DATA MINING ASSIGNMENT 9
(SVM and K-Means)

Debarati Das

1PI13CS052

OVERVIEW (PART A)

Super Vector Machine or Super Vector Network is a Supervised Learning Model with associated learning algorithm that analyzes data and recognizes patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classification.

The goal of a support vector machine is to find the optimal separating hyperplane which maximizes the margin of the training data. The first thing we can see from this definition, is that a SVM needs training data. Which means it is a supervised learning algorithm. It is also important to know that SVM is a classification algorithm. Which means we will use it to predict if something belongs to a particular class. The objective of a SVM is to **find the optimal separating hyperplane** because it correctly classifies the training data and because it is the one which will generalize better with unseen data. We can make the following observations:

- If an hyperplane is very close to a data point, its margin will be small.
- The further an hyperplane is from a data point, the larger its margin will be.

The Iris flower data set is a multivariate dataset which quantifies the structural variation of three related species of Iris flower.
Thus classification is done on the basis of flower species which are:
Iris-setosa---------->Blue
Iris-versicolor ---------->Red
Iris-verginica ---------->Cyan

The multi-class SVM will be implemented by LIBSVM library. LIBSVM implements the SMO algorithm for kernelized support vector machines(SVMs), supporting classification and regression. LIBSVM implement one against one strategy for multi-class implementation. LIBSVM to build SVM classes uses the one against one strategy, also known as "pairwise coupling", "all pairs" or "round robin", consists in constructing one SVM for each pair of classes. Thus, for a problem with c classes, c(c-1)/2 SVMs are trained to distinguish the samples of one class from the samples of another class. Usually, classification of an unknown pattern is done according to the maximum voting , where each SVM votes for one class.

ANSWERS TO QUESTIONS

1. **Show the confusion matrix.**

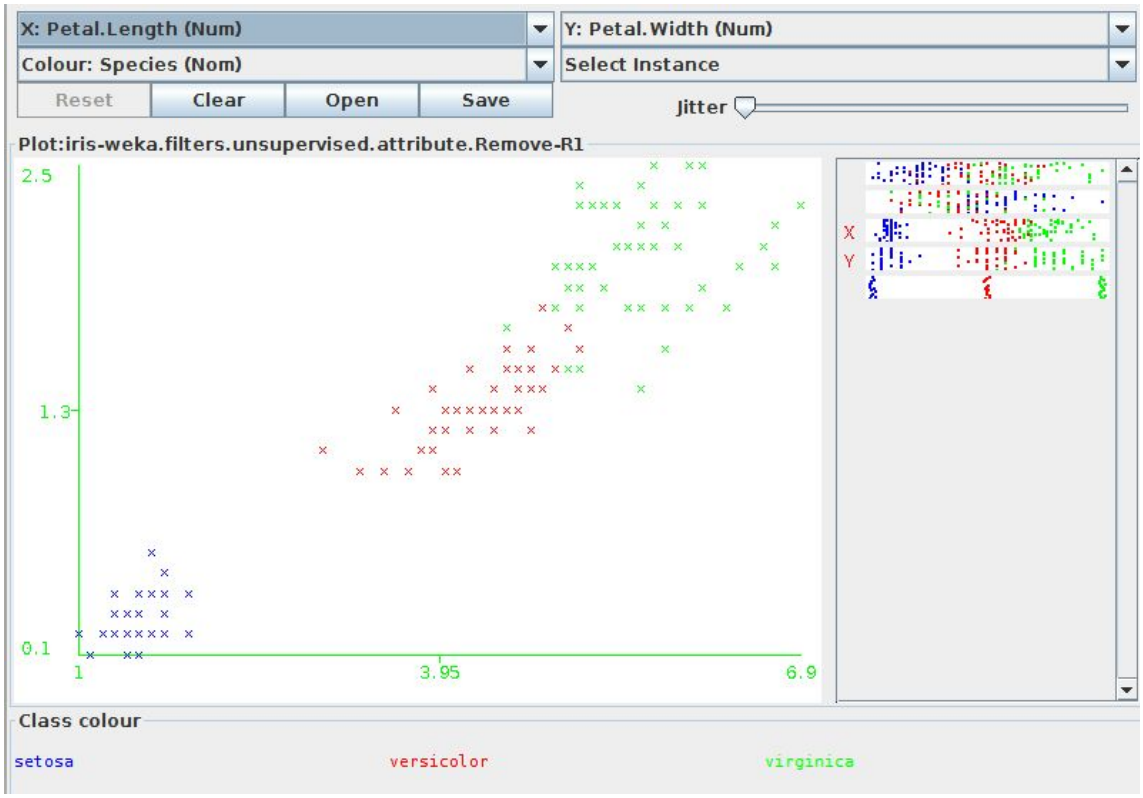
Default options : (with Radial Kernel and no Normalisation)

```
=== Confusion Matrix ===

 a  b  c   <-- classified as
50  0  0 |  a = setosa
0 47  3  |  b = versicolor
0  2 48  |  c = virginica
```

2. **Which class is best identified using the SVM? Explain.**

We can see from the below image that class 'Iris setosa' is linearly separable and other two classes are not. Thus, a dataset like Iris is linearly not separable and hence could be a best example to implement SVM. Since Iris Setosa is Linearly separable , it will be best identified by the SVM.



3. Is there any effect of the choice of the type of kernel? Explain.

Kernel is a function that returns the value of the dot product between images of two arguments. There are four kinds of kernel available for selection.

1. Linear
2. Polynomial
3. Radial
4. Sigmoid

Radial basis kernel function is most popular and most widely used from all.

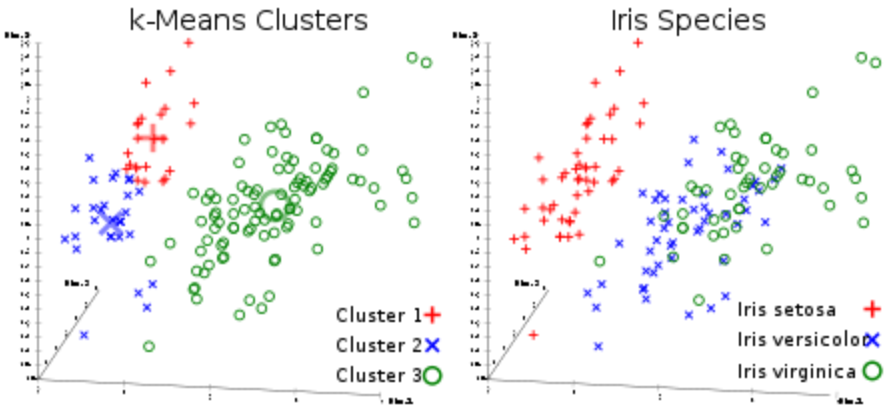
Different Kernel Functions will generate different confusion matrix. In general, the RBF kernel is a reasonable first choice. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

4. **Which kernel gives the lowest accuracy and is there any way to increase its accuracy?**

The sigmoid kernel gives the worst accuracy of 6% with the default options. The polynomial kernel also gives a high error rate but the sigmoid kernel gives the highest error rate of 78%. A non-linear or complicated kernel is actually not necessary for an easily-classified example like the iris flower data set.
The SVM only weighs instances (support vectors), not features. Normalizing the features properly is a big issue.Pre-processing of features (like normalization and various transformations) can improve accuracy. I changed the normalisation to true leading to the accuracy coming up to 92% from 6%. Further changing the SVM classification to nu-Classification, changed the accuracy to 96%.

OVERVIEW (PART B)

Clustering is the process of partitioning a group of data points into a small number of clusters. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have n data points xi,i=1...n that have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μi,i=1...k of the clusters that minimize the square of the distance from the data points to the cluster. The K-means clustering uses the Euclidean distance d(x,μi)= // x−μi // 2. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.



ANSWERS TO QUESTIONS

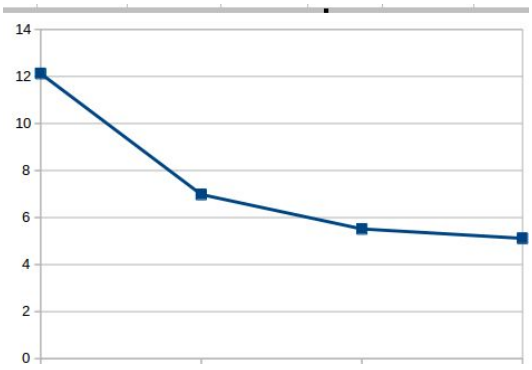
The Table expected for a) is :

K	SSW	SSB	SST	SSB/SST
2	12.1277907505	29.0383196708	41.1661104214	0.7053938148
3	6.9822164738	34.1838939476	41.1661104214	0.8303892109
4	5.516933472	35.6491769493	41.1661104214	0.8659836109
5	5.114887116	36.0512233054	41.1661104214	0.8757500511

b)What are your observations based on this table?

SSB/SST gives a measure of variance between clusters .The total within-cluster sum of square (SSW) measures the compactness of the clustering and we want it to be as small as possible. We can see that k=2 gives maximum in cluster error and minimum variance. k=3 gives a lesser error (6.9) and a higher variance which is desirable. The value of SST is remaining constant in all the cases.

On plotting SSW on the Y axis and K on the X axis we observe the "elbow" in the graph or the change is observed when k=3. This is an indication that k=3 is an optimal number of clusters.



c) For each value of k show a table that shows a cross tabulation between the cluster number and class.

K	Cluster number	Class
2	0	Versicolor
	1	Setosa
3	0	Versicolor
	1	Setosa
	2	Virginica
4	0	Versicolor
	1	Versicolor
	2	Virginica
	3	Setosa
5	0	Versicolor
	1	Versicolor
	2	Setosa
	3	Virginica
	4	Versicolor

d) Based on the table in c) above, what is the optimum number of clusters? Explain.

Since there are three classes known of the Iris Dataset, and the case k=3 yields 3 unique classes, it seems that it would be better to take k=3 as the optimal number. On checking with SSW vs k graph also yields k=3 as optimal number.
When we take k=2 class virginica is skipped out, while taking values >3 is yielding non unique classes.