NML502 - Final Project

Elliot Smith, Eugen Hruska, Varun Suriyanarayana

4/22/2018

Introduction

In this report, we will be comparing the clustering results of a dataset of different beer types using two distinct methods, K Means and Self-Organizing Maps. Our previous experience tells us that when data reaches a particular level of complexity that K Means is no longer an acceptable method for clustering; we hope to prove or debunk this assertion. In addition, given our hypothesis is correct, that K Means is not a satisfactory method, we hope to prove that using an SOM network is a significantly stronger technique for identifying clusters in multi-dimensional data. We hope that you will learn something through this process as we have.

Motivation

There were many motivations for us to perform this analysis:

- We love beer! While trying to decide on a project topic and looking for similar interests, we found that each of us enjoyed beer and that trying to cluster multi-dimensional beer data would be both interesting and fun; it certainly made working on the project very engaging as we were curious to test the results of our analysis against our previous experience.
- Our group had a strong preference to explore Unsupervies Learning (as opposed to a Supervised approach). As a team, we were certainly more interested in the concept of trying to find the solution without providing the answer; the potential power of accurate prediction without supervision was very interesting to us. However, having the answers available to us was also important, so that we could test how well our analyses performed.
- Self-Organizing Maps is a topic that deserved further exploration. We decided that an SOM was a robust tool that was perhaps best equipped to solve our problem with an optimal solution and we also had an intrinsic desire to explore the network's abilities more in a multi-dimensional setting; this project represented the perfect opportunity.
- Each of us had the desire to pursue the topic of classification; it worked to our benefit that the data we chose was very well-suited to our task. The concept that we can try to determine a observations class (in this case, a style of beer) by developing a learning algorithm was a major driver for us. Classification power is an extremely robust topic; the ability to classify observations based on their features is a topic worth exploring!
- Finally, we wanted to compare how Self-Organizing Maps compared against a more conventional clustering method (in this case, K Means). Since SOMs are a relatively new topic to us, we wanted to understand how it would compare to methods that we had more familiarity with in an effort to expand our statistical toolsets and perhaps replace our primitive tools (K Means) with a more robust one (SOMs).

Objectives

Our Data

Beer Data

Data Transformation

Our Analysis

K Means

Self-Organizing Map

The Comparison

Conclusion