Middle East Technical University Department of Computer Engineering

CENG 499

Introduction to Machine Learning

Spring 2018-2019

Homework 2 - Hierarchical Agglomerative Clustering and Decision Trees

version 1

Due date: 21 04 2019, Sunday, 23:59

1 Introduction

In this assignment you have two different tasks. The first one is on Hierarchical Agglomerative Clustering(HAC) in which the datasets are not labeled and you are expected to find the clusters in them using different distance criteria. The second one is on Decision Trees in which you are going to create different trees using different attribute selection strategies and different pruning techniques.

2 Part1: Hiearchical Agglomerative Clustering

In this part, the algorithms will run on synthetically created datasets, in which the data points can be visualized. The datasets are included in the homework files. You should try these criteria on each dataset(these criteria are taken from [1]):

1. The Single-Linkage Criterion:

$$SingleLink(\{\boldsymbol{x}_n\}_{n=1}^N, \{\boldsymbol{y}_m\}_{m=1}^M) = \min_{n,m} ||\boldsymbol{x}_n - \boldsymbol{y}_m||,$$

2. The Complete-Linkage Criterion:

$$CompleteLink(\{\boldsymbol{x}_n\}_{n=1}^N, \{\boldsymbol{y}_m\}_{m=1}^M) = \max_{n,m} ||\boldsymbol{x}_n - \boldsymbol{y}_m||,$$

3. The Average-Linkage Criterion:

$$AverageLink(\{\boldsymbol{x}_n\}_{n=1}^N, \{\boldsymbol{y}_m\}_{m=1}^M) = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} ||\boldsymbol{x}_n - \boldsymbol{y}_m||,$$

4. The Centroid Criterion:

$$Centroid(\{\boldsymbol{x}_n\}_{n=1}^N, \{\boldsymbol{y}_m\}_{m=1}^M) = || \left(\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \right) - \left(\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{y}_m \right) ||,$$

in which the norm is the Euclidean norm.
After using HAC, we get only one cluster that consists of every data point since the algorithm runs util it merges all of them. However, to see the effects of the criteria functions, in this assignment you are expected to stop the merging of the last k clusters where k is the number of clusters which may change data to data. The k value is given to you for each dataset. For each dataset, you are expected to:

- Plot the data points and colorize them according to their clusters to show the results of the algorithms. All the data provided to you are two-dimensional; therefore, it can be plotted easily.
- Try to describe the behaviour of different criteria on each dataset. A good criterion function may differ from dataset to dataset and may change according to our expectations from the task. For each dataset, give reasons for each criterion function why it is or it is not suitable for that dataset.

2.1 Data1

For Data1, stop the algorithm upon reaching 2 clusters. A visualization of it can be seen in 1. The data is normally not labeled; however, to remove ambiguity, it is colorized to depict what our expectations from the data should be.

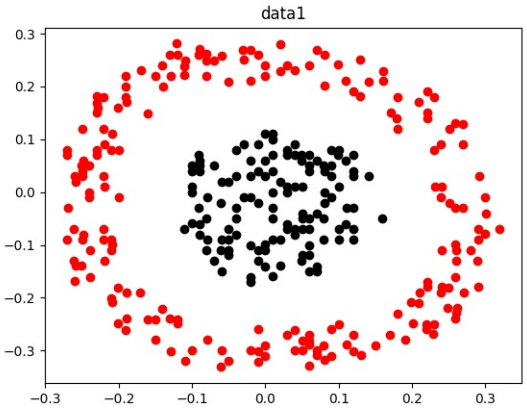

Figure 1: A plot of data1

2.2 Data2

For Data2, stop the algorithm upon reaching 2 clusters. A visualization of it can be seen in 2. The data is normally not labeled; however, to remove ambiguity, it is colorized to depict what our expectations from the data should be.

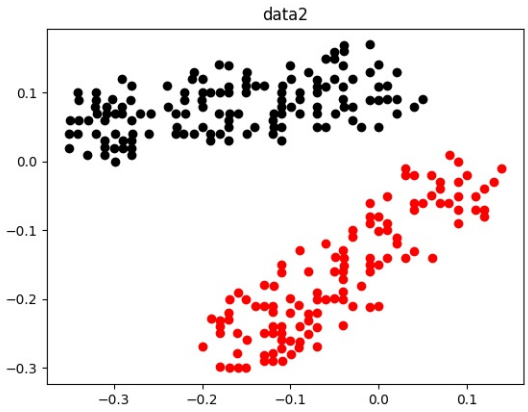

Figure 2: A plot of data2

2.3 Data3

For Data3, stop the algorithm upon reaching 2 clusters. A visualization of it can be seen in 3. The data is normally not labeled; however, to remove ambiguity, it is colorized to depict what our expectations from the data should be.

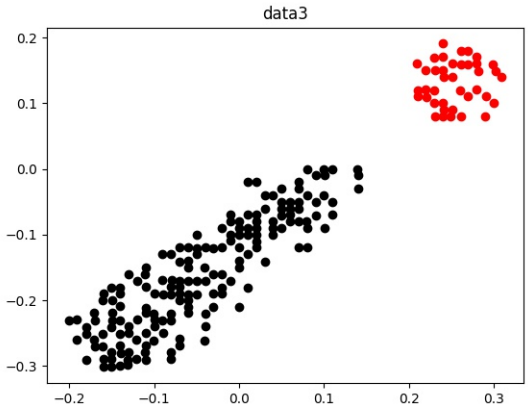

Figure 3: A plot of data3

2.4 Data4

For Data4, stop the algorithm upon reaching 4 clusters. A visualization of it can be seen in 4. The data is normally not labeled; however, to remove ambiguity, it is colorized to depict what our expectations from the data should be.

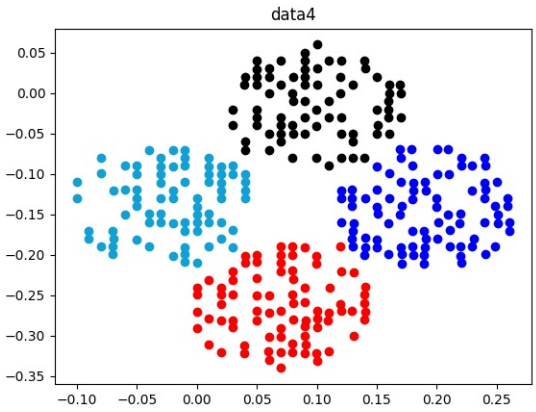

Figure 4: A plot of data4

3 Part2: Decision Tree

In this assignment, you are going to try different attribute selection strategy using id3 algorithm to create decision trees. The dataset is originally taken from `https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`. However, you should use the dataset given in the homework files(it is divided into train and test sets). Its attributes(buying, maint, doors, persons, lug_boot, safety) are categorical, and it has 4 classes(unacc, acc, good, vgood). In the given dataset files, the last value is for the class attribute.
The values that attributes can take are:

doors variables
2,3,4,5 values dedik

- buying: vhigh, high, med, low
- maint: vhigh, high, med, low
- doors: 2, 3, 4, 5more
- persons: 2, 4, more
- lug_boot: small, med, big
- safety: low, med, high

You are expected to do these for all of the subsections below:

- While growing the decision tree, you should use the training set. You should use test set only for final evaluation of the trees.
- After growing the tree, report the test accuracy.
- Draw the tree diagram and put it in the report. In the nodes, proportion of each output class on that current node should be written. If the resultant drawing becomes unreadable, you may put the images of trees in the submission file and reference them in the report. You may use any drawing library, tool, etc. for this purpose.
- Briefly comment on how that specific attribute selection or pruning strategy affected the tree.

3.1 Information Gain

You should select attributes according to the Information Gain. For a set S and an attribute A,

$$InformationGain(S, A) = E(S) - \sum_i \frac{|S_i|}{|S|} E(S_i),$$

where E is the entropy. Be careful on whether you should maximize or minimize the Information Gain.

3.2 Gain Ratio

You should select attributes according to the Gain Ratio. For a set S and an attribute A,

$$IntrinsicInformation(S, A) = -\sum_i \frac{|S_i|}{|S|} log_2 \left(\frac{|S_i|}{|S|} \right),$$
$$GainRatio(S, A) = \frac{InformationGain(S, A)}{IntrinsicInformation(S, A)}.$$

Be careful on whether you should maximize or minimize the Gain Ratio.

3.3 Average Gini Index

You should select attributes according to the Gini Index. For a set S and attribute A,

$$GiniIndex(S) = 1 - \sum_i p_i,$$
$$AvgGiniIndex(S, A) = \sum_i \frac{|S_i|}{|S|} GiniIndex(S_i),$$

where p_i is the proportion of the i^{th} class. Be careful on whether you should maximize or minimize the Average Gini Index.

3.4 Gain Ratio with Chi-squared Pre-pruning

You should stop growing the tree when you cannot reject the null hypothesis which states that there is no association between the two variable(in our case the attribute selected by Gain Ratio and the class variable) using chi-squared test. The selection of the confidence value may change. In this homework, you can use, for example, 90% confidence. The critical values can be obtained from a chi-squared table.

3.5 Gain Ratio with Reduced error post-pruning

After fully growing the tree using Gain Ratio, you should post-prune the tree. For this part, you should split the training set into train and validation sets to calculate the validation set accuracy to decide whether the pruning enhanced the performance. **You cannot use test set for this purpose.** You algorithm can be like this:

1. while a useful replacing can be done:
    1.1. Replace a node with a leaf node(subtree replacement) whenever it increases the accuracy on the validation set and it has no subtree with such property.

4 Specifications

- The codes must be in python. You are not allowed to use any other library other than draing and plotting libraries. Python 3 is preferable but you are allowed to use Python 2 as well.
- Falsifying results, changing the composition of training and test data are strictly forbidden and you will receive 0 if this is the case. Your programs will be examined to see if you have actually reached the results and if it is working correctly.
- You have total of 3 late days for **all** your homeworks. For each day you have submitted late, you will lose 10 points. The homeworks you submit late after your total late days have exceeded 3 will not be graded.
- Using any piece of code that is not your own is strictly forbidden and constitutes as cheating. This includes friends, previous homeworks, or the internet. The violators will be punished according to the department regulations.
- Follow the course page on ODTUCLASS or COW for any updates and clarifications. Please ask your questions on discussion section of ODTUCLASS or COW instead of e-mailing if the question does not contain code or solution.

5 Submission

Submission will be done via ODTUCLASS. If you do not have access to ODTUCLASS, send your homework to this address "artun@ceng.metu.edu.tr" before the deadline. You will submit a zip file called "hw2.zip" that contains all your source code, a README file explaining which file contains what, and your report in a pdf format compiled from the given latex file.

References

[1] R. P. Adams, “Hierarchical agglomerative clustering.”