DATASCI207-005/007 Applied Machine Learning

Vilena Livinsky, PhD(c)

School of Information, UC Berkeley

Week 8: 02/26/2025 & 02/27/2025

Today's Agenda

- Unsupervised Learning
  - K-Means & other clustering algos
  - Gaussian Mixture Model (GMM)
  - PCA/SVD
- Walkthroughs:
  - K-Means & other clustering algos
    - K-Means: Cereal Brands (Product)
  - Gaussian Mixture Model
    - Image Segmentation: MRI images
  - PCA/SVD
    - Dimensionality reduction/ image compression

Final Project: Step 2

Form a group	3-4 people, max 4 NO silos or groups of 2 Groups can only be composed of you and your colleagues in your section
Inform me & the class of your formed group in Slack	Include names of group members Due date: 01/24/2025 EOD
General Plan:	~~Step 1: form a group~~ Step 2: submit your group's question to answer/ goal + dataset Step 3: baseline presentation Step 4: final presentation
Dates	Step 2: 03/13/2025 EOD Step 3: 04/03/2025 Step 4: 04/17/2025

Final Project: Logistics/ Due Dates

Final Project Timeline/ Deliverables

- Step1: See groups on Slack
- **Step 2: Select dataset** and identify a leading **question/goal** for your use-case
  - notify via email of dataset + question selection for you group (vlivinsky@ischool.berkeley.edu)
  - Due date: 03/13/2025 EOD
- **Step 3: Baseline** group presentation (10 mins)
  - Due date: 04/03/2025
- **Step 4: Final** Project Presentations (15 mins)
  - Due date: 04/17/2025
- For past project examples refer to: Cornelia Paulik

Make sure your **baseline presentation** slides include:

- Title, Authors
- What is the **question** you will be working on? Why is it interesting?
- What is the **data** you will be using? Include the data source, size of dataset, main features to be used. Please also include summary statistics of your data.
- What prediction **algorithms** do you plan to use? Please describe them in detail.
- How will you **evaluate** your results? Please describe your chosen performance metrices and/or statistical tests in detail.

Refer to bcourses home page for **final presentation** guidelines and grading

- Note that the final project grade is individual and is based on each member's contribution
- Final project team member reviews to be submitted at end of class—a survey will be sent

(Some) Example Data Sources

UCI Machine Learning Repository: https://archive.ics.uci.edu/datasets,

data hosted at data gov: https://data.gov/,

you can also utilize Google's datasetsearch: https://datasetsearch.research.google.com/?source=post_page-----bb6d0dc3378b----------

and of course there's Kaggle: https://www.kaggle.com/

Do NOT use common ML datasets such as the iris data set, the mushroom data set, the titanic data set, etc. (& anything you've used in your homework assignments)

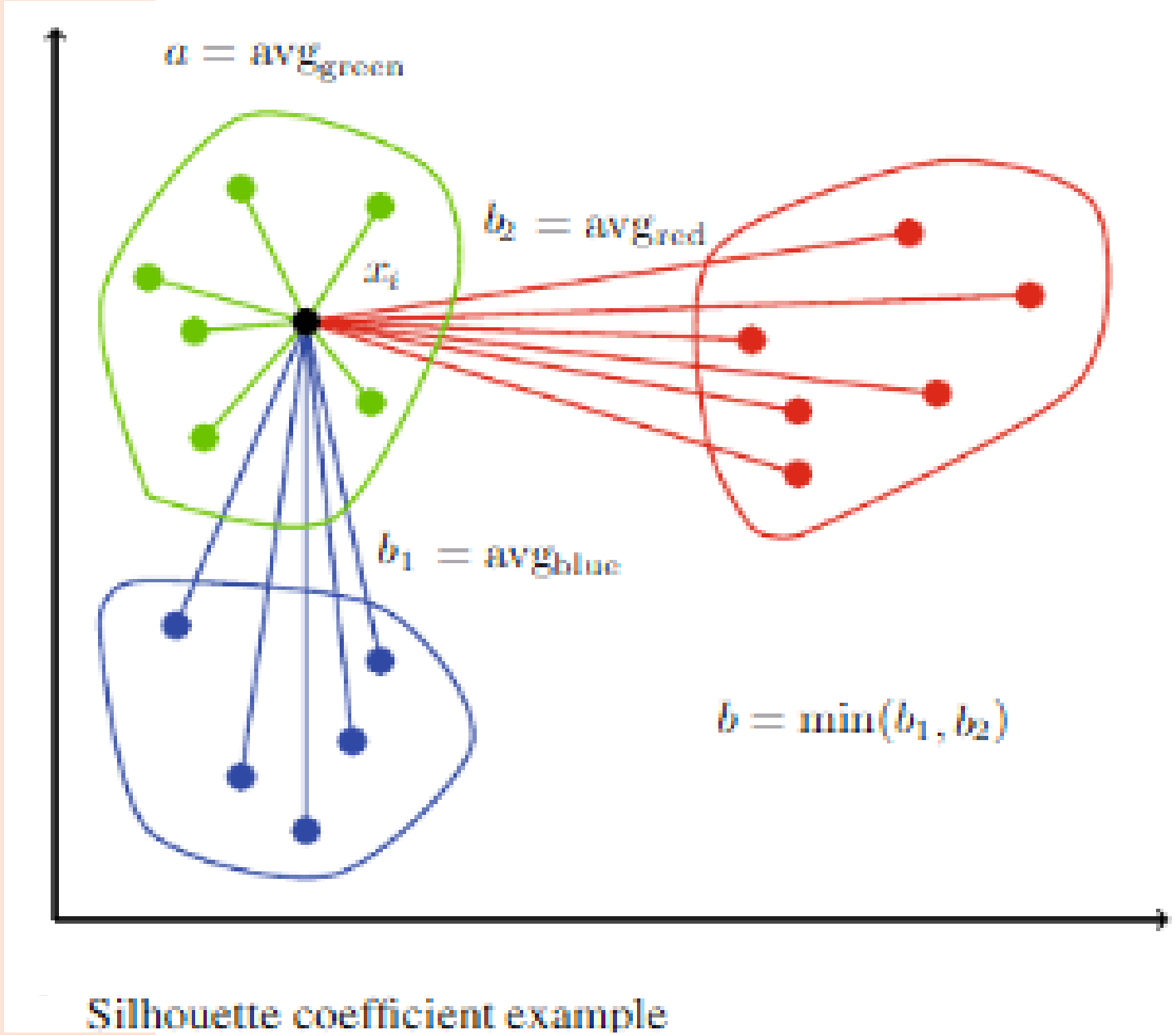
Silhoutte Coefficient

- Cluster separation
- Cluster cohesion

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{max\{ b^{(i)}, a^{(i)}\}}$$

- Properties: si ∈ [−1, 1]

- Image Ref.: Pai, S., Troia, F. D., Visaggio, C. A., Austin, T. H., & Stamp, M. (2017). Clustering for malware classification. Journal of Computer Virology and Hacking Techniques, 13, 95-107.



Silhouette coefficient example