USC ID: 9907399097

Name: Subhiksha Rani

Homework-5 Report

1. Multi-class and Multi-Label Classification Using Support Vector Machines.
    a. Downloading the data set & assigning 70% of the data to Training-set & remaining to test set.
    b. Each instance has three labels: Families, Genus, and Species. Each of the labels has multiple classes.
        i. Hamming-Loss (Example based measure): It is the fraction of labels that are incorrectly predicted, i.e., the fraction of the wrong labels to the total number of labels.

$$\frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} \text{xor}(y_{i,j}, z_{i,j}), \text{ where } y_{i,j} \text{ is the target and } z_{i,j} \text{ is the prediction.}$$

Exact Match Ratio (Subset accuracy): It indicates the percentage of samples that have all their labels classified correctly. The disadvantage of this measure is that multi-class classification problems have a chance of being partially correct, but here we ignore those partially correct matches.

$$ExactMatchRatio, MR = \frac{1}{n} \sum_{i=1}^{n} I(Y_i = Z_i)$$

        ii. SVM using Gaussian kernels and one versus all classifiers

For Family label with raw attributes:
Best C: 100
Best Gamma: 1
Hamming Loss: 0.006484483557202408
Exact Match ratio: 0.9935155164427976

For Genus label with raw attributes:
Best C: 100
Best Gamma: 1
Hamming Loss: 0.007874015748031496

Exact Match ratio: 0.9921259842519685

For Species label with raw attributes:
Best C: 10
Best Gamma: 1
Hamming Loss: 0.010653080129689671
Exact Match ratio: 0.9893469198703103

For Family label with standardized attributes:
Best C: 10
Best Gamma: 0.1
Hamming Loss: 0.007410838351088467
Exact Match ratio: 0.9925891616489115

For Genus label with standardized attributes:
Best C: 10
Best Gamma: 0.1
Hamming Loss: 0.012042612320518759
Exact Match ratio: 0.9879573876794813

For Species label with standardized attributes:
Best C: 1000
Best Gamma: 0.01
Hamming Loss: 0.010653080129689671
Exact Match ratio: 0.9893469198703103

iii. L1-penalized SVMs

For Family label:
Best C: 1
Hamming Loss: 0.07132931912922649
Exact Match ratio: 0.9286706808707735

For Genus label:
Best C: 10
Hamming Loss: 0.058360352014821676
Exact Match ratio: 0.9416396479851783

For Species label:
Best C: 1
Hamming Loss: 0.04075961093098657

Exact Match ratio: 0.9592403890690134

iv. SMOTE to deal with class imbalance - Oversampling minority class using SMOTE.
<u>For Family label:</u>
Best C: 10
Hamming Loss: 0.0921723019916628
Exact Match ratio: 0.9078276980083372

<u>For Genus label:</u>
Best C: 10
Hamming Loss: 0.0968040759610931
Exact Match ratio: 0.9031959240389069

<u>For Species label:</u>
Best C: 100
Hamming Loss: 0.042612320518758684
Exact Match ratio: 0.9573876794812413

From the above trained models, we can see that the Hamming Loss is less than or equal to 0.09 for all the models (which is closer to 0), indicates a good score. Also, the Exact Match ratio is 0.9 for almost all the models (which is closer to 1, 1 being ideal Exact Match Ratio). From this we can conclude that all the models are about 90% accurate.

2. K-Means Clustering on a Multi-Class and Multi-Label Data Set
   a. Best k obtained by using silhouette score is 4 with an average score = 0.3787509343305295

   b. The most common Family in Cluster 0 is Leptodactylidae with a count of 3467.
   The most common Family in Cluster 1 is Hylidae with a count of 1245.
   The most common Family in Cluster 2 is Dendrobatidae with a count of 500.
   The most common Family in Cluster 3 is Hylidae with a count of 590.
   The most common Genus in Cluster 0 is Adenomera with a count of 3466.
   The most common Genus in Cluster 1 is Hypsiboas with a count of 1038.
   The most common Genus in Cluster 2 is Ameerega with a count of 500.
   The most common Genus in Cluster 3 is Hypsiboas with a count of 542.
   The most common Species in Cluster 0 is AdenomeraHylaedactylus with a count of 3466.

The most common Species in Cluster 1 is HypsiboasCordobae with a count of 1018.
The most common Species in Cluster 2 is Ameeregatrivittata with a count of 500.
The most common Species in Cluster 3 is HypsiboasCinerascens with a count of 452.

c. Hamming loss: The fraction of the wrong labels to the total number of labels. This is a loss function, so the optimal value is zero.
Hamming score: Also called accuracy in the multi-label setting, is defined as the number of correct labels divided by the union of predicted and true labels.

Hamming Distance for Cluster 0: 0.6761958146487294
Hamming Loss for Cluster 0: 0.02849402092675635
Hamming Score for Cluster 0: 0.4857529895366219

Hamming Distance for Cluster 1: 0.9848637739656912
Hamming Loss for Cluster 1: 0.44483686511940795
Hamming Score for Cluster 1: 0.277581567440296

Hamming Distance for Cluster 2: 0.7032007759456839
Hamming Loss for Cluster 2: 0.5150339476236664
Hamming Score for Cluster 2: 0.24248302618816683

Hamming Distance for Cluster 3: 0.6938110749185668
Hamming Loss for Cluster 3: 0.14006514657980454
Hamming Score for Cluster 3: 0.4299674267100977

3. ISLR 10.7.2 : Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}.$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to

indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

10.7.2

(a) Step 1:- We already have the dissimilarity matrix

$$\begin{pmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{pmatrix}$$

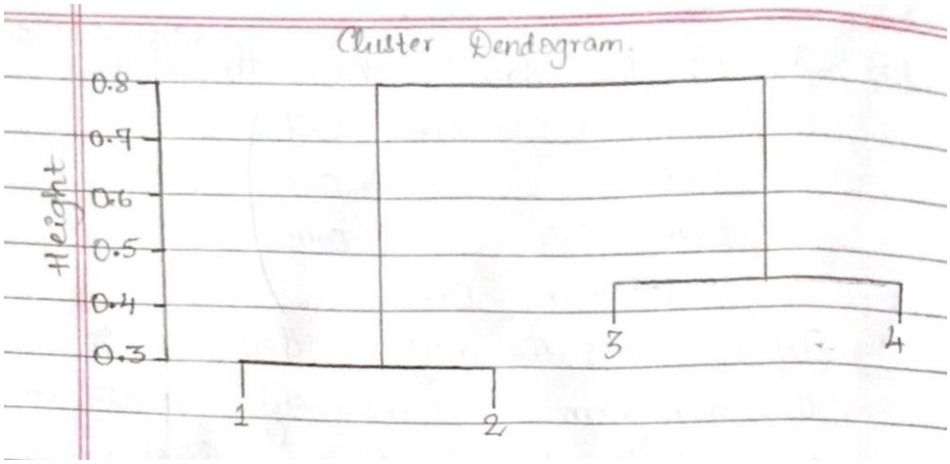
Step 2:- $i=4$: We may see that 0.3 is the minimum dissimilarity, so we fuse observations 1 & 2 to form cluster (1,2) at height 0.3. We now have the new dissimilarity matrix

$$\begin{pmatrix} & 0.5 & 0.8 \\ 0.5 & & 0.45 \\ 0.8 & 0.45 & \end{pmatrix}$$

$i=3$: We now see that the ~~matrix~~ minimum dissimilarity is 0.45, so we fuse observations 3 & 4 to form cluster (3,4) at height 0.45. We now have the new dissimilarity matrix

$$\begin{pmatrix} & 0.8 \\ 0.8 & \end{pmatrix}$$

$i=4$: It remains to fuse clusters (1,2) and (3,4) to form cluster ((1,2),(3,4)) at height 0.8.



(b) Repeat (a), this time using single linkage clustering.

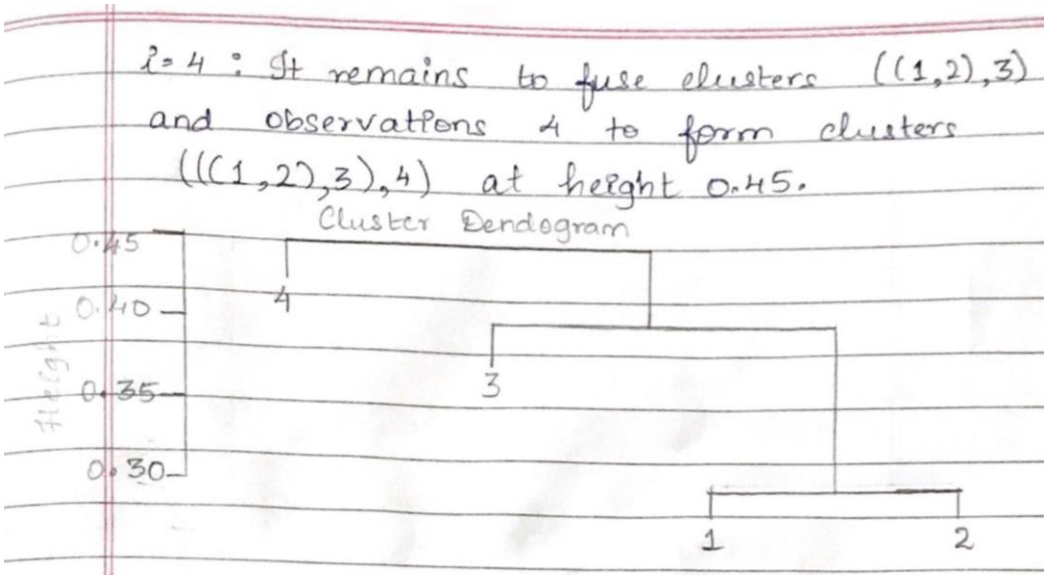
(b) Step 1:- We already have

$$\begin{pmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{pmatrix}$$

Step 2:- $i=4$: We may see that 0.3 is the minimum dissimilarity, so we fuse observations 1 and 2 to form cluster (1,2) at height 0.3. We now have the new dissimilarity matrix :

$$\begin{pmatrix} & 0.4 & 0.7 \\ 0.4 & & 0.45 \\ 0.7 & 0.45 & \end{pmatrix}$$

$i=3$: We now see that the minimum dissimilarity is 0.4, so we fuse cluster (1,2) and observation 3 to form cluster ((1,2),3) at height 0.4, We now have dissimilarity matrix $\begin{pmatrix} & 0.45 \\ 0.45 & \end{pmatrix}$



(c) Suppose that we cut the dendogram obtained in (a) such that two clusters result. Which observations are in each cluster?
Ans: In this case, we have clusters (1,2) and (3,4).

(d) Suppose that we cut the dendogram obtained in (b) such that two clusters result. Which observations are in each cluster?
Ans: In this case, we have clusters ((1,2),3) and (4).

(e) It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.