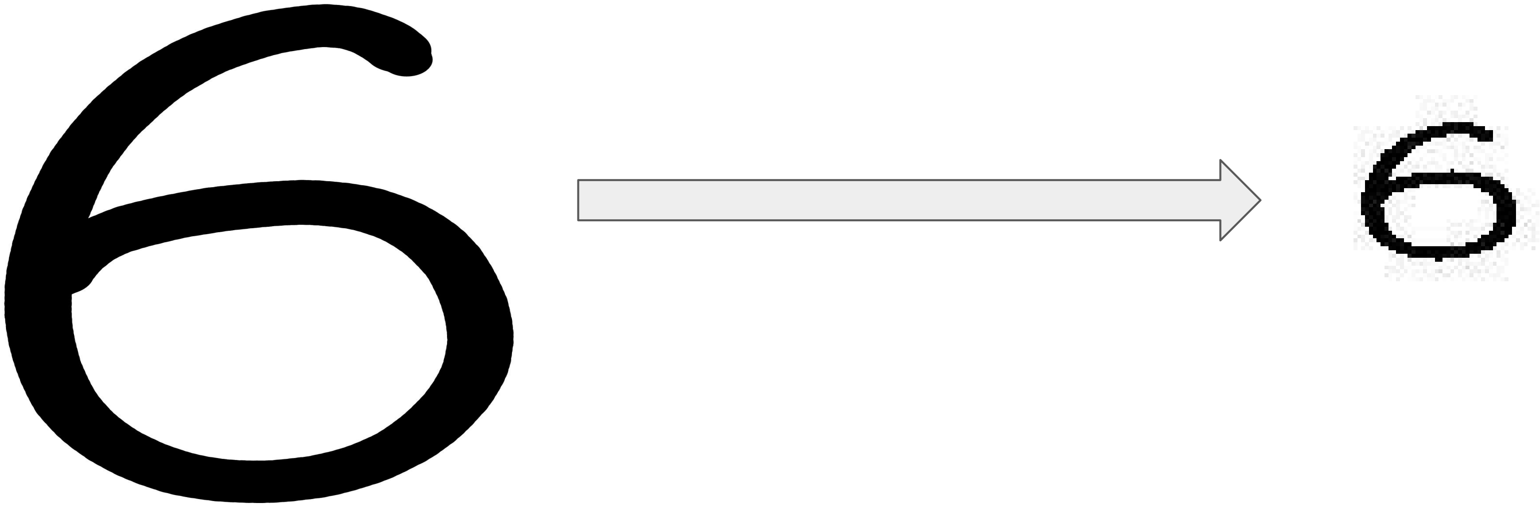




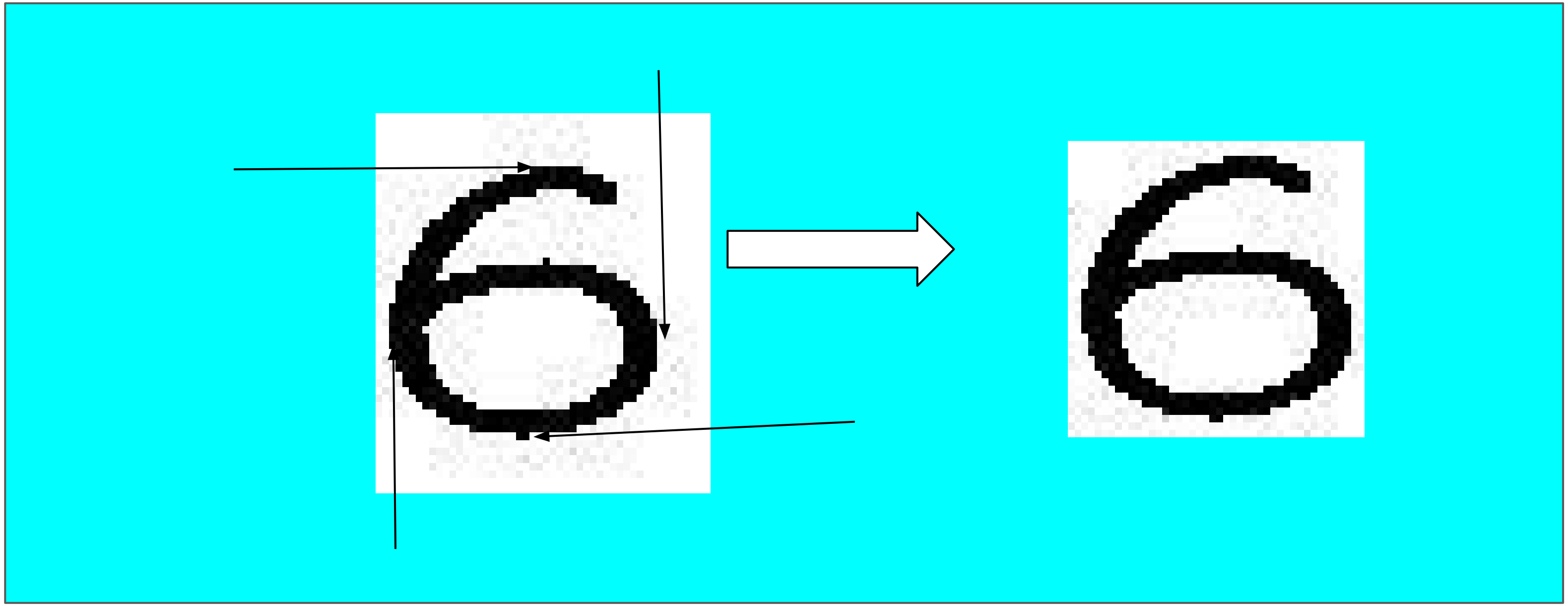




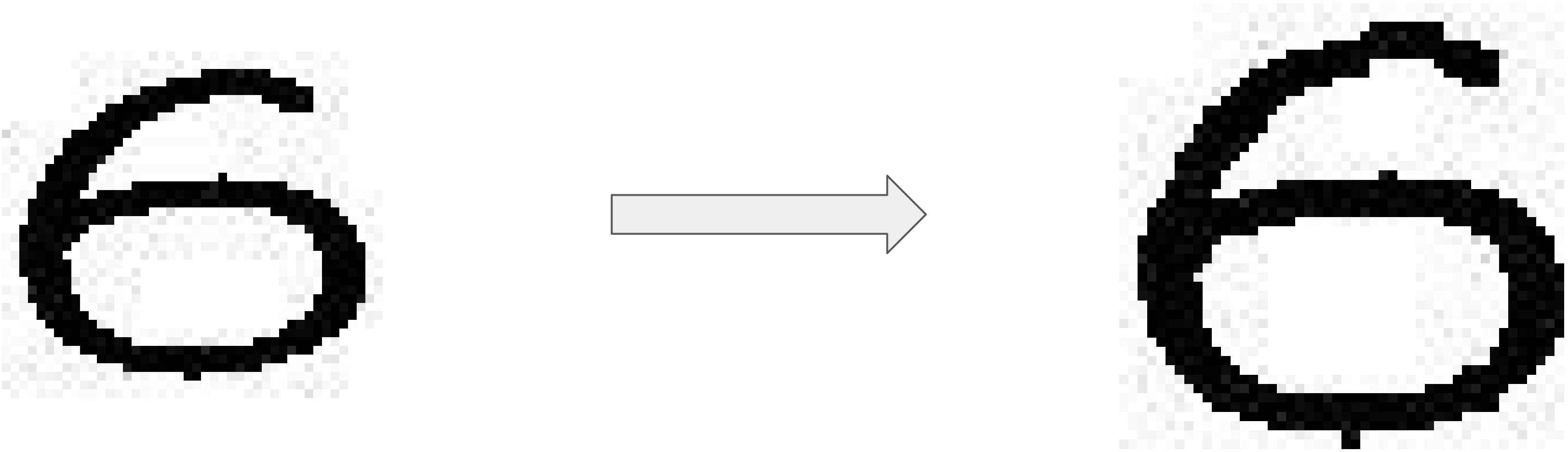
2 Resize the original image to 50 X 50 pixel image.



3 Digit Extraction from the image.

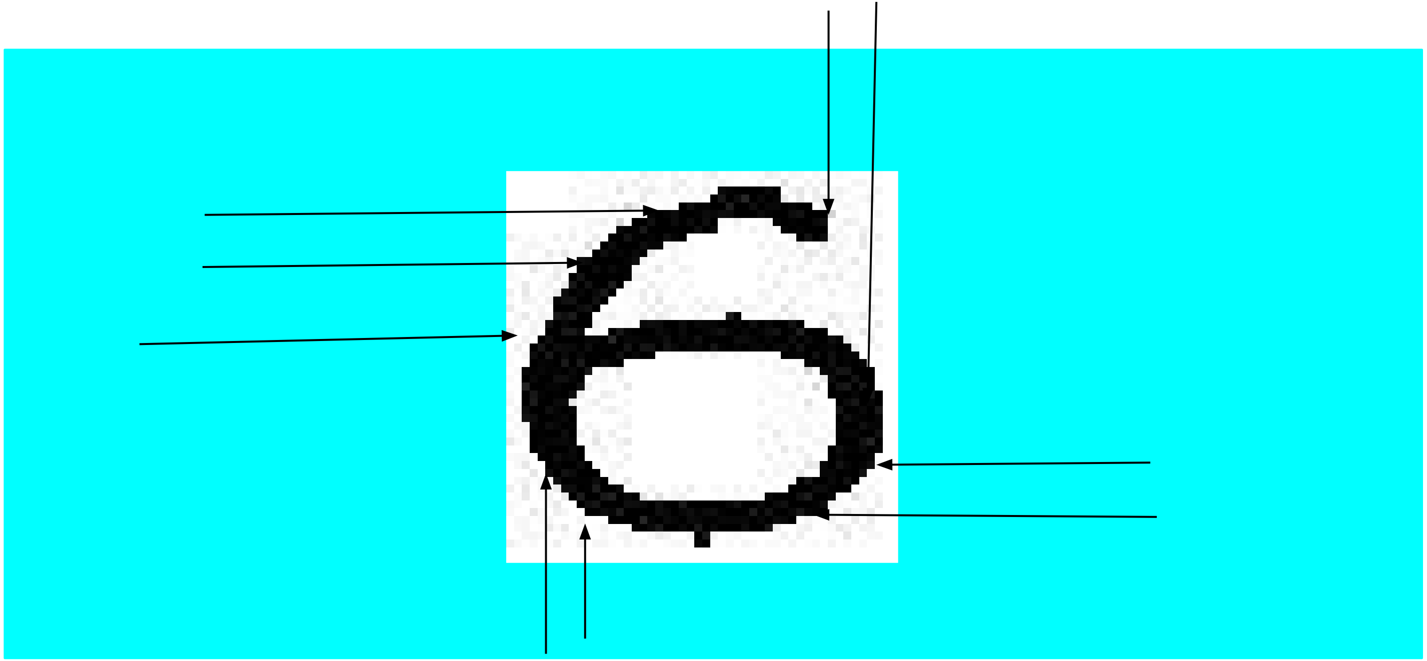


4 The cropped image was again upscaled to 50 x 50 pixel image



Feature Extraction from the generated image

Important features were extracted from the generated image. We took 4 X 50 ie 200 features out of 50 X 50 pixel size image.



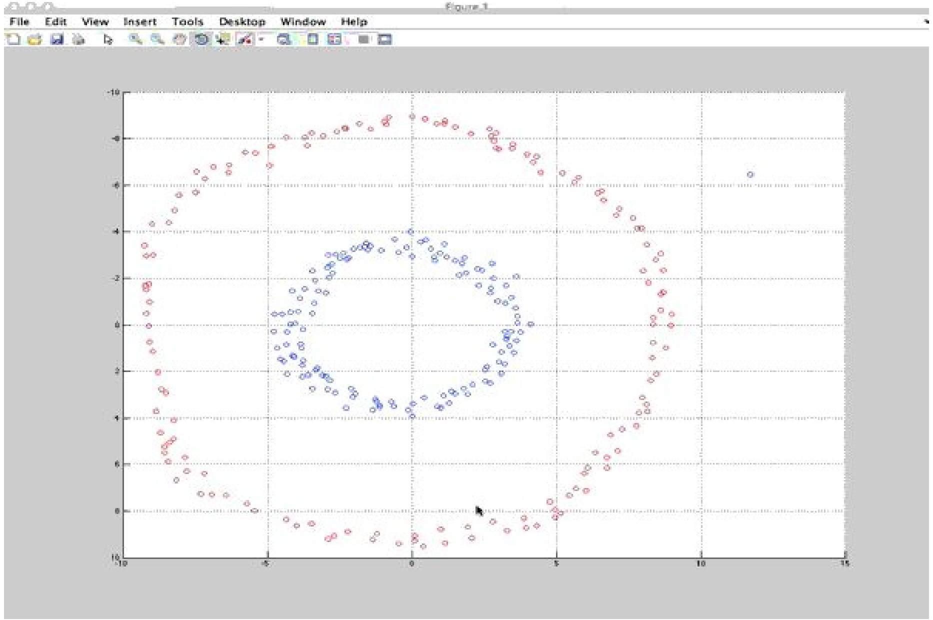
One class Classification

In machine learning one class classification tries to find objects of a specific class by learning from only objects of that class.

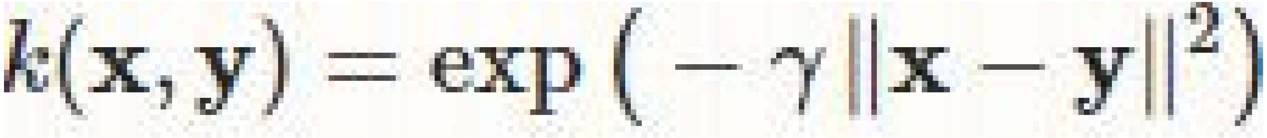
Many applications can be found for example novelty detection , outlier detection.

Why we use Kernel?

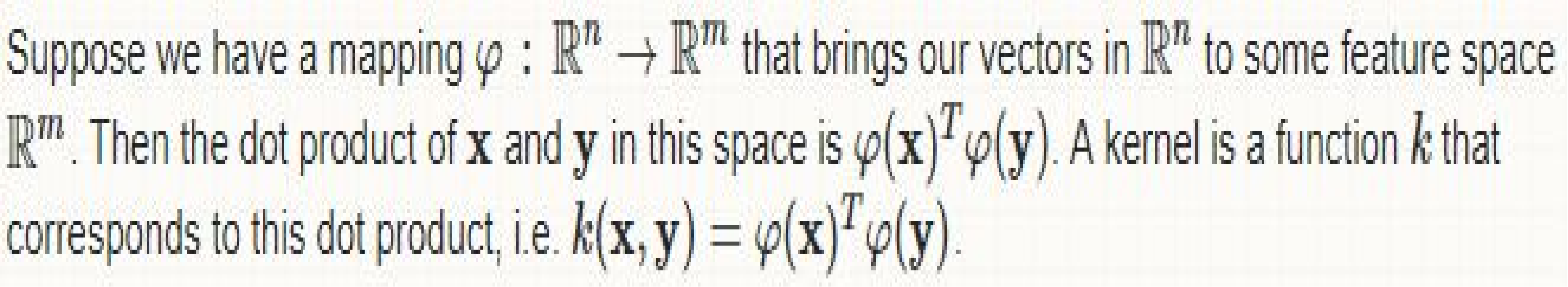
Kernel is a way of computing the dot product of two vectors x and y in some (possibly very high dimensional) feature space.



There are various types of kernel. We will be using Gaussian Kernel.



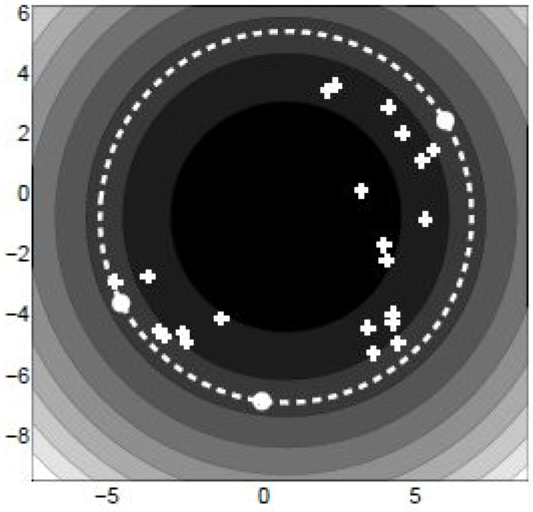
$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma\|\mathbf{x}-\mathbf{y}\|^2\right)$$

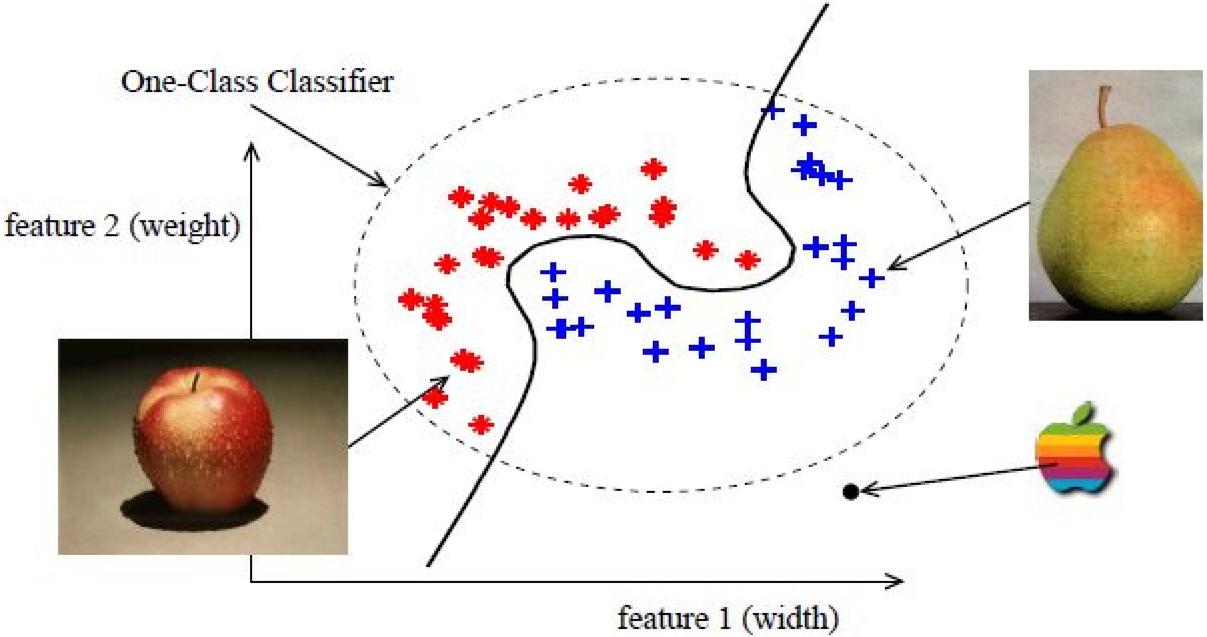


Suppose we have a mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ that brings our vectors in $\mathbb{R}^n$ to some feature space $\mathbb{R}^m$. Then the dot product of $\mathbf{x}$ and $\mathbf{y}$ in this space is $\varphi(\mathbf{x})^T\varphi(\mathbf{y})$. A kernel is a function k that corresponds to this dot product, i.e. $k(\mathbf{x},\mathbf{y}) = \varphi(\mathbf{x})^T\varphi(\mathbf{y})$.

One class classification using SVDD

SVDD is support vector data description.We need to find a sphere with minimum volume containing all the target data points.




One-Class Classifier
feature 2 (weight)
feature 1 (width)

Results using One class SVM in sklearn

Training data Size	Testing data size	Accuracy
742 (digit 0)	2689	99.2%
604 (digit 6)	2689	98.7%
632 (digit 8)	2689/1947(non 0)	72.8%/97.4%
711 (digit 9)	2689/1947(non 0)	79.2%/90.9%

Work to be done after Mid Sem

Implement One class classification from scratch using SVDD the method as proposed by Tax and Duin.

Compare our results with that of Sklearn’s.