UNSUPERVISED LEARNING PROJECT PROPOSAL

BRANDON MORETZ
SEAN PRENTIS

INTRODUCTION

There are many times when we look to a dataset to answer specific, narrowly targeted, questions we have concering some variable of interest. For example, to accurately forecast the value of a home, we must find a relevant dataset that contains accurate information of comparable inventory so that we can explore the significant variables of a home which ultimately determine the sale price of the residence which would be our targeted variable of interest. However, in these situations we can be too quick to jump to a specific conclusion about what information is contained in a dataset by not fully exploring the dataset to its fullest without algorithmic aid.

The key task in this assignment is to use unsupervised learning methods to assist our research methods by looking for the presence of latent factors or hidden market segment clusters in the underlying data that we might not find ourselves with more traditional methods of exploratory data analysis. In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis, variable derivation and selection to ultimately produce a set of market segments we could use for a targeted advertising campaign based upon the results of our analysis.

DATA SURVEY

The Ames housing dataset contains approximately three-thousand observations of eighty-two variables collected from the Ames Assessor's Office specifically assess the value of individual residential properties sold in Ames, Iowa from 2006 to 2010. This data was originally collected to assess home values in an observational study; however, we can repurpose it here and leverage many of the home value indicators as socioeconomic variables that could help us partition these residences into similar categories for our marketing campaign.

The main task we aim to achieve with this analysis is the in-depth exploration of the interdependencies and relationships in the variables under study, and how the attributes of a residence are interrelated such they form similar segments.

METHODOLOGY

In this analysis, we will use all the information presented in the dataset and then weed out anything that is likely a non-contributing factor in our analysis. We will take a breath over depth approach, at least initially, using traditional exploratory data analysis along with dimensionality reduction techniques to construct the core set of variables to be used throughout this analysis. Once we have obtained a concise set of variables for analysis, we will attempt to uncover the underlying structure of these homes, or how they form clusters which can potentially used to perform market segmentation analysis.

RESEARCH QUESTIONS

1.) What are the characteristics of these homes that have the most distinguishing characteristics?

2.) How can we form a concise set of descriptors that accurately reflects the variation in the homes, minimizing the number of individual variables?