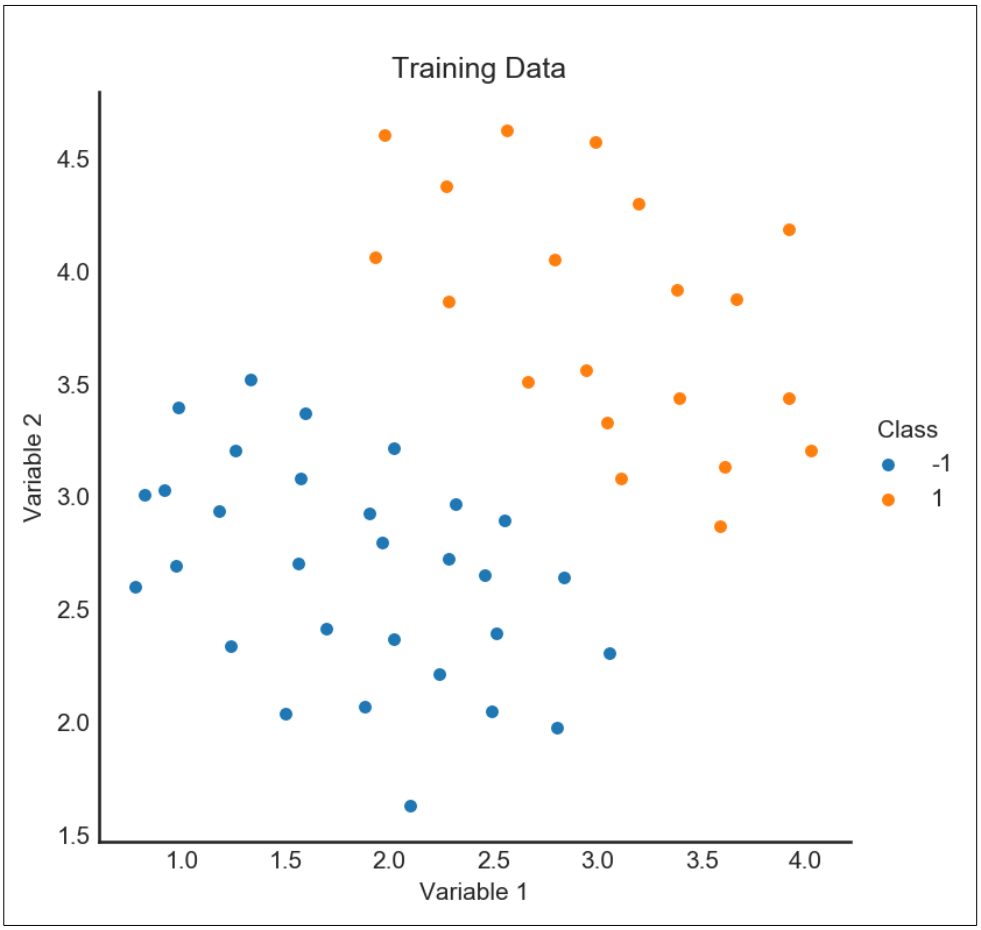
CLL 788 – Assignment 3
Solution

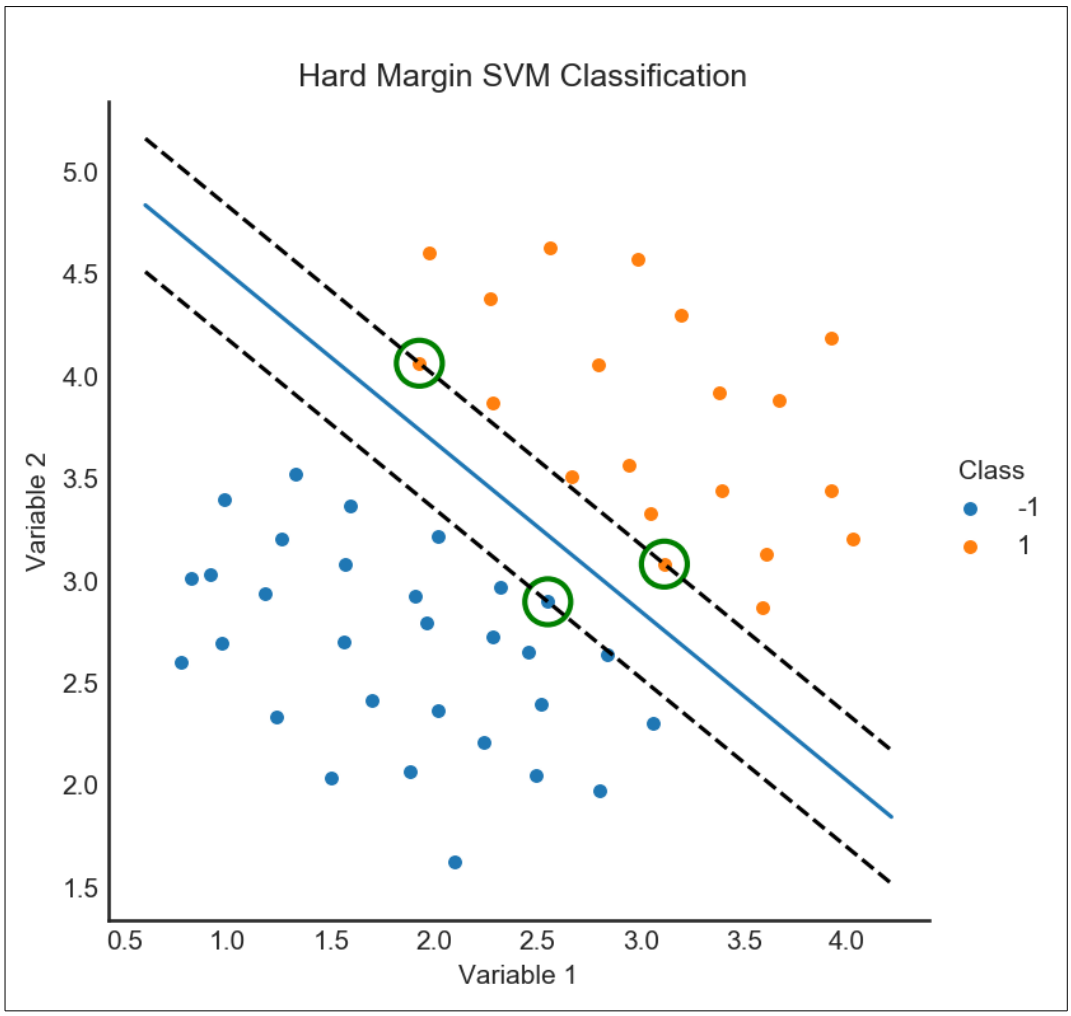
Q1

While Visualizing the Data 1.xlsx using seaborn, it can be seen that the two feature sets can be classified linearly with the help of SVM. Since there is no overlapping of features , the regularization is not required. (C=0)



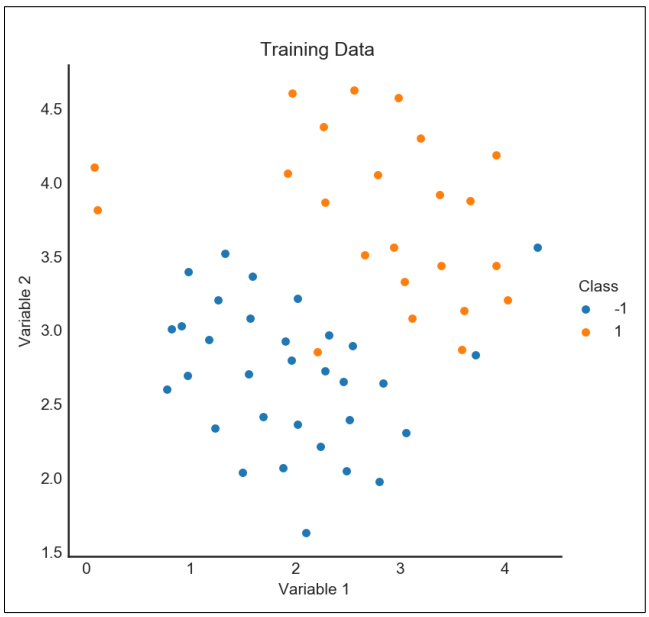
Q2.

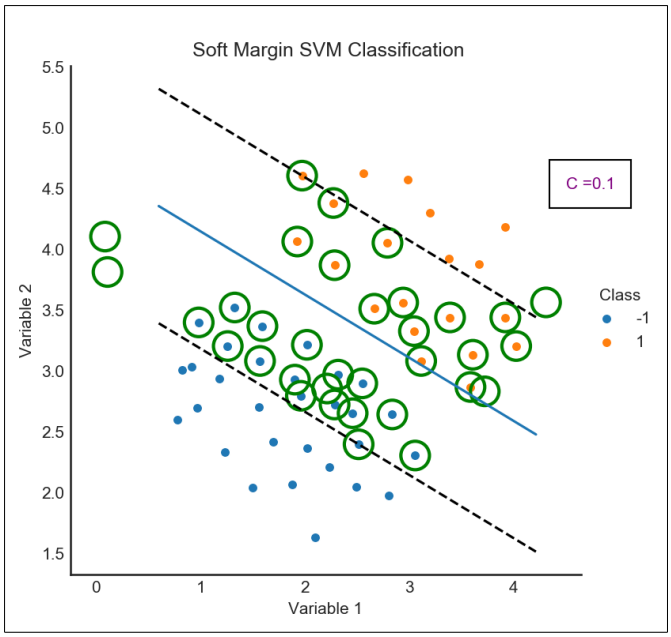
The SVM is trained using CVXOPT library available in python which is a quadratic optimizer for complex sets. It can be seen that after SVM training and classification, the number of support vectors are only **3 out of 50** sample feature points. The parameters for decision function can also be calculated and it can be plotted as SVM.

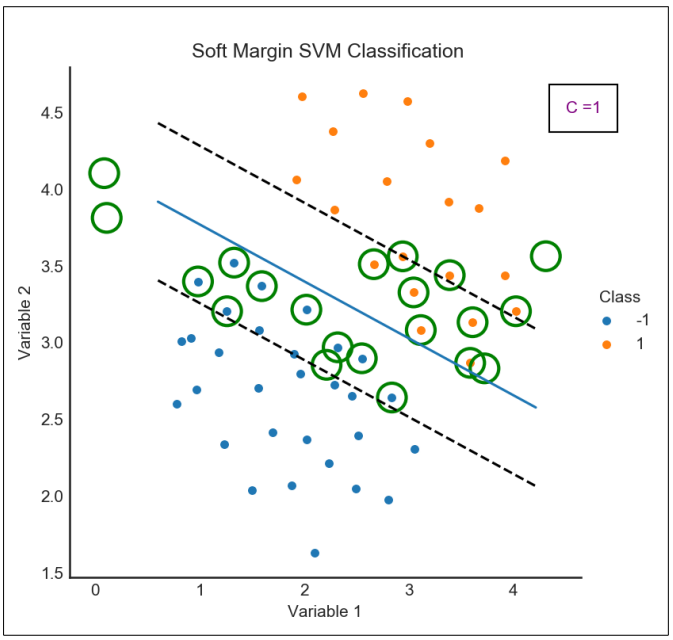


Q3.

While visualizing the samples points in Data2.xslsx , it can be seen that some feature overlapping exists. Hence, we need to train a SVM with soft margin where regularization paramter C needs to be set for margin relaxations. A large value of C results in smaller margins and less number of support vectors. A small value of C results in larger margins and higher number of support vectors. It can be seen that for the given data set a value of C=0.1 results in 21 support vectors whereas for C=1, there are 14 support vectors.







Q4.

iit delhi

ASSIGNMENT-3

Q4 :- Classify Red Domestic Suv using Naive Bayes classifier

Solution :- Given the data, let's calculate all the probabilities :-

1. $P(\text{Stolen} = \text{Yes}) = 5/10$ $\quad P(\text{Stolen} = \text{No}) = 5/10$
2. $P(\text{Red} = \text{Yes}) = 5/10$ $\quad P(\text{Yellow} = \text{Yes}) = 5/10$
3. $P(\text{Sports} = \text{Yes}) = 6/10$ $\quad P(\text{Suv} = \text{Yes}) = 4/10$
4. $P(\text{Domestic} = \text{Yes}) = 5/10$ $\quad P(\text{Imported} = \text{Yes}) = 5/10$

Now, let's calculate all conditional probabilities :-

$P(\text{Stolen} = \text{Yes}) = 5/10$ $\quad P(\text{Stolen} = \text{No}) = 5/10$

1. Color

$P(\text{Red} = Y \mid \text{Stolen} = Y) = 3/5$
$P(\text{Red} = N \mid \text{Stolen} = Y) = 2/5$
$P(\text{Yellow} = Y \mid \text{Stolen} = Y) = 2/5$
$P(\text{Yellow} = N \mid \text{Stolen} = Y) = 3/5$

1. Color

$P(\text{Red} = Y \mid \text{Stolen} = N) = 2/5$
$P(\text{Red} = N \mid \text{Stolen} = N) = 3/5$
$P(\text{Yellow} = Y \mid \text{Stolen} = N) = 3/5$
$P(\text{Yellow} = N \mid \text{Stolen} = N) = 2/5$

2. Type

$P(\text{Sports} = Y \mid \text{Stolen} = Y) = 4/5$
$P(\text{Sports} = N \mid \text{Stolen} = Y) = 1/5$
$P(\text{Suv} = Y \mid \text{Stolen} = Y) = 1/5$
$P(\text{Suv} = N \mid \text{Stolen} = Y) = 4/5$

2. Type

$P(\text{Sports} = Y \mid \text{Stolen} = N) = 2/5$
$P(\text{Sports} = N \mid \text{Stolen} = N) = 3/5$
$P(\text{Suv} = Y \mid \text{Stolen} = N) = 3/5$
$P(\text{Suv} = N \mid \text{Stolen} = N) = 2/5$

3. Origin

$P(D = Y \mid \text{Stolen} = Y) = 2/5$
$P(D = N \mid \text{Stolen} = Y) = 3/5$
$P(I = Y \mid \text{Stolen} = Y) = 3/5$
$P(I = N \mid \text{Stolen} = Y) = 2/5$

3. Origin

$P(D = Y \mid \text{Stolen} = N) = 3/5$
$P(D = N \mid \text{Stolen} = N) = 2/5$
$P(I = Y \mid \text{Stolen} = N) = 2/5$
$P(I = N \mid \text{Stolen} = N) = 3/5$

Now :- for a Red, domestic, SUV :-

$$P(\text{stolen} = \text{Yes}) = \left(\frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{5}{10}\right) \div \left(\frac{5}{10} \times \frac{5}{10} \times \frac{4}{10}\right) = 0.24$$

$$P(\text{stolen} = \text{No}) = \left(\frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{5}{10}\right) \div \left(\frac{5}{10} \times \frac{5}{10} \times \frac{4}{10}\right) = 0.72$$

Since $P(\text{stolen} = \text{No}) > P(\text{stolen} = \text{Yes})$
we can say that Red, domestic, SUV will not be stolen.

Q5.

iit delhi

ASSIGNMENT - 3

Qs. Manually perform K Means clustering on 10 data points into two clusters.

Solution :- Given :-

S.No/Sample	Var 1	Var 2
1	-1.54	2.29
2	-0.44	2.34
3	0.03	0.41
4	1.2	1.87
5	0.65	2.39
6	-4.67	-4.8
7	-3.37	-5.41
8	-3.93	-4.64
9	-4.78	-4.96
10	-4.12	-5.36

EU1 : Eucledian distance of a sample point from mean of cluster 1
EU2 : Eucledian distance of a sample point from mean of cluster 2.

Step 1 : Separating two points on the basis of distance between the sample points ie creating two cluster centres

Step 2 : Calculate EU1 and EU2 from means of both clusters and find the shortest distance. Classify the sample to the cluster with shortest distance.

Step 3 : Recalculate the mean of new cluster.

Step 4 : Repeat Step 2 for next sample point.

Step 5 : Repeat step 3 till all samples are classified.

Cluster 1		Cluster 2	
Sample No.	Mean	Sample No.	Mean
4	1·2, 1·87	10	-4·12, -2·36
1, 4	-0·17, 2·08	10	-4·12, -2·36
1, 2, 4	-0·26, 2·16	10	-4·12, -2·36
1, 2, 3, 4	-0·18, 1·72	10	-4·12, -2·36
1, 2, 3, 4, 5	-0·02, 1·86	10	-4·12, -2·36
1, 2, 3, 4, 5	-0·02, 1·86	6, 10	-4·39, -3·58
1, 2, 3, 4, 5	-0·02, 1·86	6, 7, 10	-4·05, -4·19
1, 2, 3, 4, 5	-0·02, 1·86	6, 7, 8, 10	-4·02, -4·30
1, 2, 3, 4, 5	-0·02, 1·86	6, 7, 8, 9, 10	-4·17, -4·43

* Sample 4 and 10 are selected as initial two points and then the classification of other samples into two clusters is done.

Sample No.	Var 1	Var 2	EU1	EU2	Cluster
1	-1·54	2·29	2·77	5·32	1
2	-0·44	2·34	0·37	5·97	1
3	0·03	0·41	1·78	4·99	1
4	1·2	1·87	1·39	6·79	1
5	0·65	2·39	1·06	6·73	1
6	-4·67	-4·8	8·10	2·5	2
7	-3·37	-5·41	8·00	2·00	2
8	-3·97	-4·64	7·58	0·46	2
9	-4·78	-4·96	8·31	1·00	2
10	-4·12	-5·36	8·30	0·92	2

Formula:-

$$EU = \sqrt{(X_1 - CMx_1)^2 + (X_2 - CMx_2)^2}$$

where X_1, X_2 = Sample point, CMx_1, CMx_2 = Cluster Mean

New Mean = $\frac{\text{Old points + New point}}{\text{No. of points}}$

Hence the two clusters are

Cluster	Samples	Mean of Cluster
1	1, 2, 3, 4, 5	-0.02, 1.86
2	6, 7, 8, 9, 10	-4.17, -4.43

As a check, it can also be seen that if the Euclidean distance from new means of finalized clusters is calculated, the sample points should have smaller distance from the mean of their own clusters.