Bronco Learn

A K-means Algorithm
For a 13 Dimension Dataset

Erali Shah

Rohan Oswal

Brett Castello

Introduction

Algorithm
High Level Design
Low Level Design
Code
Optimization

Problem / Solution

Problem

- Dataset of 178 rows
- Thirteen different dimensions
- Skewed with outliers
- No labels or outputs

Solution

- Isolate clusters
- Unsupervised Algorithm
- Number of clusters
- Optimize / Measure Performance

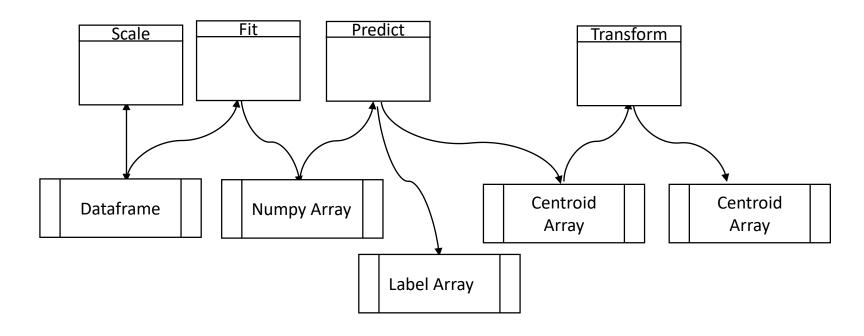
Algorithm

Unsupervised Learning

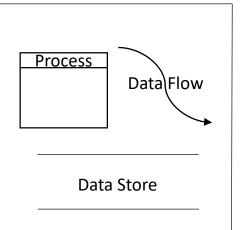
K "means"

• Five different distance formulas

Design – High Level



Legend



Design – Low Level Fit MinMax Set Rand. Set Convert 2 Load DS **Iterate/** Scaler NP Array Center Inertia **Update** Dataframe Numpy Array Centroid Array Inertia Array Numpy Array Dataset **Predict** Calculate Legend Distances Process Data\Flow **Label Distance Array**

Data Store

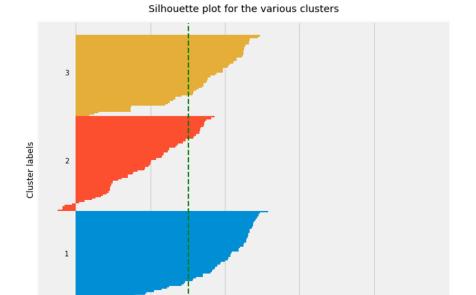
Optimization

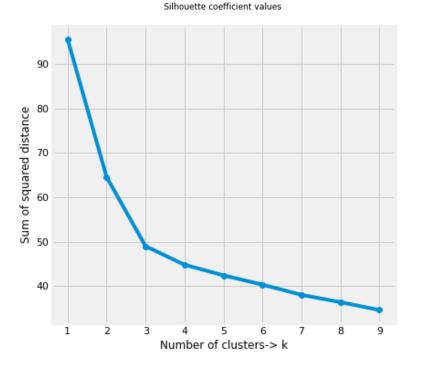
Four different metrics

- Inertia score
- Silhouette score
- Calinski harabaz index
- Davies bouldin score

Best: k=3

Best Distance Metric: Euclidean





Questions?