Homework 2

Wei Xu, Student No:117034910134

I. Problem 1

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

In this question, it is a three classification problem and we adopt the one-vs-rest (ovr) strategy. It means we should train three SVM models to classify. The ovr strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label; discrete class labels alone can lead to ambiguities, where multiple classes are predicted for a single sample. We first classify the data as follows:

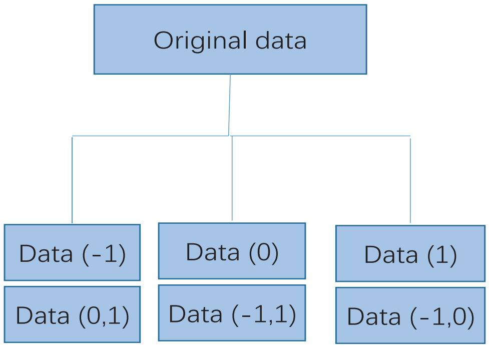

Fig. 1: data classify

Then we use preprocessing.StandardScaler() in sklearn to standardize features by removing the mean and scaling to unit variance. This will improve the performance for SVM. We use the standard algorithm to train three SVM models. We first show the original data in 2-dimension by using dimensionality reduction.

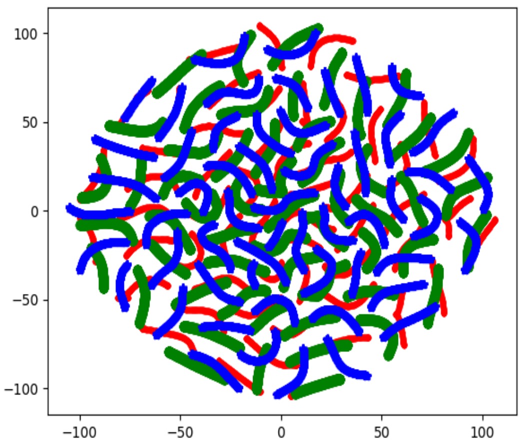

Fig. 2: Original data after dimensionality reduction

We clearly can see after dimensionality reduction into 2-dimension, the classification is hard to solve. It seems all the data are mixed together.

After training three models, we can get the accuracy of our SVM based on ovr. Our model can get 43.5% for this three classification problem.

II. Problem 2

For the three-class classification problem, we use Min-Max-Module SVM and part-vs-part task decomposition method. We first divide the three-class problem into three two-class problems using one-vs-rest method and then decompose these imbalance two-class problems into balance two-class problems following random task decomposition.

In our simulation, we choose randomly 1000 sample of raw data five times and randomly choose 1000 sample of label data two times. Thus, we should train 10 SVM models and use Min-Max-Module to get the result. The scheme of Min-Max-Module can be shown as follows:

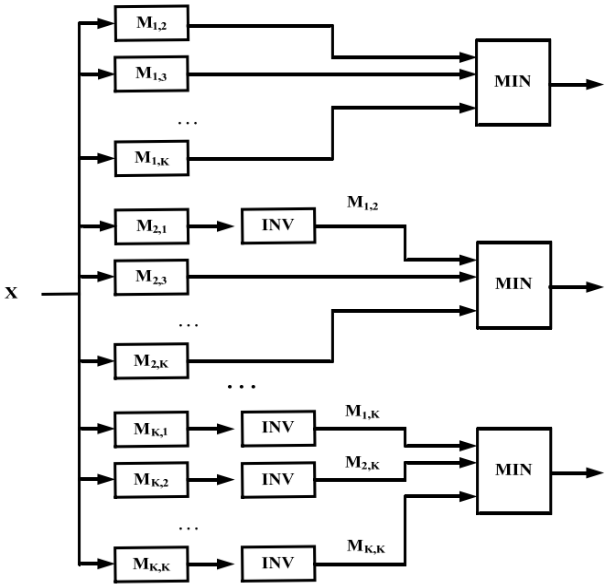

Fig. 3: Min-Max-Module

After training these models, we can get the accuracy of our SVM based on part-vs-part task decomposition and Min-Max-Module. Our model can get 47.5% for this three classification problem. Compared with Problem 1, we improve 4$\sim$5% of accuracy.