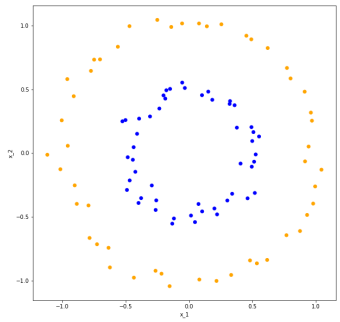
Assignment 2 Report

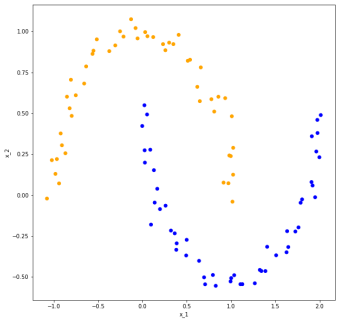
Machine Learning

Submitted by: Himanshu Aggarwal (MT17015)

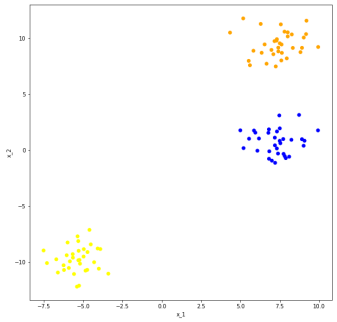
1. Visualisations of the given datasets:



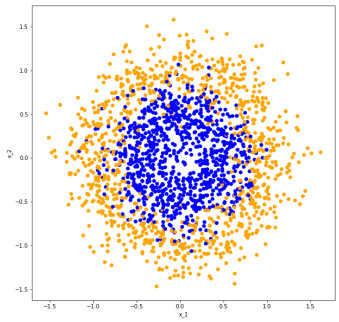
Dataset 1



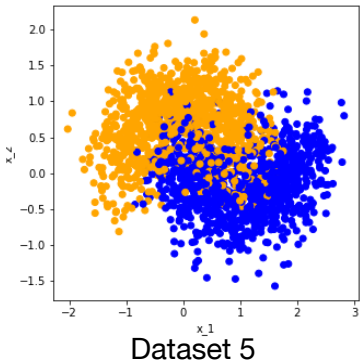
Dataset 2



Dataset 3



Dataset 4



Dataset 5

i. Dataset 1
   Data given is of the circular form in terms of visualisation, i.e., it is not linearly separable.

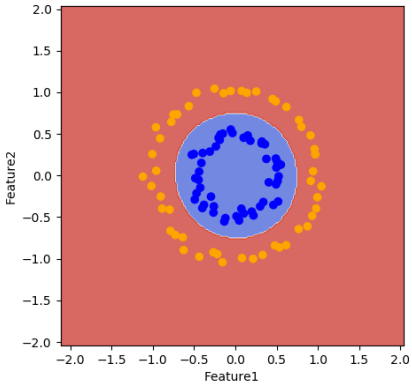
ii. Dataset 2
   Moon shape data, again, not linearly separable.

iii. Dataset 3
   Data is already clustered into 3 groups. Data is linearly separable.

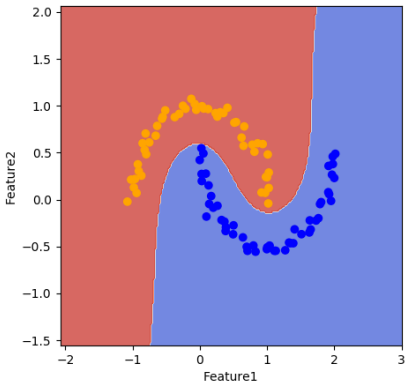
iv. Dataset 4
   Data is visually similar to dataset 1, but contains a lot of noise.

v. Dataset 5
   Data is visually similar to dataset 2, but contains a lot of noise.

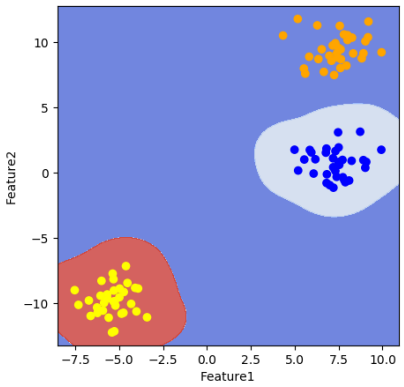
2. Visualisation of the datasets using SVM (with custom built kernel and predict function)



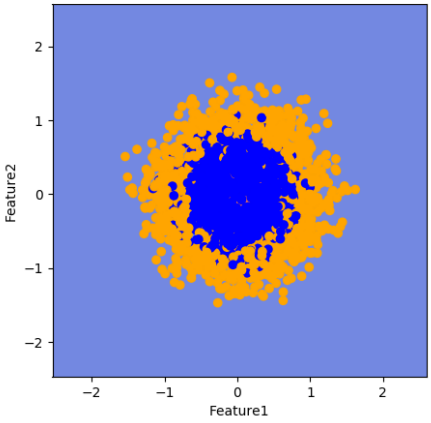
Dataset 1



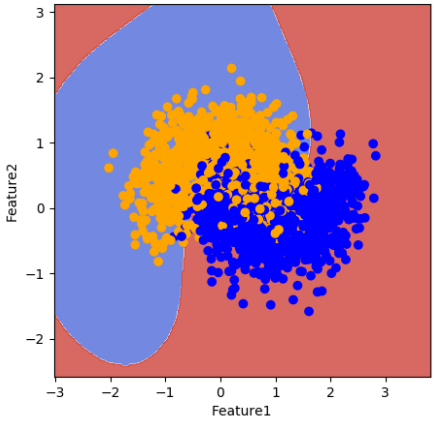
Dataset 2



Dataset 3



Dataset 4



Dataset 5

3. SVM results:

	Linear SVM (MAE)	RBF SVM (MAE)
Dataset 1	0.68421052631	0.0
Dataset 2	0.15789473684	0.0
Dataset 3	0.0	0.0
Dataset 4	0.526315789474	0.107769423559
Dataset 5	0.172932330827	0.16290726817

Above results are MAE on the given datasets using SVM technique with linear as well as RBF kernel.
SVM was trained using the 80% of the given data, and was tested on the remaining 20% of the data.

Dataset 1
For this dataset, SVM with RBF kernel gives the best results with 0 MAE. However, the data was not linearly seperable, so the predictions made by linear SVM are not so good.

Dataset 2
For this dataset, Linear SVM doesn't perform that bad, since the data was not linearly separable but a linear boundary can be made between the classes with some error. However, SVM with RBF kernel performs best here as well.

Dataset 3
Since the data was already linearly separable, both the kernels give best results.

Dataset 4
This dataset was visually similar to the dataset1, therefore, the results with the linear SVM for this dataset are similar to the dataset1. However, SVM with RBF kernel gives much better results. The lack in accuracy of the RBF kernel is only due to the noise/outliers present in the data.

Dataset 5
This dataset was visually similar to the dataset2 , and its results are again similar to that of dataset2. Again, lack in accuracy of SVM with RBF kernel is due to the noise/outliers present in the data.

4. Question on Kaggle

The data given consisted of variable length feature vectors and label vectors with 5 labels.

Current MAE on kaggle : 0.80921

Approach:
- Remove the same data (same features and same corresponding label)
- Treat the feature vectors as strings of variable length.
- Use bag of words model to vectorise the features.
- After vectoring the features, classify them
- Make predictions on the test data.

Preprocessing used:
Removal of redundant data, i.e., data samples which consisted of same feature vector and corresponding label, were removed.

Techniques:
CountVectorizer Classifier of sklearn module was used for training of model.

Other approach tried:
- Treat each integer present in the feature vectors as a different feature.
- Create a matrix for all those features(>90000) along with the samples.
- Using sklearn feature selection techniques(like SelectKfeatures) to reduce the number of features.
- Training SVM on this final data.
- Predicting labels for test data.

Problem with this approach was the amount of memory and computation required to train the model.