CSE 411: Machine Learning

Clustering

Dr Muhammad Abul Hasan

Department of Computer Science and Engineering
Green University of Bangladesh
muhammad.hasan@cse.green.edu.bd

Fall 2023

Outline

1. Unsupervised Learning

2. Clustering

3. K-Means Clustering

4. K Means Clustering Example

5. Finding optimum K

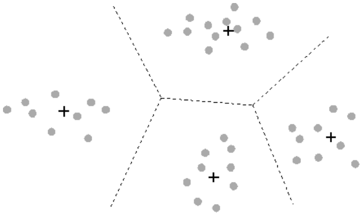
> *“People worry that computers will get too smart and take over the world, but the real problem is that they’re too stupid and they’ve already taken over the world.*
>
> *– Pedro Domingos”*

Unsupervised Learning

- **Supervised learning**:
  - Predict target value y given features $\mathbf{x}$.
- **Unsupervised learning**:
  - Understand patterns of data (just $\mathbf{x}$)
  - Useful for many reasons:
    - Data mining (“explain”)
    - Missing data values (“impute”)
    - Representation (feature generation or selection)
- Example of unsupervised learning: Clustering

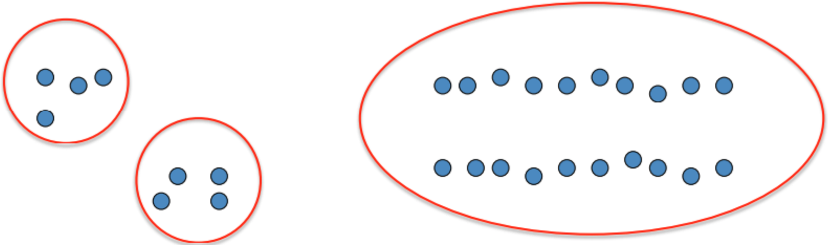
Clustering and Data Compression

- Clustering is related to vector quantization
  - Dictionary of vectors (the cluster centers)
  - Each original value is represented using a dictionary index
  - Each center "claims" a nearby region (Voronoi region)



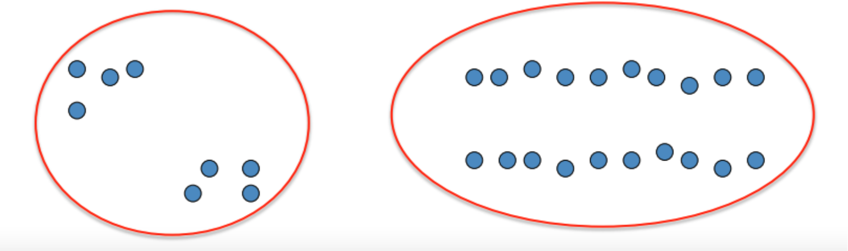
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



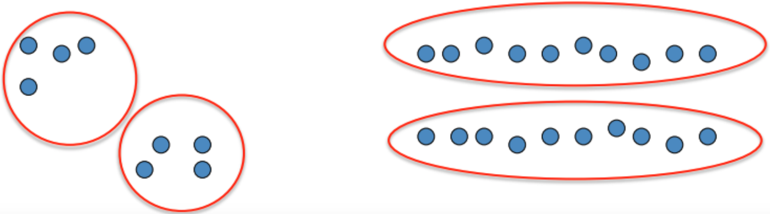
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



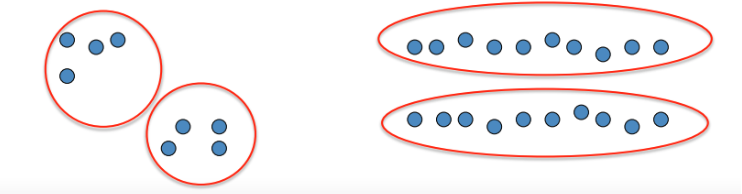
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns

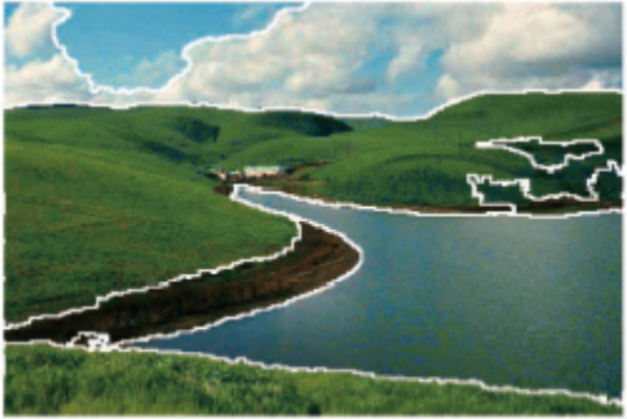


What could "similar" mean?

- One option: small Euclidean distance (squared)
- Clustering results are crucially dependent on the measure of similarity (or distance) between "points" to be clustered

Clustering examples

- Image segmentation
- Goal: Break up the image into meaningful or perceptually similar regions

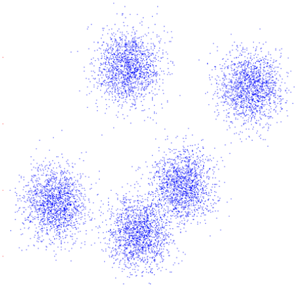


K-Means Clustering

- A simple clustering algorithm
- Iterate between
  - Updating the assignment of data to clusters
  - Updating the cluster's summarization
- Suppose we have K clusters $c = 1 \ldots K$

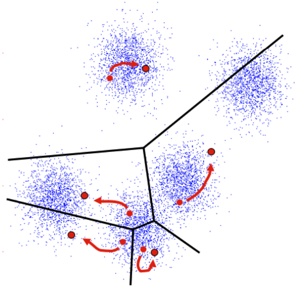
K-Means

- An iterative clustering algorithm
    - Initialize: Pick K random points as cluster centers
    - Alternate:
        1. Assign data points to the closest cluster center
        2. Change the cluster center to the average of its assigned points
    - Stop when no points assignments change

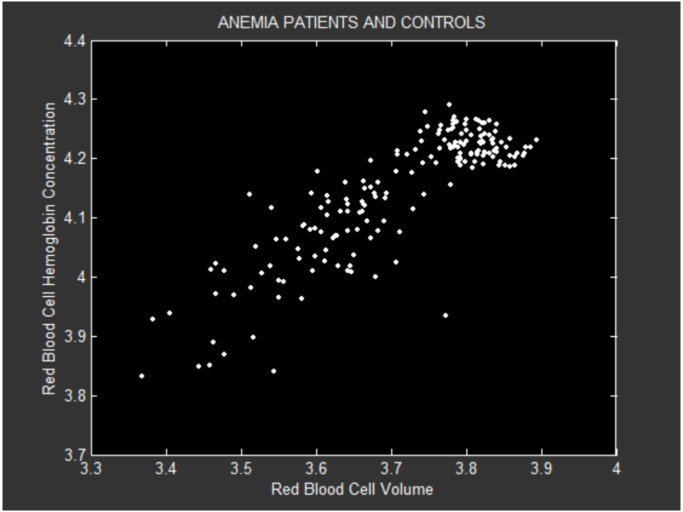


K-Means

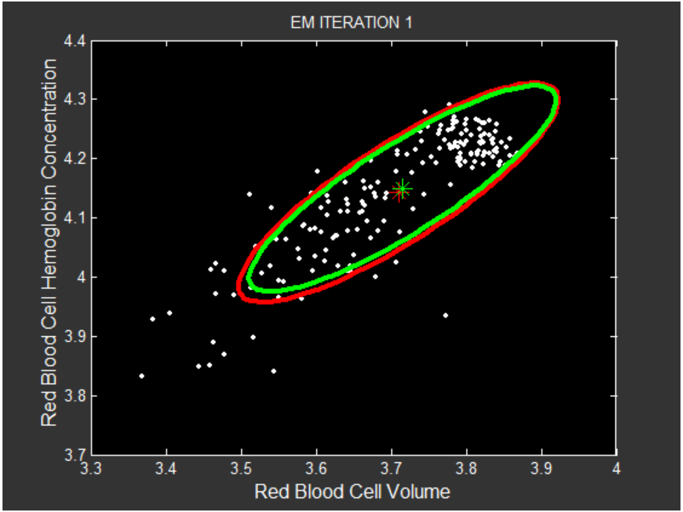
- An iterative clustering algorithm
  - Initialize: Pick K random points as cluster centers
  - Alternate:
    1. Assign data points to the closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points assignments change



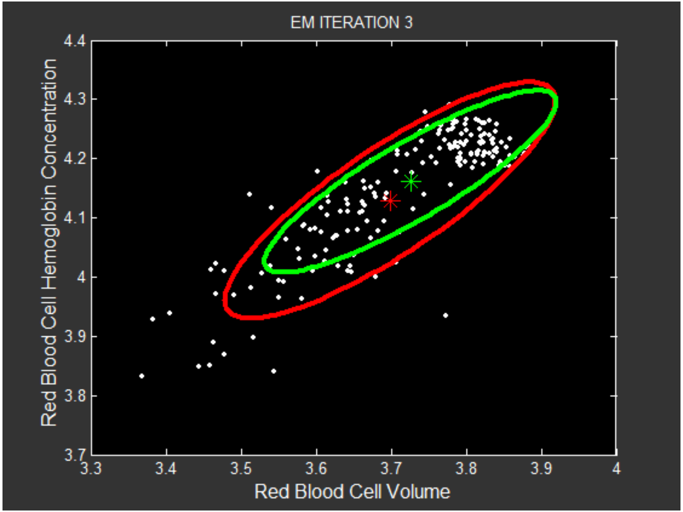
Clustering Example



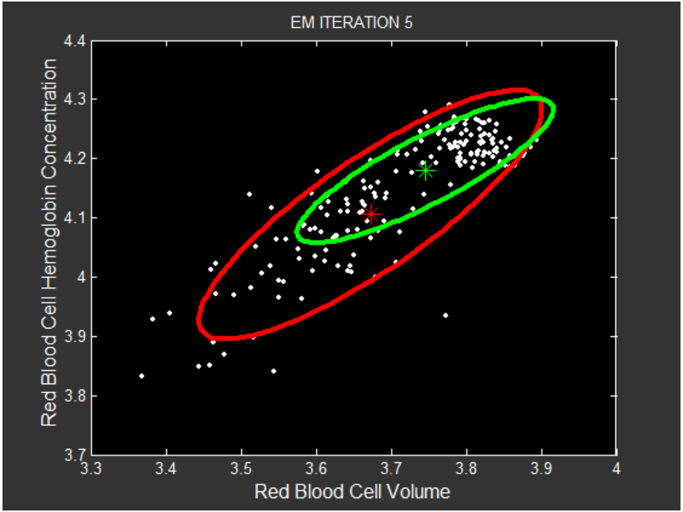
Clustering Example



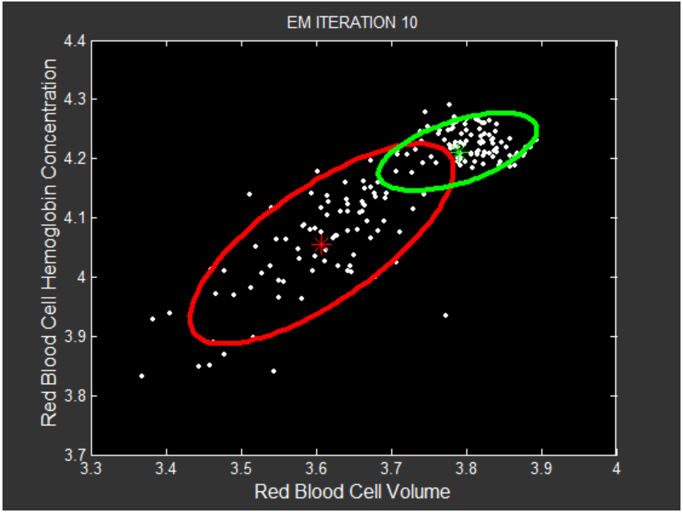
Clustering Example



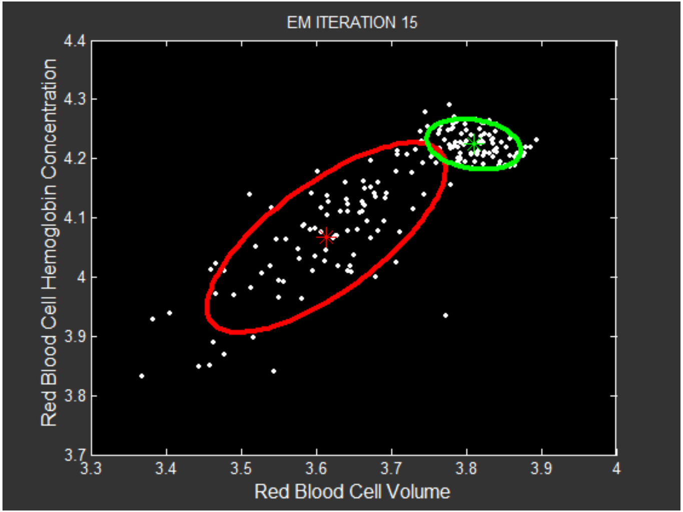
Clustering Example



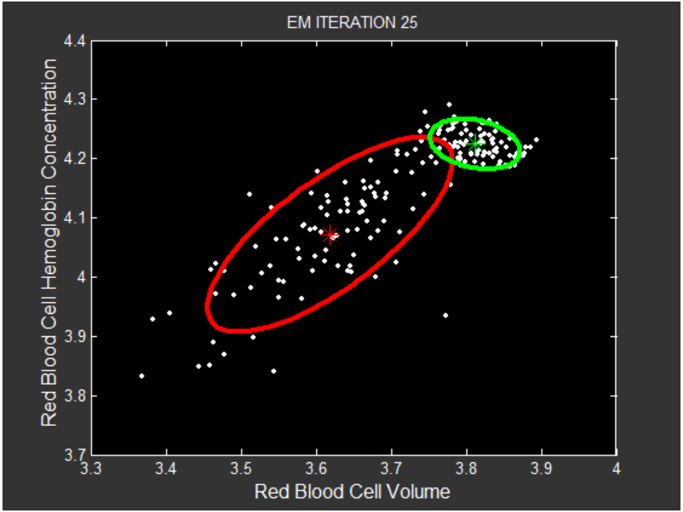
Clustering Example



Clustering Example



Clustering Example



Properties of K means algorithm

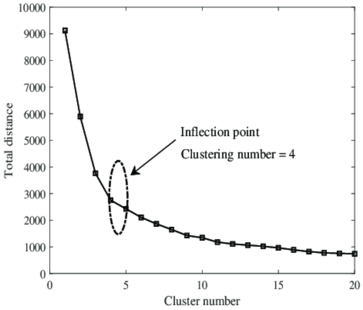
- Guaranteed to converge in a finite number of iterations.
- Running time per iteration:
  1. Assign data points to the closest cluster center $\mathcal{O}(KN)$ time
  2. Change the cluster center to the average of its assigned points $\mathcal{O}(N)$

How to find the number of clusters in K-means?

- K is a hyperparameter to the k-means algorithm.
- In most cases, the number of clusters K is determined in a heuristic fashion.
- Most strategies involve running K-means with different values of K – and finding the best value using some criteria. One of the two most popular criteria used is the elbow method.

Elbow Method

- The elbow method involves finding optimum values of K and finding the elbow point.
- At first, the quality of clustering improves rapidly when changing the value of K, but eventually stabilizes.
- The elbow point is where the relative improvement is not very high anymore.



Thank You!