

Statistics 503 Exam 2

SOLUTION Open notes

1. (3pts) The following code fits a neural network model to the data:

```
tmp.dat<-d.olive[d.olive[,1]==3,]
indx<-c(sample(c(1:51),35),sample(c(52:101),34),sample(c(102:151),34))
tmp.dat.tr<-tmp.dat[indx,]
tmp.dat.ts<-tmp.dat[-indx,]
olive.nn<-nnet(tmp.dat.tr[,3:10],tmp.dat.tr[,2],size=6,linout=T,decay=5e-4,
  range=0.6,maxit=1000)
olive.nn.ts<-nnet(tmp.dat.ts[,3:10],tmp.dat.ts[,2],size=6,linout=T,decay=5e-4,
  range=0.6,maxit=1000)

table(tmp.dat.tr[,2],round(predict(olive.nn,tmp.dat.tr[,3:10])))
table(tmp.dat.ts[,2],round(predict(olive.nn.ts,tmp.dat.ts[,3:10])))
```

It results in these missclassification tables:

Training				Test			
7	8	9		7	8	9	
7	34	0	0	7	16	0	0
8	0	34	0	8	0	16	0
9	0	0	35	9	0	0	16

giving training and test error of zero. Why is the test error so small?

The test error is calculated from a model built on the test data instead of the training data.

2. (3pts) This missclassification table for the test data results from running SVM to classify cancer in samples of tissue.

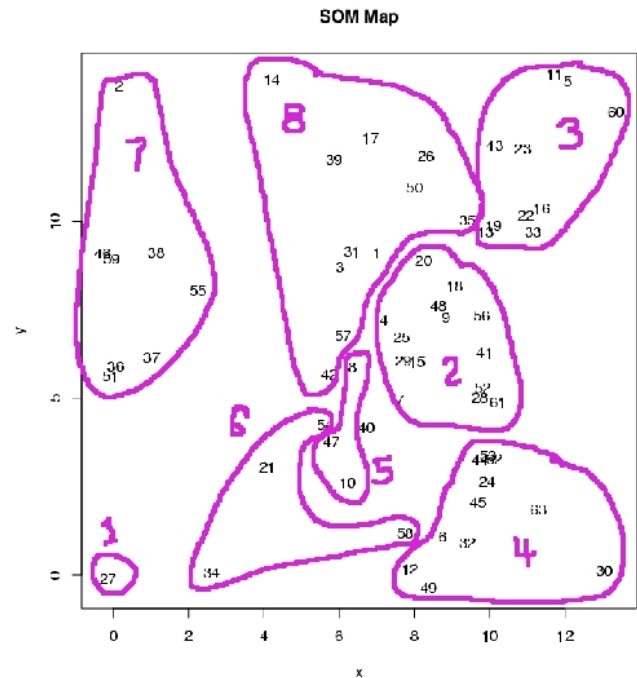
		Predicted	
		Normal	Cancerous
True	Normal	980	20
	Cancerous	100	100

Compute the overall test error. Compute the test error separately for normal and cancerous tissue. What's wrong with the classifier?

Overall test error is 10%, for normal tissue the test error is 2%, and for cancerous tissue the test error is 50%. There is too much error in predicting cancerous tissue. This error needs to be 0 for a useful classification rule.

3. (2pts) Roughly circle the results of k -means clustering on the SOM map.

1	2	3	
[27] Eternal Sunshine of the Spotless Mind	[7] Being Julia [9] Birth [15] Cellular [18] Club Dread [20] Cursed [25] Ella Enchanted [28] Eulogy [29] EuroTrip [40] I Heart Huckabees [41] Incident at Loch Ness [42] Jersey Girl [48] Laws of Attraction [56] Mix [61] Ono	[5] At Night with No Curtains [11] Camp Slaughter [13] Catch That Kid [16] Charlie [19] Confessions of a Teenage Drama Queen [22] Decoys [23] Drum [33] Full Clip [43] Johnson Family Vacation [60] November	
4	5	6	
[6] Before Sunset [12] Cape of Good Hope [24] Eating Out [30] Exquisite Corpse [32] Four Dead Batteries [44] Josh Jarman [45] Khwaab [49] Less Like Me [53] Mean Creek [62] Papa [63] Red Cockroaches	[8] Beyond the Sea [10] Cachimba [31] Flight of the Phoenix [47] Kinsey [57] Modigliani	[21] Dawn of the Dead [34] Garden State [54] Mean Girls [58] Napoleon Dynamite	
7	[2] Alexander [36] Harry Potter and the Prisoner of Azkaban [37] Hellboy [38] Hidalgo [46] King Arthur [51] Man on Fire [55] Meet the Fockers [59] National Treasure		
8	[1] After the Sunset [3] Alfie [4] Along Came Polly [14] Catwoman [17] Christmas with the Kranks [26] Envy [35] Garfield [39] Home on the Range [50] Little Black Book [52] Mars		



4. (2pts) We generate data from two concentric circles which form two clusters, and fit a radial kernel SVM classifier. Cases 1-15 are in class 1, the inner circle, and cases 16-40 are in class 2, the outer circle. As best you can, identify which points are the unbounded support vectors from this model fit.

```
> tmp.svm
```

Call:

```
svm(formula = factor(cl) ~ ., data = x4, kernel = "radial")
```

Parameters:

```
SVM-Type: C-classification
SVM-Kernel: radial
cost: 1
gamma: 0.5
```

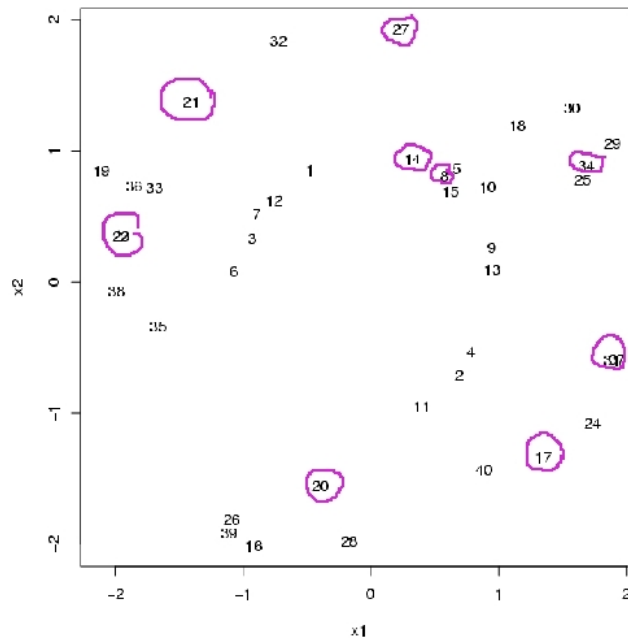
Number of Support Vectors: 22

```
> tmp.svm$index
```

```
[1] 2 3 4 5 6 7 8 10 11 14 17 18 20 21 23 25 27 31 33 34 35 40
```

```
> tmp.svm$coefs
```

```
      [,1]
[1,] 1.0000000
[2,] 1.0000000
[3,] 1.0000000
[4,] 1.0000000
[5,] 1.0000000
[6,] 1.0000000
[7,] 0.4096640
[8,] 1.0000000
[9,] 1.0000000
[10,] 0.6790273
[11,] -0.1466926
[12,] -1.0000000
[13,] -0.9612008
[14,] -0.6772297
[15,] -0.2331215
[16,] -1.0000000
[17,] -0.7256547
[18,] -0.9897454
[19,] -1.0000000
[20,] -0.3550466
[21,] -1.0000000
[22,] -1.0000000
```



Isn't it interesting to find that the support vectors are not the most "extreme" in the circle?" For example, I would have expected that 10 was chosen instead of 8 or 14, and 18 chosen in preference to 34!

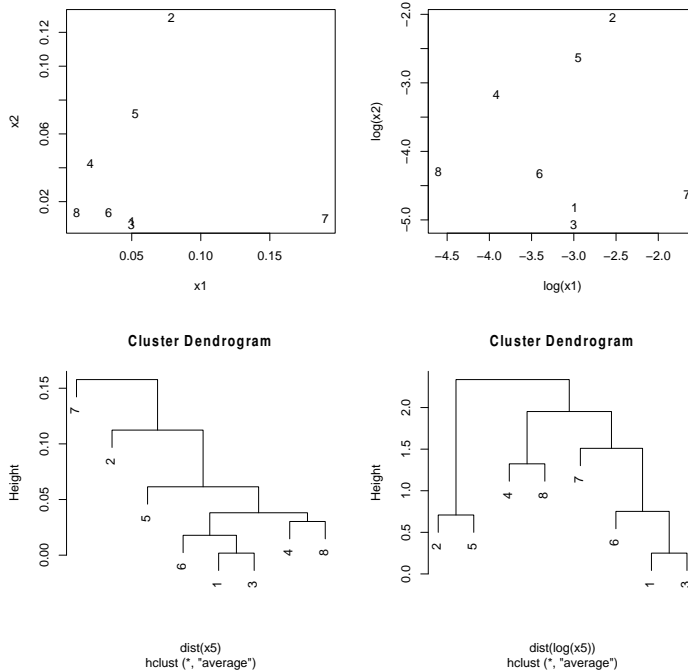
5. (3pts) Match up the clusters as best possible for the two clustering methods in this confusion table. Write down the mapping of cluster labels, and the rearranged confusion matrix.

	km			
hclust	1	2	3	4
1	5	23	0	0
2	0	0	8	0
3	2	0	0	4
4	21	0	0	0

	km			
hclust	1	2	3	4
4	21	0	0	0
1	5	23	0	0
2	0	0	8	0
3	2	0	0	4

$1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4, 4 \rightarrow 1.$

6. (3pts) In the music data, and the movies data that we've looked at this semester, several of the variables were highly skewed. In the plot below, we have conducted cluster analysis on both raw data where the variables have skewed distributions and the logged data. Why do the results differ? What is the impact of skewed variables on the results of cluster analysis? Which result do you prefer in this example?



The results differ because the euclidean distance between points is different. When variables are skewed there will be more singleton clusters in the solution, and more large clusters.

7. We constructed a simulated data set to test the effect of the number of variables relative to the number of cases and model complexity. The data contains separated classes in the first variable (X_1) but all other variables ($X_2 - X_{50}$) are samples from Gaussian (normal) distributions, all similar to X_2 plotted below. There are 20 observations.

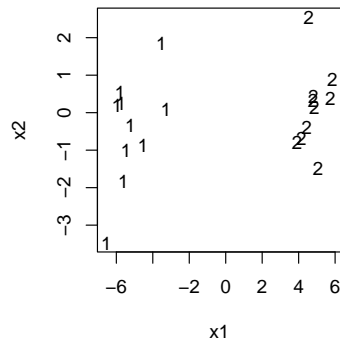
(a) (2pts) We fit the SVM model to variables X_1, X_2 . Three cases are chosen as support vectors. Write down the SVM model for this data.

```
> tmp.svm$index
[1] 5 13 18
> tmp.svm$coefs
      [,1]
[1,] 0.999
[2,] -0.763
[3,] -0.236
> tmp.svm$SV
      x6      V2
5 -0.598  0.185
13  0.787 -0.492
18  0.909  2.091
>
> tmp.svm$rho
[1] -0.144
```

$$\hat{y} = -0.144 - 1.41x_1 + 0.067x_2$$

(b) (2pts) Following this we fit the SVM model for increasing numbers of variables. This table shows the number of support vectors as number of variables increases. What is the pattern in the number of support vectors relative to the number of variables? Why do you think this occurs? Would you expect the test error would increase or decrease with the increasing number of variables?

	Number of variables				
	2	5	10	25	50
Num SVs	3	6	11	14	18



The number of support vectors increases with increasing number of variables. Its not clear to me why this is. One possible explanation is that SVM borrows a little from each new variable to improve the fit to the training data. But what it borrows is due to random sampling which should make the test error worse. The test error would likely increase with increasing complexity.