1. **Multi-class and Multi-Label Classification Using Support Vector Machines**

   (a) Download the Anuran Calls (MFCCs) Data Set from: `https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29`. Choose 70% of the data randomly as the training set.

   (b) Each instance has three labels: Families, Genus, and Species. Each of the labels has multiple classes. We wish to solve a multi-class and multi-label problem. One of the most important approaches to multi-label classification is to train a classifier for each label (binary relevance). We first try this approach:

      i. Research exact match and hamming score/ loss methods for evaluating multi-label classification and use them in evaluating the classifiers in this problem.

      ii. Train a SVM for each of the labels, using Gaussian kernels and one versus all classifiers. Determine the weight of the SVM penalty and the width of the Gaussian Kernel using 10 fold cross validation.[1] You are welcome to try to solve the problem with both standardized [2] and raw attributes and report the results.

      iii. Repeat 1(b)ii with $\mathscr{L}_1$-penalized SVMs.[3] Remember to standardize[4] the attributes. Determine the weight of the SVM penalty using 10 fold cross validation.

      iv. Repeat 1(b)iii by using SMOTE or any other method you know to remedy class imbalance. Report your conclusions about the classifiers you trained.

      v. Extra Practice: Study the Classifier Chain method and apply it to the above problem.

      vi. Extra Practice: Research how confusion matrices, precision, recall, ROC, and AUC are defined for multi-label classification and compute them for the classifiers you trained in above.

2. **K-Means Clustering on a Multi-Class and Multi-Label Data Set**

   Monte-Carlo Simulation: Perform the following procedures 50 times, and report the average and standard deviation of the 50 Hamming Distances that you calculate.

---

[1] How to choose parameter ranges for SVMs? One can use wide ranges for the parameters and a fine grid (e.g. 1000 points) for cross validation; however, this method may be computationally expensive. An alternative way is to train the SVM with very large and very small parameters on the whole training data and find very large and very small parameters for which the training accuracy is not below a threshold (e.g., 70%). Then one can select a fixed number of parameters (e.g., 20) between those points for cross validation. For the penalty parameter, usually one has to consider increments in $\log(\lambda)$. For example, if one found that the accuracy of a support vector machine will not be below 70% for $\lambda = 10^{-3}$ and $\lambda = 10^6$, one has to choose $\log(\lambda) \in \{-3, -2, \ldots, 4, 5, 6\}$. For the Gaussian Kernel parameter, one usually chooses linear increments, e.g. $\sigma \in \{.1, .2, \ldots, 2\}$. When both σ and λ are to be chosen using cross-validation, combinations of very small and very large λ's and σ's that keep the accuracy above a threshold (e.g. 70%) can be used to determine the ranges for σ and λ. Please note that these are very rough rules of thumb, not general procedures.

[2] It seems that the data are already normalized.

[3] The convention is to use $\mathscr{L}_1$ penalty with linear kernel.

[4] It seems that the data are already normalized.

(a) Use k-means clustering on the whole Anuran Calls (MFCCs) Data Set (do not split the data into train and test, as we are not performing supervised learning in this exercise). Choose $k \in \{1, 2, \ldots, 50\}$ automatically based on one of the methods provided in the slides (CH or Gap Statistics or scree plots or Silhouettes) or any other method you know.

(b) In each cluster, determine which family is the majority by reading the true labels. Repeat for genus and species.

(c) Now for each cluster you have a majority label triplet (family, genus, species). Calculate the average Hamming distance, Hamming score, and Hamming loss[5] between the true labels and the labels assigned by clusters.

3. ISLR 12.6.2

4. Extra Practice: The rest of problems in 12.6.

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.

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

(c) Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which observations are in each cluster?

(d) Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster?

(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.

ıre Survey on Algorithms for

ıvıuıu ıavcı ıcaııııığ, ıy ıvıuıamaıau ıuıuwcı.

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.

According to hierarchically clustering algorithm, the dendrogram is:
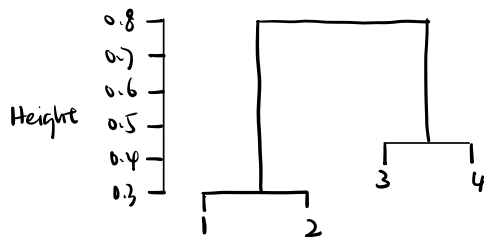
(b) Repeat (a), this time using single linkage clustering.

According to hierarchically clustering algorithm, the dendrogram is:

(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.