Term Project Report for STAT745

STAT745 Term Project - Part 2
Unsupervised Dataset: BOSTON-HOUSING

Dec 10, 2010

Abstract In this part we are going to be interested in the unsupervised dataset. The dataset that was given to me was Boston-Housing. The problem with the unsupervised datasets is that there is not supervisor giving us the labels (for classification) or the numeric values (for regression). To derive a meaning out of the data we will use multi-dimensional scaling and rank the data instances to provide an interpretation.

1 Introduction

We are using an unsupervised dataset called Boston-Housing or Housing. Similar to my approach to the previous case, I have firstly manually inspected the data to get an understanding of the features as well as to see if there are any missing values. Boston-Housing data consists of 506 instances, that are defined by 14 attributes.

We will use multi-dimensional scaling (MDS) on the dataset to observe similarities between instances and to provide a ranking of the data. MDS algorithm uses the proximities between the instances and by using these proximities tell us about the similartiy and dissimilarity of instances. For the purpose of finding proximities, R provides a function called *"dist"*, which returns a distance matrix reporting the distances between every pair of instances in the dataset. Furthermore, MDS is a dimensionality reduction technique, that helps us visualize the data in lower dimensions. R language, provides us *"cmdscale"* that implements the dimensionality reduction according to MDS.

In our case, we will represent our dataset in 2 dimensions. In Figure 1 the representation of 506 instances are shown on a 2 dimensional plot. The x and y axis of Figure 1 are the new dimensions. As can be seen, when reduced to a 2 dimensional representation, the instances (a.k.a. houses) show proximity to one another: Notice how the instances from 400 to 500 are aligned on a line, whereas the remaining of the houses align on another line and notice how those two lines (or those two clusters) are separate from one another.

Ekrem Kocaguneli
Lane Department of Computer Science and Electrical Engineering
West Virginia University
Morgantown, WV 26505, USA
E-mail: ekocagun@mix.wvu.edu

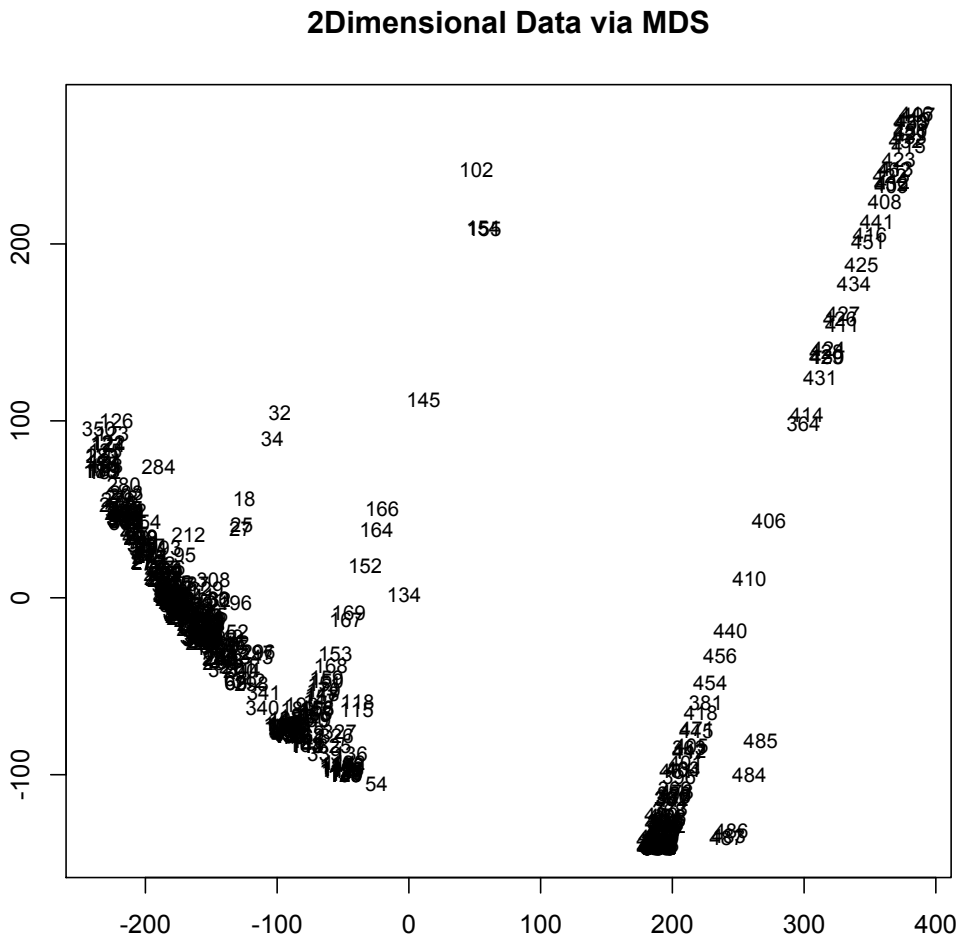

Fig. 1: The 2-dimensional representation of Boston-Housing data. Notice how there are two clusters of houses and each cluster is like a straight line. Also notice that when projected on a 2D plot, it is easier to see the proximity between the houses. The numbers in the plot correspond to instance ID's in the actual dataset.