DM Assignment 09

15.3.2016

Adithya Abraham Philip

1PI13CS008

Overview

This document answers questions related to the use of LibSVM and Simple K Means clustering on the Iris dataset, using Weka.

Questions & Answers

PART A

a. Show the confusion matrix

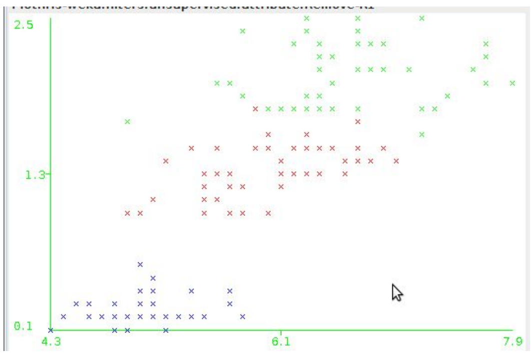
Ans:

setosa (classified as)	versicolor (classified as)	virginica (classified as)	
50	0	0	setosa (actual)
0	47	3	versicolor (actual)
0	2	48	virginica (actual)

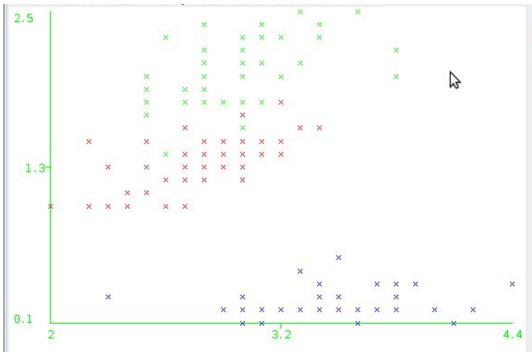
b. Which class was best identified by the SVM? Explain.

Ans: Setosa was best identified by the SVM, as seen in the confusion matrix (50 out of 50 correctly classified. The reason is that objects of Setosa class are clustered together, and their cluster is at a larger distance (with respect to sepal length, width and petal length, width) from instances of the other two classes. On the other hand, the boundary between the instances of the other two classes are a little blurred as they are not clearly separable visually. This can be seen in the following images: (Setosa - blue, Versicolor - red, Virginica - green)

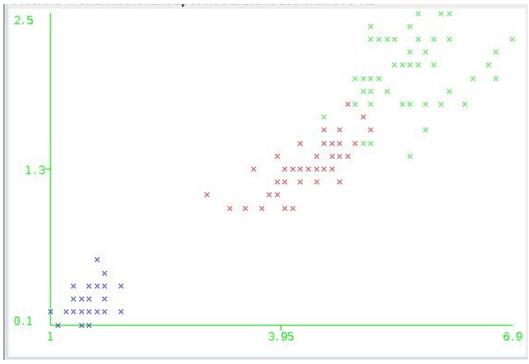
Sepal length



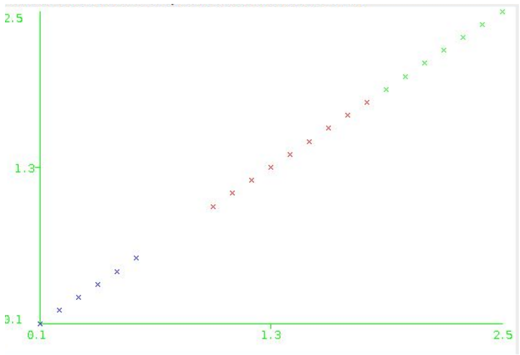
Sepal Width



Petal Length



Petal Width



c. Does the choice of kernel have an effect? Explain.

Ans: Yes, different kernels misclassify differently, and some have worse accuracy than others. (in this case all except the sigmoid function has the same accuracy, while the sigmoid function has very low accuracy initially). This is expected as choosing a different kernel implies a different function is used to draw boundaries between the different classes of data, and consequently decide to which class an unknown instance is classified. This function may do better or worse depending on the nature and distribution of data from the different classes.

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

The sigmoid kernel gives the lowest accuracy. Its accuracy can be improved by normalizing the values of the attributes being used for the distance measure (accuracy increases from 6.67% to 92.67%). This is because the sigmoid function maps values between 0 and 1, and consequently works best with values in that range.

PART B

a. Report the results of the K-Means clustering.

Ans: **This table shows when values are not normalized before calculating Euclidian distance**

No. of Clusters	SSW	SSB	SST	SSB/SST
2	154.95	526.42	681.37	0.77
3	79.33	602.04	681.37	0.88
4	60.06	621.31	681.37	0.91
5	55.44	625.93	681.37	0.92

This table shows when values are normalized the before calculating Euclidian Distance

No. of Clusters	SSW	SSB	SST	SSB/SST
2	12.13	29.04	41.17	0.71

3	6.98	34.18	41.17	0.83
4	5.51	35.65	41.17	0.87
5	5.11	36.05	41.17	0.88

b. What are your observations based on this table?

Ans: As the number of clusters increases the cohesion (can be thought of as inversely proportional to SSW) and separation increases.

c. For each value of k show a cross tabulation between cluster number and class.

Ans:

K = 2

Class (down), Cluster (across)	0 (No class)	1 (Setosa)
Setosa	0	50
Versicolor	50	0
Virginica	50	0

K = 3

Class (down), Cluster (across)	0 (Versicolor)	1 (Setosa)	2 (Virginica)
Setosa	0	50	0
Versicolor	47	0	3
Virginica	14	0	36

K = 4

Class (down), Cluster (across)	0 (Versicolor - if multiple assignments allowed)	1 (Versicolor)	2 (Virginica)	3 (Setosa)
Setosa	0	0	0	50
Versicolor	23	27	0	0
Virginica	19	2	29	0

K = 5

Class	0 (Virginica -	1	2 (Virginica)	3 (Setosa)	4 (Versicolor

(down), Cluster (across)	if multiple assignments allowed)	(Versicolor)			- if multiple assignments allowed)
Setosa	0	0	0	50	0
Versicolor	10	25	0	0	15
Virginica	17	1	27	0	5

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

Ans: **If more than one cluster can be associated with the same class** k=5 gives us the most accuracy as assigning 0 to Virginica and 4 to Versicolor gives us 15 misclassified tuples which is the least percentage of misclassification.

If only one cluster can be associated with a given class then k=3 gives us the most accuracy with 17 misclassified tuples. This is generally desired.