# Model AI Assignment: An Introduction to *k*-Means Clustering

Todd W. Neller, Gettysburg College
Laura E. Brown, Michigan Technological University

# Why Clustering?

- In the Model AI assignment repository (http://modelai.gettysburg.edu), assignments available for:
  - search, genetic algorithms, constraint satisfaction, supervised learning, reinforcement learning, etc.
- Clustering is a major topic not included
- Informal EAAI-14 poll indicated greatest need for
  - **unsupervised learning** teaching support materials, and
  - ***k*-Means Clustering** was the **best representative algorithm**.
- **Goal:** Collect and collate resources on clustering from textbooks, general web resources, and coverage in MOOCs; develop assignments for experiential learning

# Assignment Learning Objectives

- Define unsupervised learning and distinguish it from supervised learning
- Define and implement *k*-means clustering
- Understand the limitations of *k*-means clustering
  - what are the assumptions
  - when does the method fail
- Implementation considerations
  - how to initialize cluster centers
  - how to select *k*
- Allow for instructor extensions
  - use of other clustering methods (hierarchical, spectral, *k*-medoids), apply to real-world data, etc.

# Clustering Problem

- **Clustering** is grouping a set of objects such that objects in the same group (i.e. cluster) are more similar to each other in some sense than to objects of different groups
- Our specific clustering problem:
    - Given: a set of *n* observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, where each observation is a *d*-dimensional real vector
    - Given: a number of clusters *k*
    - Compute: a cluster assignment mapping $C(\boldsymbol{x}_i) \in \{1, \ldots, k\}$ that minimizes the **within cluster sum of squares (WCSS)**

# $k$-Means Clustering Algorithm

- General algorithm:
  - Randomly choose k cluster centroids **μ**_1,**μ**_2, … **μ**_k and arbitrarily initialize cluster assignment mapping C.
  - While remapping C from each **x**_i to its closest centroid **μ**_j causes a change in C:
    - Recompute **μ**_1,**μ**_2, … **μ**_k according to the new C
- In order to minimize the WCSS, we alternately:
  - Recompute C to minimize the WCSS holding **μ**_j fixed.
  - Recompute **μ**_j to minimize the WCSS holding C fixed.
- In minimizing the WCSS, we seek a clustering that minimizes Euclidean distance *variance* within clusters.

# Assignment Details

Part 1

- Students will implement *k*-means clustering method
- Run the implementation repeatedly over a set of test cases
  - Objective 1: Define and implement *k*-means clustering
  - Objective 2: Understand the limitations of *k*-means clustering
    - what are the assumptions
    - when does the method fail

# Assignment Details

Part 2

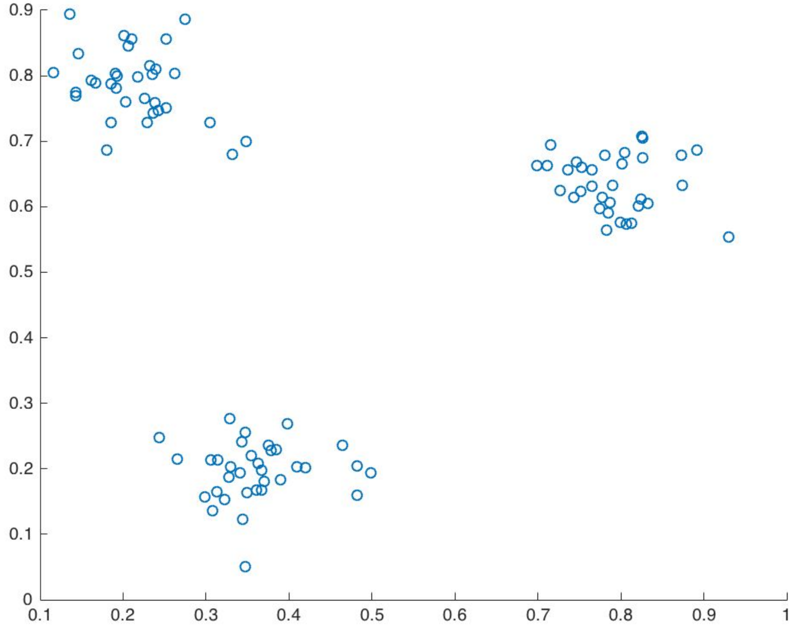
- Implement iterated *k*-means (10 runs, returns clustering with the lowest WCSS)
- Run the implementation repeatedly over a set of test cases
    - Objective 2: Understand the limitations of *k*-means clustering
        - what are the assumptions
        - when does the method fail
    - Objective 3: Implementation considerations

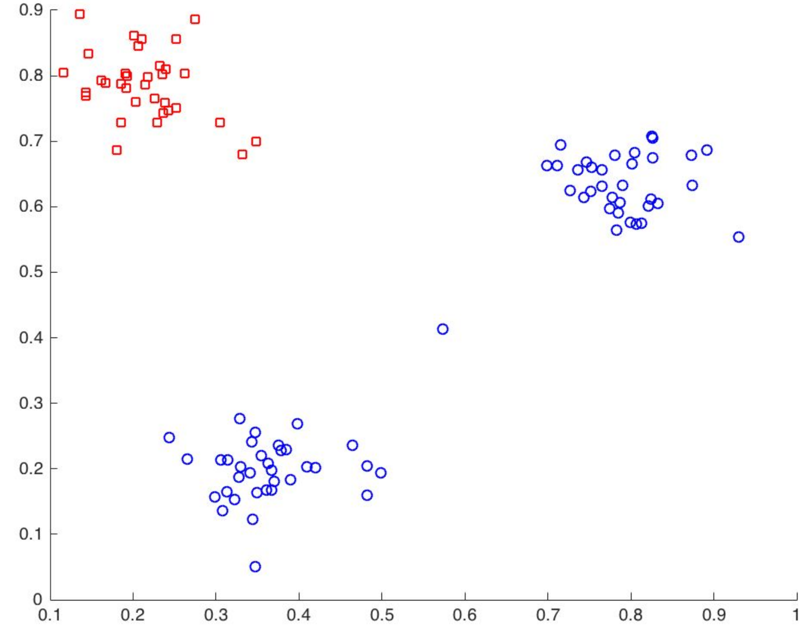
# Assignment Details

Part 3

- Select best *k* value using simplified gap statistic (most significant logarithmic difference between uniform WCSS and observed WCSS)
- Run the implementation repeatedly over a set of test cases
    - Objective 2: Understand the limitations of *k*-means clustering
        - what are the assumptions
        - when does the method fail
    - Objective 3: Implementation considerations
        - how to select *k*

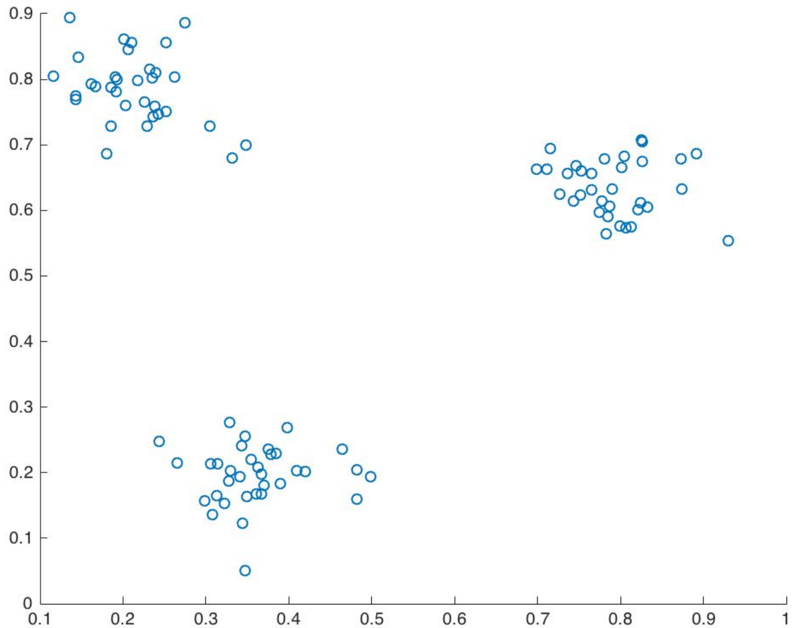
# Example Test Cases

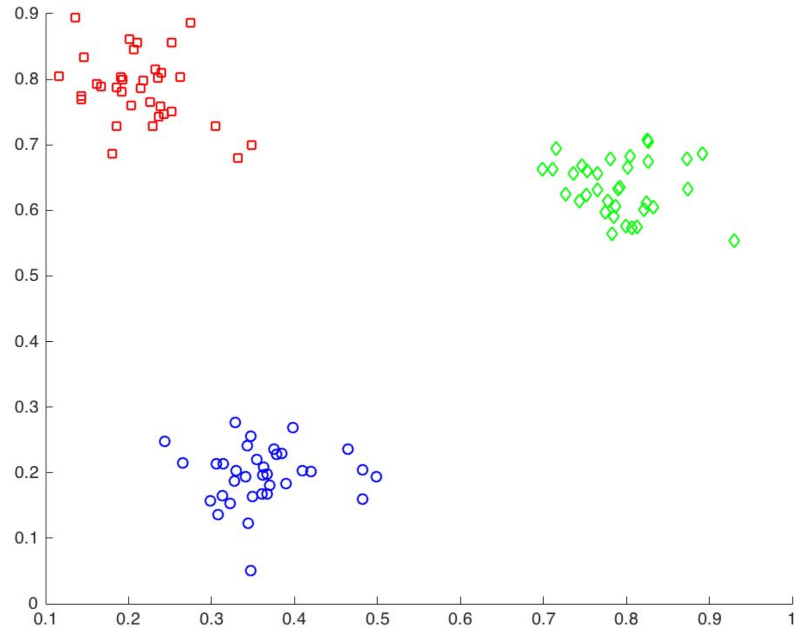



Easy Gaussian, k=2

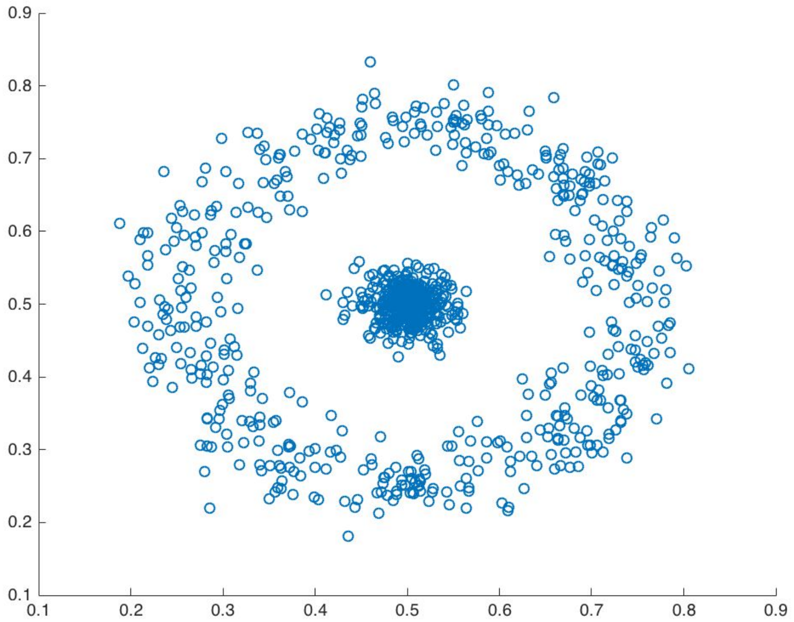
# Example Test Cases

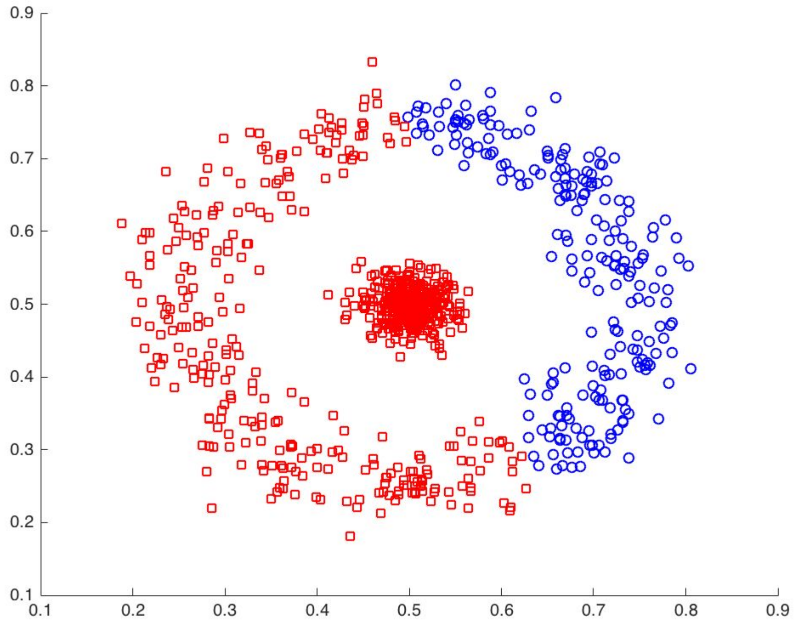



Easy Gaussian, k=3

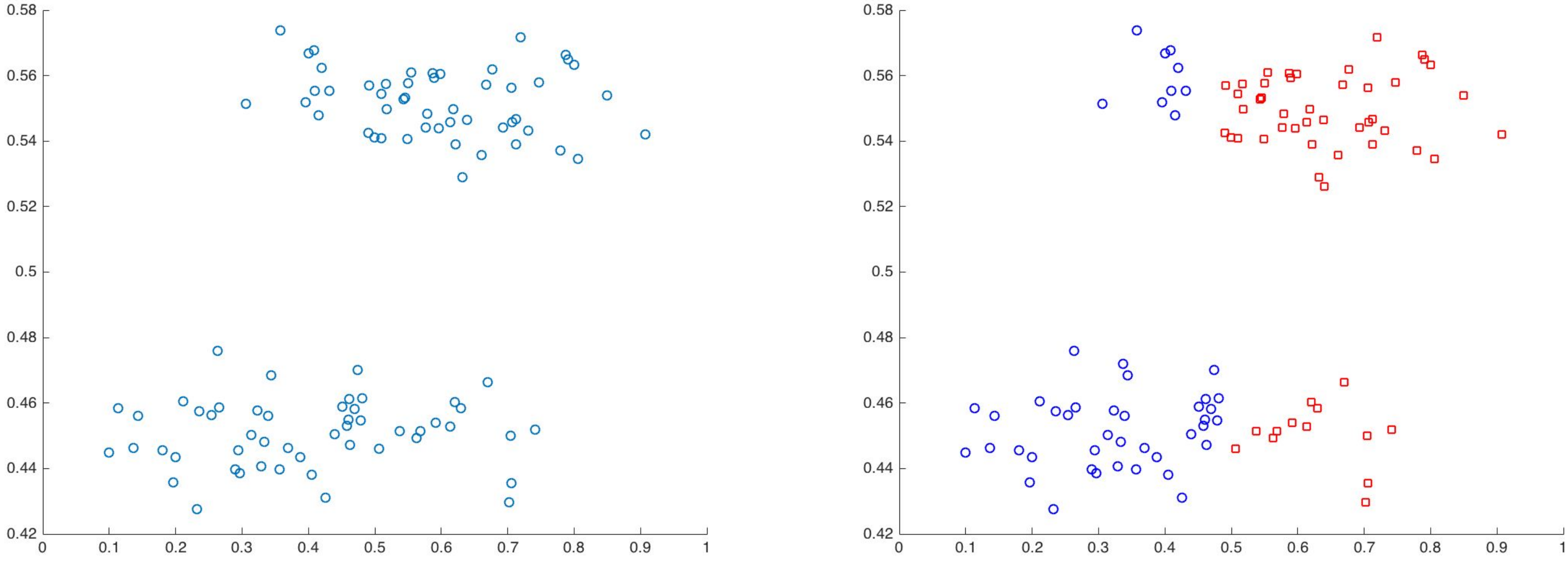
# Example Test Cases




Bullseye

# Example Test Cases




Stretched Distribution

# Summary of Resources

- Slides introducing clustering
- Assignment
  - k-Means, iterated k-Means, iterated k-Means with Gap Statistic
  - data sets showing strengths and weaknesses
  - MATLAB/Octave visualization scripts
  - Learning objectives and mapping to ACM/IEEE CS2013 Curricula
- Index to excellent pre-existing resources online
  - Textbooks, websites, demos, software, videos, MOOCs
  - K-Means Clustering Notation Guide PDF to translate
- Weka tutorial with iris data demonstrating feature selection
- Real-world data sets