Machine Learning - CS 7641 Assignment 3

Anthony Menninger
Georgia Tech OMSCS Program
amenninger3
tmenninger@gatech.edu

Abstract

This paper explores several unsupervised learning algorithms. It uses two datasets: Census Data labeled to income above or below $50,000 and MNIST handwritten digit image data for digits 0-9. Clustering is performed on the original datasets, then several dimensionality techniques are used to reduced the feature set and then clustering is preformed again. Finally, the reduced datasets are tested using learning algorithms.

Introduction

The first data set is from 1994 and 1995 current population surveys conducted by the U.S. Census Bureau [2] from the UC Irvine Machine learning site. The classification task is to determine if the user type for each instance has an income above $50,000 or below ($94,000 in today's dollars). There are 40 usable features, with both continuous features, such as age, and discrete, such as married status or sex. The training set has 199,523 instances and the test set has 99,762 instances. Because many of the features in the file are string based, I created transformed instances using the sklearn preprocessing LabelEncoder to translate discrete string values into consistent numeric values. This was necessary for algorithms that only used numeric features. I chose this data set because it the training dataset seemed relatively large and I expect somewhat noisy with a smallish number of features.

The second data set is The MNIST Database of Handwritten Digits [2]. This consists of instances of 28 X 28 greyscale images of the digits 0-9. One transformation performed was that the values where scaled to 0 to 1, from 0 to 255. The images were flattened, creating 784 features, one for each pixel. There are 60,000 training instances and 10,000 test instances. I chose this because it is a good comparison to the first data set, with a very different type of data (images), which leads to significantly more features (784). In addition, each of the features can be thought of as related to each other, as they are all positional in a two dimensional grid, while the first data set features do not have any necessary relation between themselves ie: age is not related to sex.

For this assignment, the core python library used was sklearn [3], which has many unsupervised learning algorithms. In addition, an enhancement Yellowbrick [4] was also used for some visualizations.

Clustering

K Means

K Means creates K clusters based on the distance of instances to cluster centers. A key question is what is the optimal number of clusters. An Elbow Plot looks at the intra distance within clusters at different K, and then picks the K where intra-distance change slows down, or the "elbow" of the plot. A silhouette plot charts the intra cluster distance and references it vs the next nearest cluster intra cluster distance. This produces a chart that can provide visual insight. Both methods were used in choosing an optimal K.

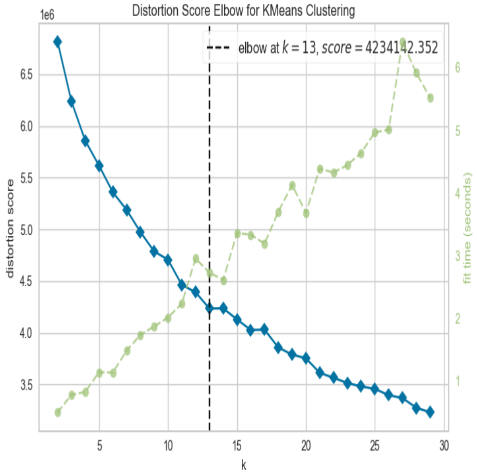

Figure 1: K Means Census Elbow: .

Figure 1 shows the elbow chart score

Figure 2 shows the silhoette score

Figure 2 shows the silhoette score **Figure 3** shows the silhoette score

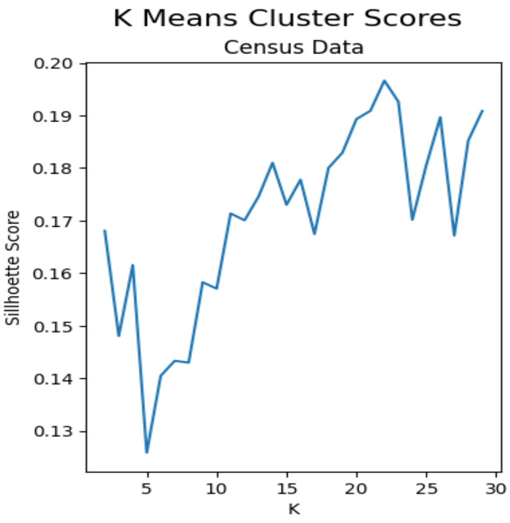

Figure 2: K Means Census Data Silhouette Scores: .

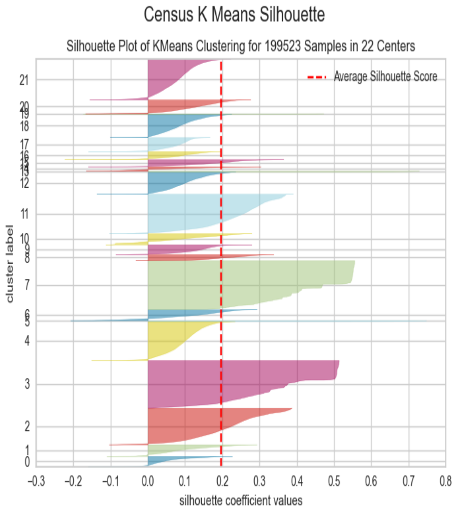

Figure 3: K Means Census Best Silhouette: .

References

1 The MNIST Database of Handwritten Digits. url: http://yann.lecun.com/exdb/mnist/.

2 Census Income (KDD). url: https://archive-beta.ics.uci.edu/ml/datasets/census+income+kdd.

3 Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

4 Bengfort et al., (2019). Yellowbrick: Visualizing the Scikit-Learn Model Selection Process. Journal of Open Source Software, 4(35), 1075, https://doi.org/10.21105/joss.01075.

4 Menninger, A. (2022) Code created for this experiment https://github.com/BigTMiami/ML_Assign_2 Created: 2022.