CS/ENGR M148 L12:
Clustering

Sandra Batista

Administrative News

This week in discussion section:

No lab this week.

Project check-in this week on unsupervised learning.

Midterm grading underway. We hope to be done by Wednesday next week. ~98% done.

We'll be sharing a mid-quarter survey

Sorry, we did not offer extra credit, so that we could preserve anonymity

PS3 data posted.

Join our slido for the week…

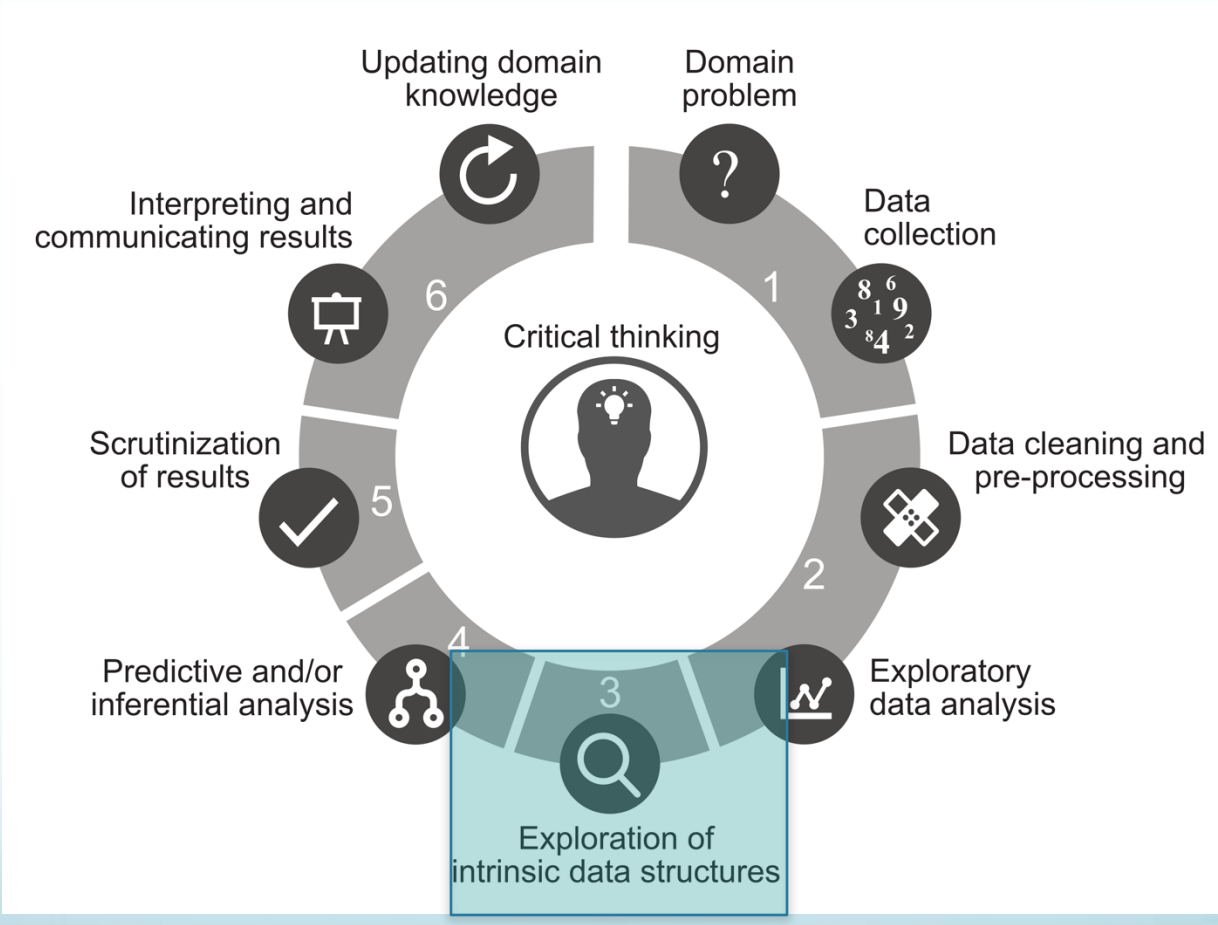
https://app.sli.do/event/nCV57u4mC7eUMit9euSBr2

Today's Learning Objectives

Students will be able to:

- Review: Using SVD to perform PCA
- Understand what agglomerative clustering is
- Understand what agglomerative clustering is
- Evaluate clustering using quantitative metrics
- Apply clustering to a real LPGA data set

Data Science Life Cycle (DSLC)



[Yu, Barter 2024]

Dimensionality reduction

High-dimensional data is data with many features (such as thousands of variables) e.g. gene expression data, nutrition data

Dimensionality Reduction is the process of creating a lower-dimensional representation of a dataset.

1. Summarize the strongest patterns and relationships between the *variables* in the data (
2. Make computation on the data easier by reducing its size.

E.g.: Principal component analysis summarizes low-dimensional linear relationships in high-dimensional datasets

[Yu, Barter 2024]

Principal Component Analysis (PCA)

Principal component analysis is an algorithm that computes a series of "orthogonal" linear combinations that have the maximum possible variance relative to the origin

By default, the origin is the data point whose measurements across all variables equal zero

If data has been mean-centered, the origin is the mean of all the measurements

[Yu, Barter 2024]

Singular value decomposition

- **Input**: an *m x n* matrix
- **Output**: a set of numbers called *singular values* and two collections of vectors: a set of *right singular vectors* and another set of *left singular vectors*.

$$X = UDV^T.$$

Singular value decomposition

$$X = UDV^T.$$

The left matrix, U, contains the ***left-singular vectors***

The matrix D is a **diagonal matrix** whose entries correspond to the magnitude or strength of the corresponding principal component directions. **Singular values are diagonal entries of D.**

V contains the **right-singular vectors** of the data matrix and each right-singular vector corresponds to a principal component.

Each column of V contains the coefficients of the corresponding principal component linear combination.

Variable Loadings

Variable loadings on corresponds to the coefficient (or weight) of the variable in the linear combination that defines the principal component.

Variable Loadings can be used for feature importance if variables on the same scale.

The variable loadings for each PC are extracted from the relevant column of the right-singular vector matrix, V.

[Yu, Barter 2024]

Variable Loadings

$$V = \begin{array}{r} \\ \text{(Sodium)} \\ \text{(Potassium)} \\ \text{(Calcium)} \\ \text{(Phosphorus)} \\ \text{(Magnesium)} \\ \text{(Total choline)} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{(PC1)} & \text{(PC2)} & \text{(PC3)} & \text{(PC4)} & \text{(PC5)} & \text{(PC6)} \end{array} \\ \begin{bmatrix} 0.42 & 0.63 & 0.23 & -0.58 & 0.15 & -0.1 \\ 0.48 & -0.28 & -0.43 & -0.3 & -0.64 & 0.1 \\ 0.29 & -0.24 & 0.79 & 0.22 & -0.35 & -0.27 \\ 0.49 & -0.06 & 0.12 & 0.28 & 0.3 & 0.76 \\ 0.38 & -0.51 & -0.16 & -0.11 & 0.59 & -0.46 \\ 0.35 & 0.45 & -0.33 & 0.66 & -0.08 & -0.35 \end{bmatrix} \end{array}.$$

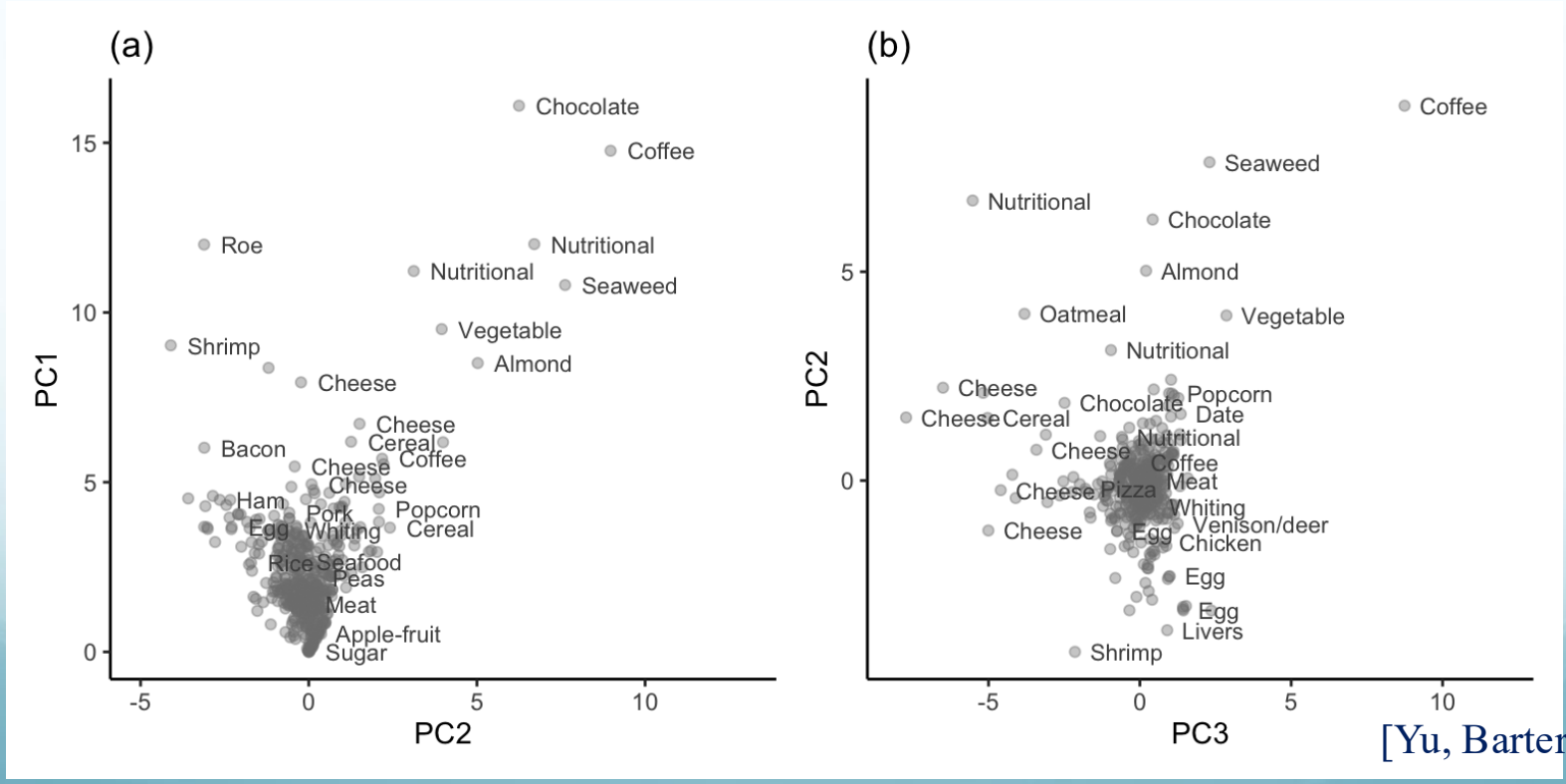
$$\begin{aligned} \text{PC1} = {} & 0.42\,\text{sodium} + 0.48\,\text{potassium} + 0.29\,\text{calcium} \\ & + 0.49\,\text{phosphorus} + 0.38\,\text{magnesium} \\ & + 0.35\,\text{total choline}. \end{aligned}$$

[Yu, Barter 2024]

PC Data set

To calculate the PC data set, multiply the original data set and the ***right-singular vectors, V***

$$X^{PC} = XV.$$



[Yu, Barter 2024]

Variance Explained

*D contains the **singular values** on its diagonal*

The magnitude of the singular values measures the variability in the original data that is being captured by each principal component.

$$D = \begin{bmatrix} 252.84 & 0 & 0 & 0 & 0 & 0 \\ 0 & 97.36 & 0 & 0 & 0 & 0 \\ 0 & 0 & 91.13 & 0 & 0 & 0 \\ 0 & 0 & 0 & 80.76 & 0 & 0 \\ 0 & 0 & 0 & 0 & 62.3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 45.01 \end{bmatrix}.$$

[Yu, Barter 2024]

Variance Explained

The sum of the variance of the columns in the original dataset equals the sum of the squared singular values:

$$\sum_j Var(X_j) = \sum_j d_j^2,$$

$$\text{prop of variability explained by PC}_j = \frac{d_j^2}{\sum_i d_i^2}.$$

[Yu, Barter 2024]

Numpy SVD

```
pca_U, pca_d, pca_V =
np.linalg.svd(food_fndds_scaled)
```


[Yu, Barter 2024]

Today's Learning Objectives

Students will be able to:

- ✓ Review: Using SVD to perform PCA
- Understand what agglomerative clustering is
- Understand what agglomerative clustering is
- Evaluate clustering using quantitative metrics
- Apply clustering to a real LPGA data set

What is unsupervised learning?

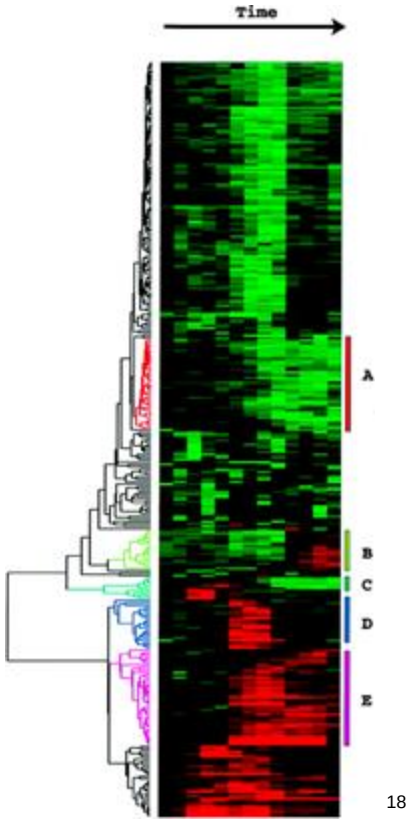
- Organize or describe data when no labels are available.
- A major area: clustering
- Organizing data into collections of points that are "close" is known as **clustering**.
- PCA is another example of unsupervised learning

[Shah 2020]

Hierarchical clustering

- Agglomerative clustering starts with many individual clusters and merges closest pairs of clusters until only one cluster remains (bottom up)
- Uses a distance metric such as Euclidean or correlation
- Many linkage approaches:

  single – compute distances for the most similar



[Pachter 2021]

Hierarchical clustering

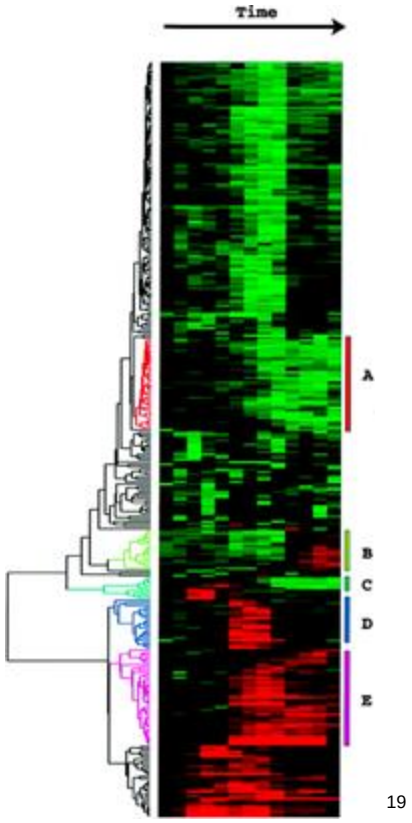
- Many linkage approaches:

  single – compute distances for the most similar elements

  complete – compute distances for most dissimilar elements

  average – computes average distances

  Ward – applies total within cluster sum of squares



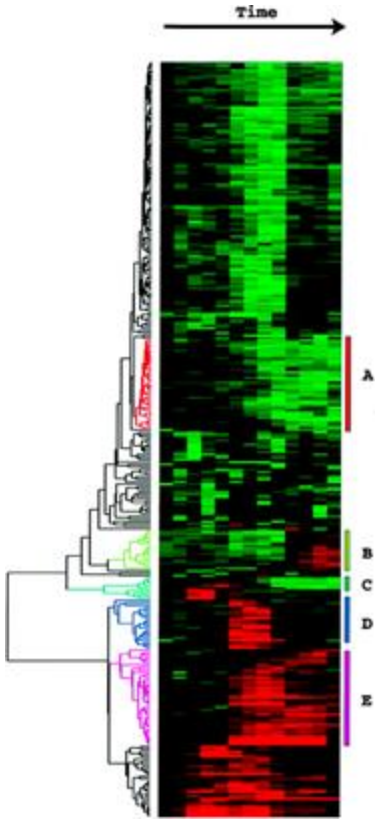
[Pachter 2021]

Hierarchical clustering

Iterative algorithm

1. Compute pairwise distance
2. Represent each point as cluster
3. Merge two closest clusters based on distance and linkage
4. Update cluster linkage matrix
5. Repeat steps 2-4 until one cluster remains

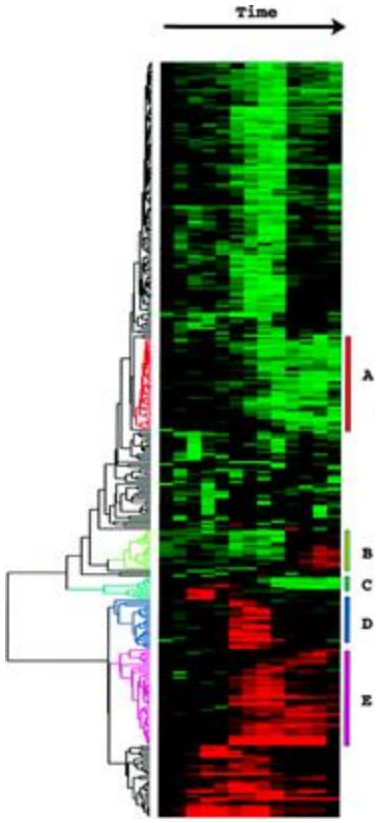
Sebastian Raschka, Yuxi (Hayden) Liu, and Vahid Mirjalili. *Machine Learning with PyTorch and Scikit-Learn*. Packt Publishing, 2022, pgs 320-321



[Pachter 2021]

Origins of hierarchical clustering for single-cell RNA-seq

- Correlation is frequently used to identify groups of genes that interact with each other, and to construct gene regulatory networks.
- (Eisen et al., 1998) used correlations of gene expression measurements as a measure of interaction and applied this to constructing **(hierarchical) clusters** of genes in a time-course of serum stimulation of primary human fibroblasts:
  - A: cholesterol biosynthesis
  - B: the cell cycle
  - C: the immediate-early response
  - D: signaling and angiogenesis
  - E: wound healing and tissue remodeling



[Pachter 2021]

Hierarchical clustering Challenges

Parameter explosion: The coupling of clustering to dimension reduction, filtering of genes, and other data processing, results in an explosion of parameters and complex heuristics.

Computational challenges: Clustering algorithms do not typically scale well since the computation of pairwise distances between cells is quadratic in the number of cells.

[Pachter 2021]

Agglomerative clustering

1. Use any computable cluster similarity measure $sim(C_i, C_j)$ e.g., Euclidean distance,
2. For n objects $v_1, \ldots, v_n$, assign each to a singleton cluster $C_i = \{v_i\}$
3. Repeat {

   - identify two most similar clusters C_j and C_k (could be ties-chose one pair)

   - delete C_j and C_k and add $(C_j \cup C_k)$ to the set of clusters.

   } until just one cluster.
4. Dendrograms diagram the sequence of cluster merges.

[Shah 2020]

Example tracing agglomerative clustering

	1	2	3	4	5
1	0				
2	8	0			
3	3	6	0		
4	5	5	8	0	
5	13	10	2	7	0

[Shah 2020]

Example tracing agglomerative clustering

[Shah 2020]

Example tracing agglomerative clustering

[Shah 2020]

Example tracing agglomerative clustering

[Shah 2020]

Drawing the dendrogram

[Shah 2020]

Today's Learning Objectives

Students will be able to:

- ✓ Review: Using SVD to perform PCA
- ✓ Understand what agglomerative clustering is
- ✓ Understand what agglomerative clustering is
- Evaluate clustering using quantitative metrics
- Apply clustering to a real LPGA data set

Divisive clustering

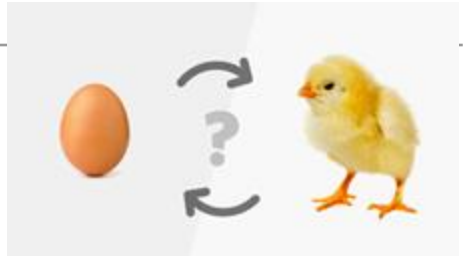
1. Put all objects in one cluster
2. Repeat until all clusters are singletons {

- choose a cluster to split based on some criterion.
- replace the chosen cluster with sub-clusters.

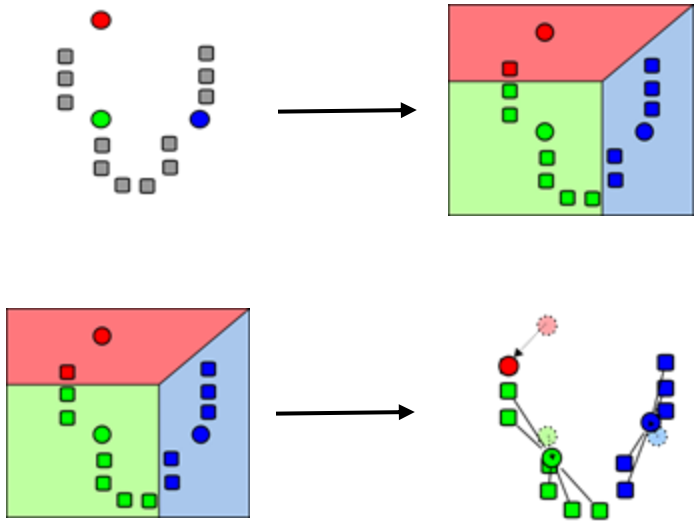
}

[Shah 2020]

Clustering: k-means



- *Lloyd's algorithm:*
  - A set of *n* points represent the clusters. Then other points can be assigned to clusters based on the closest representative point
  - Assign points to clusters. Then cluster representatives are determined by the mean distances of the points in each cluster.



[Pachter 2021]

The k-means clustering algorithm

- Given observations $(x_1,x_2,\ldots,x_n)$, partition the observations into k sets $C_1,C_2\ldots C_k$ to minimize the total within cluster sum-of-squares distances:

$$\operatorname{argmin}_S \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2.$$

- The iterative algorithm of starting with initial (randomly chosen) representatives, and then alternately assigning points to clusters, and re-computing cluster representatives, is known as *Lloyd's algorithm*.

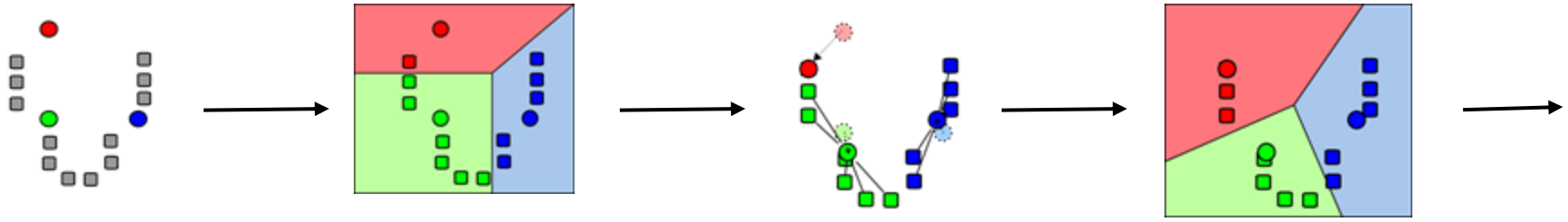
[Pachter 2021]

Lloyd's algorithm

- Randomly choose k points as representatives
- Assign points to clusters to minimize sum-of-squares distances
- Re-computing cluster representatives
- Repeat alternating assigning points to clusters and choosing representatives

[Pachter 2021]

Lloyd's algorithm



- Requires the value *k* as input.
- The algorithm may not converge to the optimal solution.
- Not the fastest approach to optimize for the *k-means* objective function.
- Finding the optimal solution to the *k-means* problem is *NP-hard* (in *k* and the dimension).

[Pachter 2021]

K-means algorithm

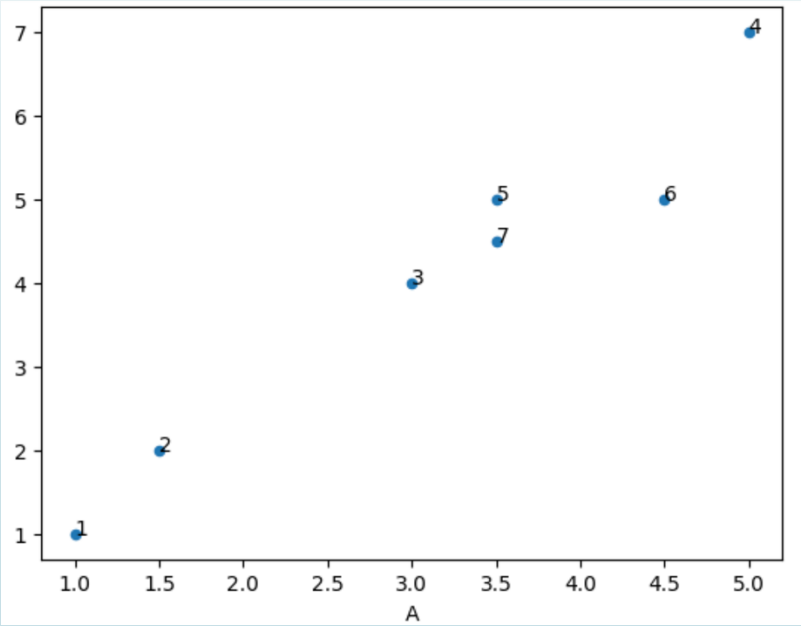
1. Begin with a decision on the value of K = number of clusters.
2. Put any initial partition that classifies the data into K clusters. You may assign the training samples randomly or systematically.
3. Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Repeat the above three steps until convergence is achieved.

[Shah 2020]

Tracing k-means clustering

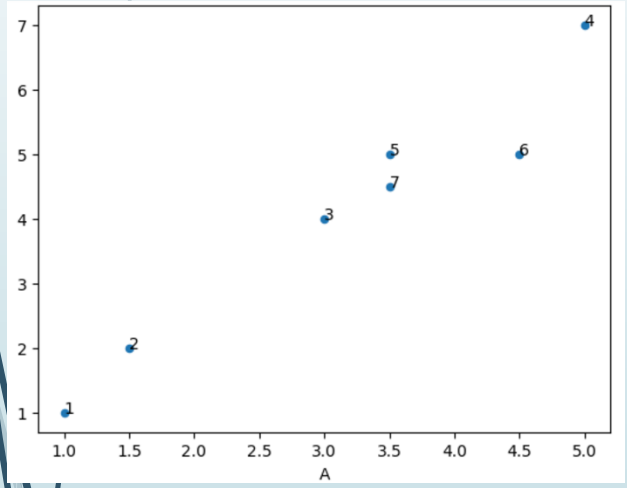
A	B	ID
1.0	1.0	1
1.5	2.0	2
3.0	4.0	3
3.5	4.5	7
3.5	5.0	5
4.5	5.0	6
5.0	7.0	4



[Shah 2020]

Tracing k-means clustering

A	B	ID
1.0	1.0	1
1.5	2.0	2
3.0	4.0	3
3.5	4.5	7
3.5	5.0	5
4.5	5.0	6
5.0	7.0	4



[Shah 2020]

Tracing k-means clustering

ID	Dist Cluster 1 centroid	Dist Cluster 2 centroid
1	1.5	5.4
2	.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	.7
6	3.8	.6
7	2.8	1.1

[Shah 2020]

Tracing k-means clustering

A	B	ID
1.0	1.0	1
1.5	2.0	2
3.0	4.0	3
3.5	4.5	7
3.5	5.0	5
4.5	5.0	6
5.0	7.0	4

[Shah 2020]

K-means algorithm

1. Begin with a decision on the value of K = number of clusters.
2. Put any initial partition that classifies the data into K clusters. You may assign the training samples randomly or systematically.
3. Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Repeat the above three steps until convergence is achieved.

[Shah 2020]

Today's Learning Objectives

Students will be able to:

- ✓ Review: Using SVD to perform PCA
- ✓ Understand what agglomerative clustering is
- ✓ Understand what agglomerative clustering is
- Evaluate clustering using quantitative metrics
- Apply clustering to a real LPGA data set

Within Sum of Squares (WSS)

Within-Cluster Sum of Squares:

$$WSS = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{i,j} - c_{k,j})^2$$

- **WSS** quantifies cluster tightness
- WSS decreases with number of clusters
- WSS increases with number of data points

[Yu, Barter 2024]

Silhouette score

For a single data point:

Let a_i be average distance from point i to other points in same cluster

Let b_i be average distance from point i to points in nearest cluster

Silhouette score:

$$\text{silhouette score}_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

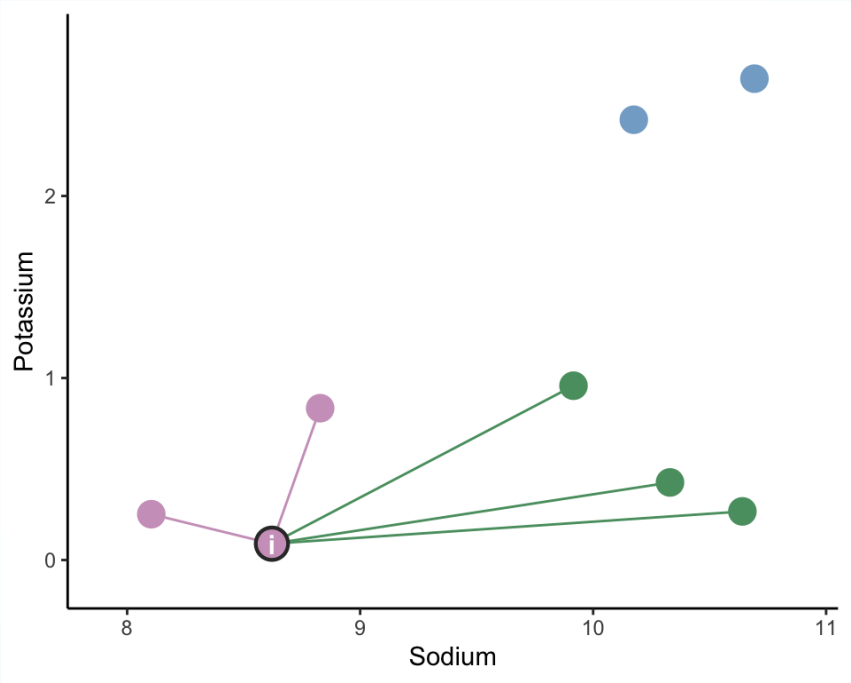
- **Silhouette** quantifies how tightly each data point is clustered and how distinctly each data point is clustered
- Average silhouette scores for each point to get clustering silhouette score

[Yu, Barter 2024]

Silhouette score

For a single data point:

$$\text{silhouette score}_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$



- 1 means well clustered
- 0 means equally close to either cluster
- -1 means poorly clustered

[Yu, Barter 2024]

Rand Index

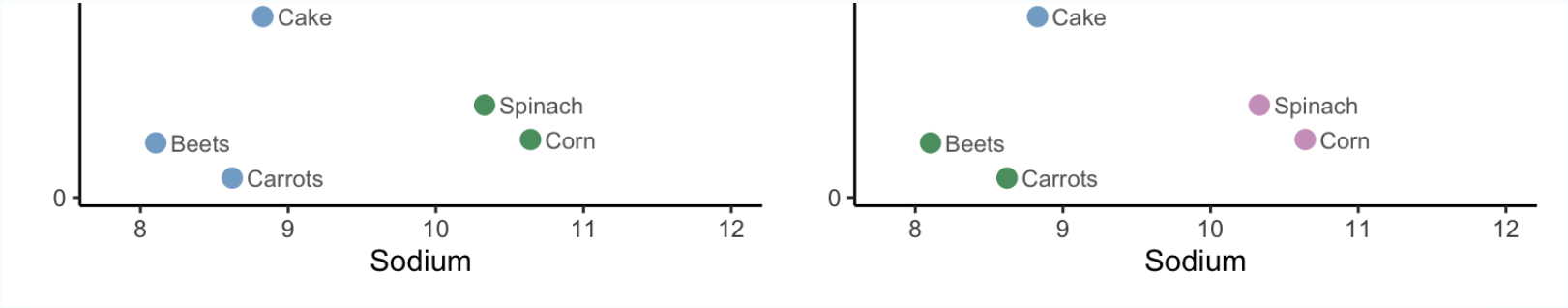
How to compare different sets of clusters?

For each pair of data points, consider if the clusters put the Same pair of points in the same cluster or different clusters.

Rand Index = # pairs in agreement/ total # of pairs

[Yu, Barter 2024]

Calculating Rand Index



[Yu, Barter 2024]

Today's Learning Objectives

Students will be able to:

- ✓ Review: Using SVD to perform PCA
- ✓ Understand what agglomerative clustering is
- ✓ Understand what agglomerative clustering is
- ✓ Evaluate clustering using quantitative metrics
- Apply clustering to a real LPGA data set

Your turn:
Clustering on LPGA data

Please get the Jupyter notebook for LPGA 2008 season data:

Go to:

The data file on BruinLearn Week 7 Module:

lpga2008.dat.txt

Notebook:
https://colab.research.google.com/drive/1LFqbyV7mr6jZup16zVfB1Fgh9TwxU-7f?usp=sharing

Save a copy to your Google Drive and keep notes there...

[Shah 2020]

Today's Learning Objectives

Students will be able to:

- ✓ Review: Using SVD to perform PCA
- ✓ Understand what agglomerative clustering is
- ✓ Understand what agglomerative clustering is
- ✓ Evaluate clustering using quantitative metrics
- ✓ Apply clustering to a real LPGA data set

Citations:

Shah. C. (2020) A hands-on introduction to data science. Cambridge University Press.

Yu, B., & Barter, R. L. (2024). Veridical data science: The practice of responsible data analysis and decision making. The MIT Press.

Shah. C. (2020) A hands-on introduction to data science. Cambridge University Press.

Thank You