Clustering Education Data

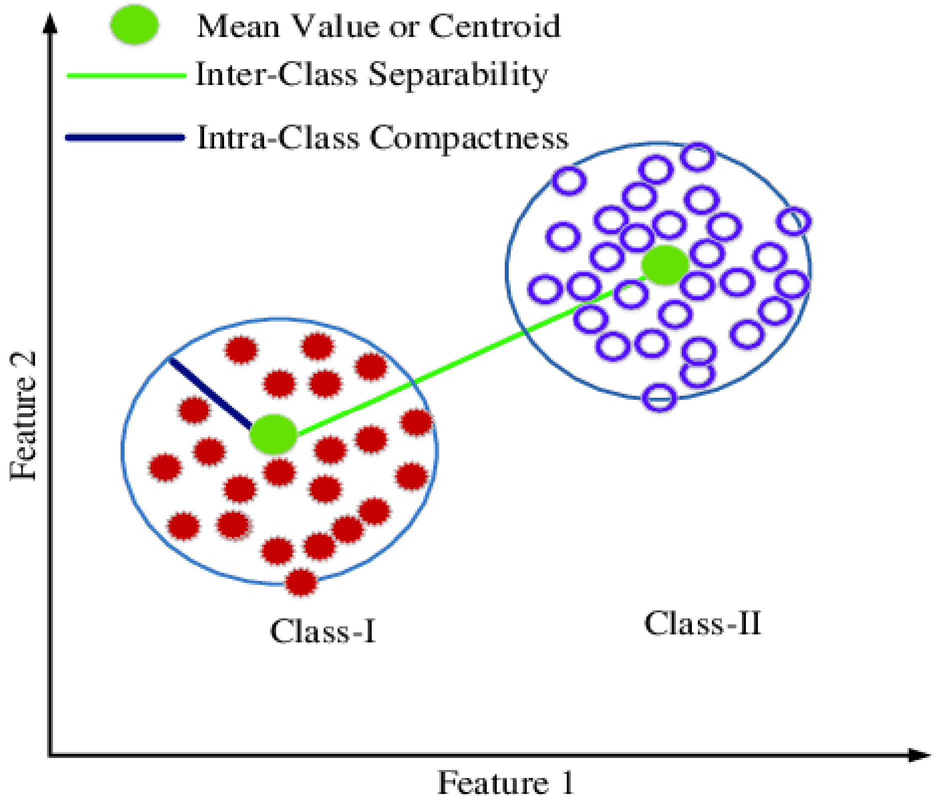
- Yale Quan
  - yalequan@uw.edu
  - M&S Seminar
  - December 3, 2020

Agenda

- Introduction to K-Medoids and PAM
- Determining the number of clusters
- Gower’s Distance
- Potential Problems with Gower’s Distance
- The PAM Algorithm
- Cluster Validation
- Excerpt on LDA

Introduction to K-Medoids and Partitioning Around the Medoid (PAM)

- The objective of cluster analysis is to create two or more partitions of the data such that
  - objects within a cluster are similar (**intra-class compactness**)
  - objects in different clusters are dissimilar when compared to any other cluster (**inter-class separability**).



Introduction to K-Medoids and Partitioning Around the Medoid (PAM)

- The goal of the PAM algorithm is to find the best observations called **medoids** that are centrally located in clusters such that the total distance between observations within a cluster and the medoid is minimized.
  - One of the main differences between PAM and other common K-clustering techniques such as K-Nearest-Neighbors (KNN) is that PAM uses an observation as the medoid or center of a cluster, while KNN uses a point in $\boldsymbol{R}^2$, commonly denoted as μ_i**,** which can be estimated using Lloyd's Algorithm (Lloyd, 1982)

Determining the number of clusters to use for PAM

- Kaufman and Rosseeuw (2009) defined a value called *silhouettes* that are calculated to determine the optimal number of clusters for the PAM algorithm. The silhouette values are calculated as follows:

1. Select the K many clusters you want to test. Conventionally this is chosen to be 12, but due to the increasing processing power higher values are possible.
2. Let i represent an object that is in cluster A. A value a is computed which is the average dissimilarity of i to all other object inside cluster A. If cluster A only contains i, then $a = 0$.
3. For every other cluster that is not equal to cluster A, compute the average dissimilarity for objects in that cluster to i. Find the cluster, B, that has the smallest average dissimilarity measurement, b, to i.

Determining the number of clusters to use for PAM

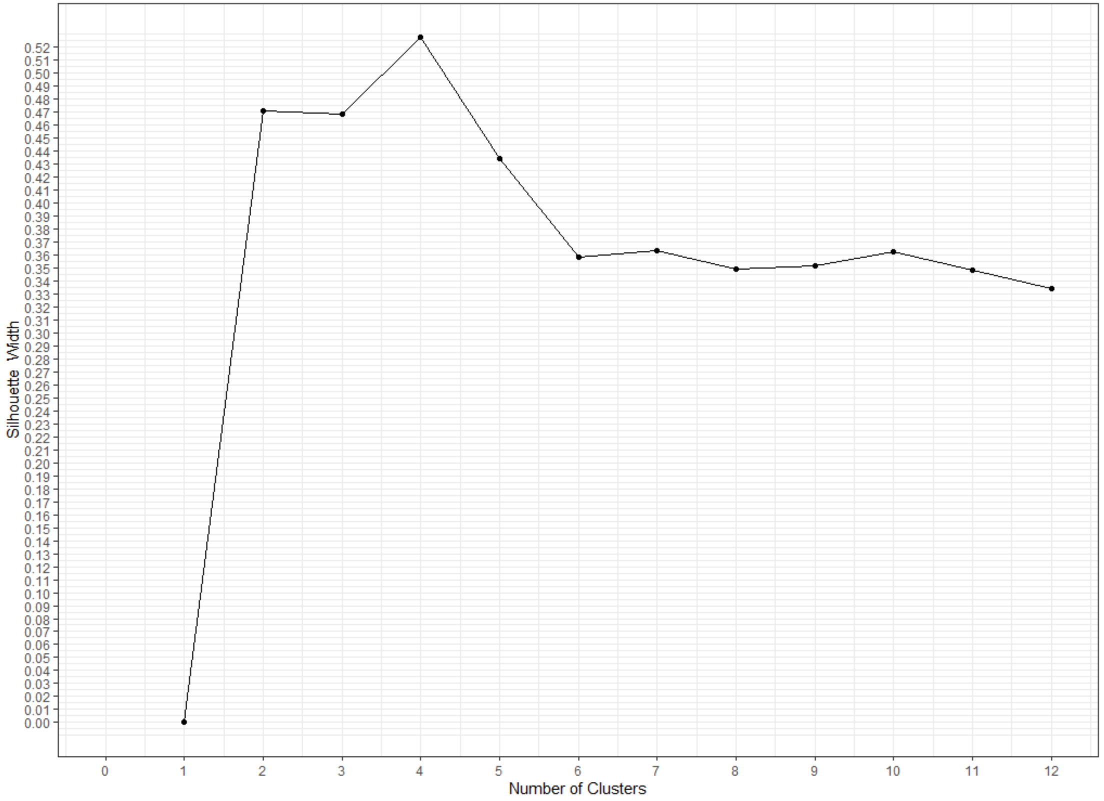
Having clusters A and B and their respective average dissimilarity measurements a and b, the silhouette value, s, is now calculated for i.

$$s_i \begin{cases} 1 - \dfrac{b}{a}, & a > b \\ 0, & a = b \\ \dfrac{a}{b} - 1, & a < b \end{cases}$$

This process is repeated for all items within a cluster.

Cluster Silhouettes

- After the silhouette values are calculated for each cluster, the average silhouette value is calculated across clusters. This is called the **cluster silhouette**
  - At the completion of the algorithm there will be K many cluster silhouettes
- Kaufman and Ross suggest that the optimal number of clusters, K, is when K maximizes the Cluster Silhouette value.
  - This is commonly determined using a **cluster silhouette graph**



Gower’s Distance

- One of the challenges with clustering education data is that education data commonly consists of mixed data types.
- This presents a challenge when using traditional linear and non-linear clustering methods like K-Means which traditionally relies on either Euclidean Distance, Manhattan Distance, or Chebychev Distance all of which rely on numeric data.
- Gower (1971) proposed a general (dis)similarity method for clustering mixed data and has been shown to be effective at clustering education data.

Gower's Distance

- $GD_{ij} = \frac{\sum_{Z=1}^{N} s_{ijz}\, \delta_{ijZ}}{\sum_{Z=1}^{N} \delta_{ijz}}$
  - If both x_{iZ} and x_{jZ} exist then $\delta_{ijZ} = 1$, else δ_{ijZ} = 0.
  - If Z is a continuous random variable the value of $s_{ijz} = 1 - \frac{|x_i - x_j|}{R_x}, \forall\, x_i \neq x_j$ where R_x is the range of the variable Z. When $x_i = x_j$ then $s_{ijz} = 1$, and when x_i and x_j are at opposite ends of the range s_{ijz} is minimized with respect to Z.

Gower's Distance

- If Z is a binary random variable with 1 indicating membership in the group and x_{iZ} and x_{jZ} exist, then **Table 2** is used to calculated the value of s_{ijz} and δ_{ij}. If either x_{iZ} or x_{jZ} do not exist, then $s_{ijz} = 0$.

TABLE 2

SCORES AND VALIDITY OF DICHOTOMOUS CHARACTER COMPARISONS

	Values of character k			
Individual i	+	+	−	−
j	+	−	+	−
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

- If Z is a nominal random variable and x_{iZ} and x_{jZ} exist and are equal, then $s_{ijz} = 1$. If either x_{iZ} or x_{jZ} do not exist, or x_{iZ} and x_{jZ} are not equal then $s_{ijz} = 0$.

Gower’s Distance - Potential Problems

- When using Gower’s Distance, I noticed a potential problem when clustering data that included nominal data:
  - If Z is a nominal random variable and x_{iZ} and x_{jZ} exist and are equal, then $s_{ijz} = 1$. If either x_{iZ} or x_{jZ} do not exist, or x_{iZ} and x_{jZ} are not equal then $s_{ijz} = 0$.
- Gower’s Distance only checks for matching nominal data. What happens if there is potential overlap in categories?
- Gower’s Distance does not include a “partial distance
  - Students changing majors from Applied Math into Pure Math
- Gower’s Distance does not include “long distance”
  - Students changing majors from Biology into English

The Kaufman and Rosseeuw (2009) Partitioning Around the Medoid (PAM) algorithm

The goal of the PAM algorithm is to minimize the average dissimilarity of objects within a cluster to their medoid (center)

Phase 1

1. K many objects are randomly selected, to serve as the medoid of each of the K clusters.
2. An additional $K - 1$ objects are incrementally placed into a cluster to minimize the distance between objects. **(This can be very computationally heavy for large datasets)**
3. The total distance between objects, D_k, is stored for each cluster

Phase 2

1. Objects within a cluster are rotated to become the medoid and a new distance measurement is calculated, $D_{k'}$, and compared with D_k
2. If $D_{k'} < D_k$ the new object becomes the medoid
3. Repeat for all cluster
4. If at least one medoid changed, begin Phase 1 again with the new medoids

This process is repeated until all potential medoids are considered and the objective function for each medoid is reduced as far as possible.

Cluster Validation

1. Objects within a cluster are similar (**intra-class compactness**)
   - Calculate the variance between items within a cluster

1. Objects in different clusters are dissimilar when compared to any other cluster (**inter-class separability**).
   - Calculate the average distance between clusters
   - Calculate the variance between clusters

2. If possible, plot the clusters and visually inspect them
   - Usually not possible with high-dimensional clustering
   - You can project them down into R^2 but it's challenging to

Classifying Graduation Times At CSULB

- I worked on a classification project at CSULB where I worked on developing classification algorithms for predicting graduation times. This ended up becoming my MS Thesis.
- My goal was to develop a classification algorithm that could be updated with new information each semester and would provide timely and meaningful information to Academic Advising on predicted graduation
- My final models presented were Multinomial Logistic Model and Fishers Multipopulation LDA. Here’s an excerpt on my results for LDA.

Classifying Observations

Goal: Derive a classification algorithm that can be used to classify new observations into “Did Not Graduate”, “Four-Year Graduate”, or “Six-Year Graduate”.

Methodology: Fishers Linear Multi-Population Discriminant Analysis

- Assumptions of Fishers multi-population LDA
  1. Covariance matrices for each population are approximately equal and full rank.
  2. The populations are Multivariate Normally distributed.

- See Gilbert (1968), Marr and Hume (1996) , and Tripp and Duffey (1981) for examples of applying Fishers LDA without meeting these assumptions.

Classifying Observations

Goal: Derive a classification algorithm that can be used to classify new observations into “Did Not Graduate”, “Four-Year Graduate”, or “Six-Year Graduate”.

Methodology: Fishers Linear Multi-Population Discriminant Analysis

Without any prior knowledge about the distribution of graduation rates for first-time freshman at CSULB, the prior probabilities were set to the proportion of each population:

π_1 = **“Did Not Graduate”,**	π_2 = **“Four-Year Graduate”**	π_3 = **“Six-Year Graduate”**
$p_{\pi_1} = 0.35357$	$p_{\pi_2} = 0.1611$	$p_{\pi_3} = 0.4854$
$n_{\pi_1} = 2{,}406$	$n_{\pi_2} = 1{,}096$	$n_{\pi_3} = 3{,}303$

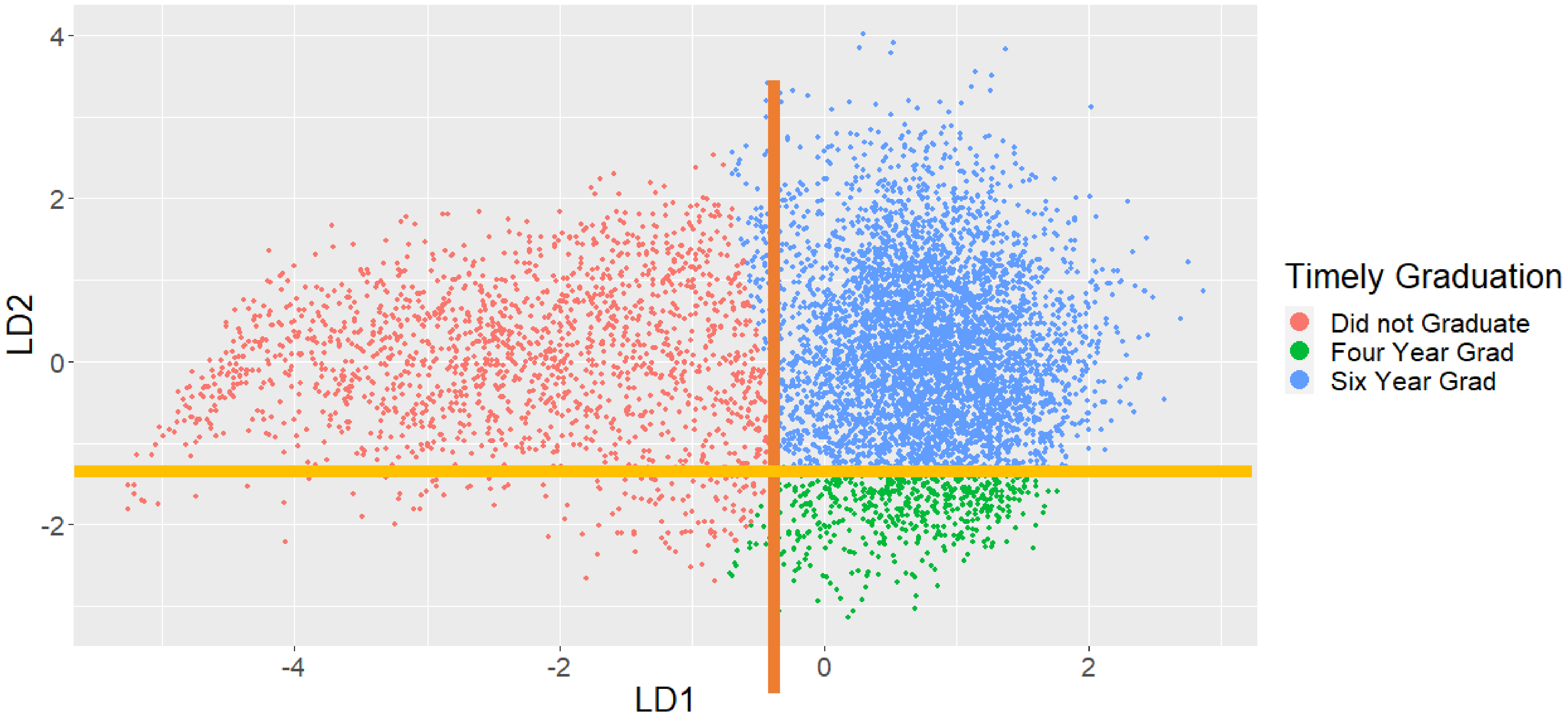
Note: I assumed equal misclassification costs.

Classifying Observations: Fishers Linear Discriminant Analysis (LDA)

Coefficients of Linear Discriminants

Variable Name	Coefficient LD1	Coefficient LD2
STEM Admission	0.0533	0.4355
Student Gender	−0.1157	−0.2910
Undeclared Admission	0.1355	0.5943
Pell Eligibility	0.0662	0.2684
Local Admission	0.1112	0.2374
Minority Admission	0.1722	−0.0856
First Generation Student	0.0130	0.0835
Number of Major Changes	0.2955	0.2159
Factor 1 - Academic Preparation	0.2198	−0.2859
Factor 2 - CSULB Academic Information	−0.7131	0.5219
Factor 3 - Credits Taken	1.0944	0.2711
Factor 4 - Math Standardized Test Score	0.0874	−0.2126
Factor 5 - Reading Standardized Test Score	0.1205	−0.5139

Classifying Observations:
Fishers Linear Discriminant Analysis (LDA) – Decision Boundaries



Classifying New Observations:
Fishers Linear Discriminant Analysis (LDA)

Categorical Variables	Continuous and Discrete Variables
STEM Admission	0
Student Gender	0
Pell Eligibility	1
Local Admission	0
Minority Admission	0
First Generation Student	0
Factor 1 - Academic Preparation	0.3834
Factor 2 - CSULB Academic Information	-0.1500
Factor 3 - Credits Taken	0.0178
Factor 4 - Math Standardized Test Score	1.1365
Factor 5 - Reading Standardized Test Score	1.9898
Number of Major Changes	2
True Classification	Six-Year Graduate

LD1	LD2
1.2071	1.5173

Classifying New Observations:
Fishers Linear Discriminant Analysis (LDA)