CPSC 340: Machine Learning and Data Mining

Density-Based Clustering

Fall 2015

Admin

- Tutorials today.
- Office hours tomorrow
- Assignment 2 due Friday.

K-Means++

- Steps of k-means++:
  1. Select initial mean μ_1, from among the object x_i.
  2. Compute distance d_{ic} of object x_i to each mean μ_c.

     $$d_{ic} = \|x_i - \mu_c\| = \sqrt{\sum_{j=1}^{d} (x_{ij} - \mu_c)^2}$$

  3. For each object set d_i to the minimum distance across all clusters c.

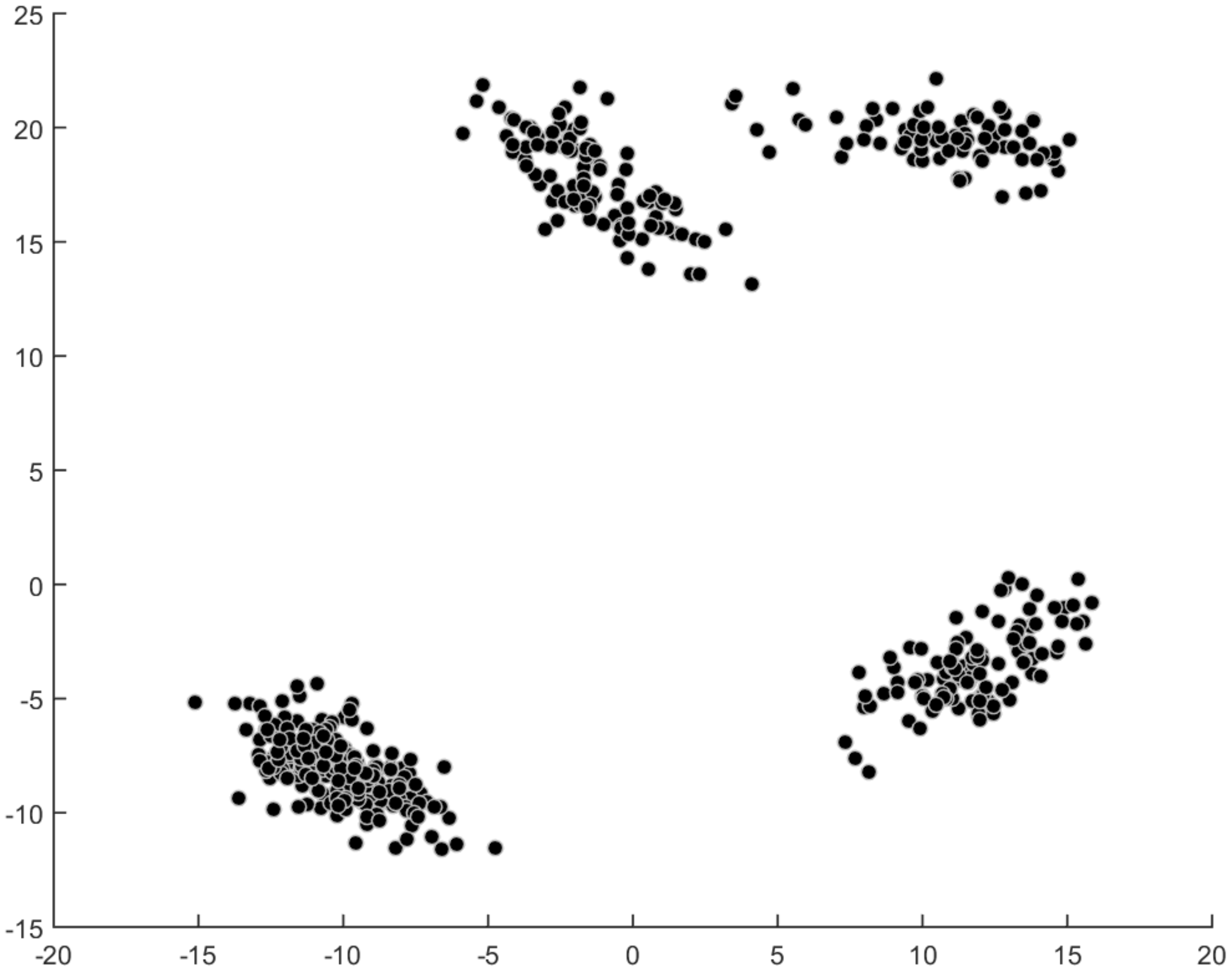
     $$d_i = \min_c \{d_{ic}\}$$

  4. Choose next mean by sampling proportional to $(d_i)^2$.

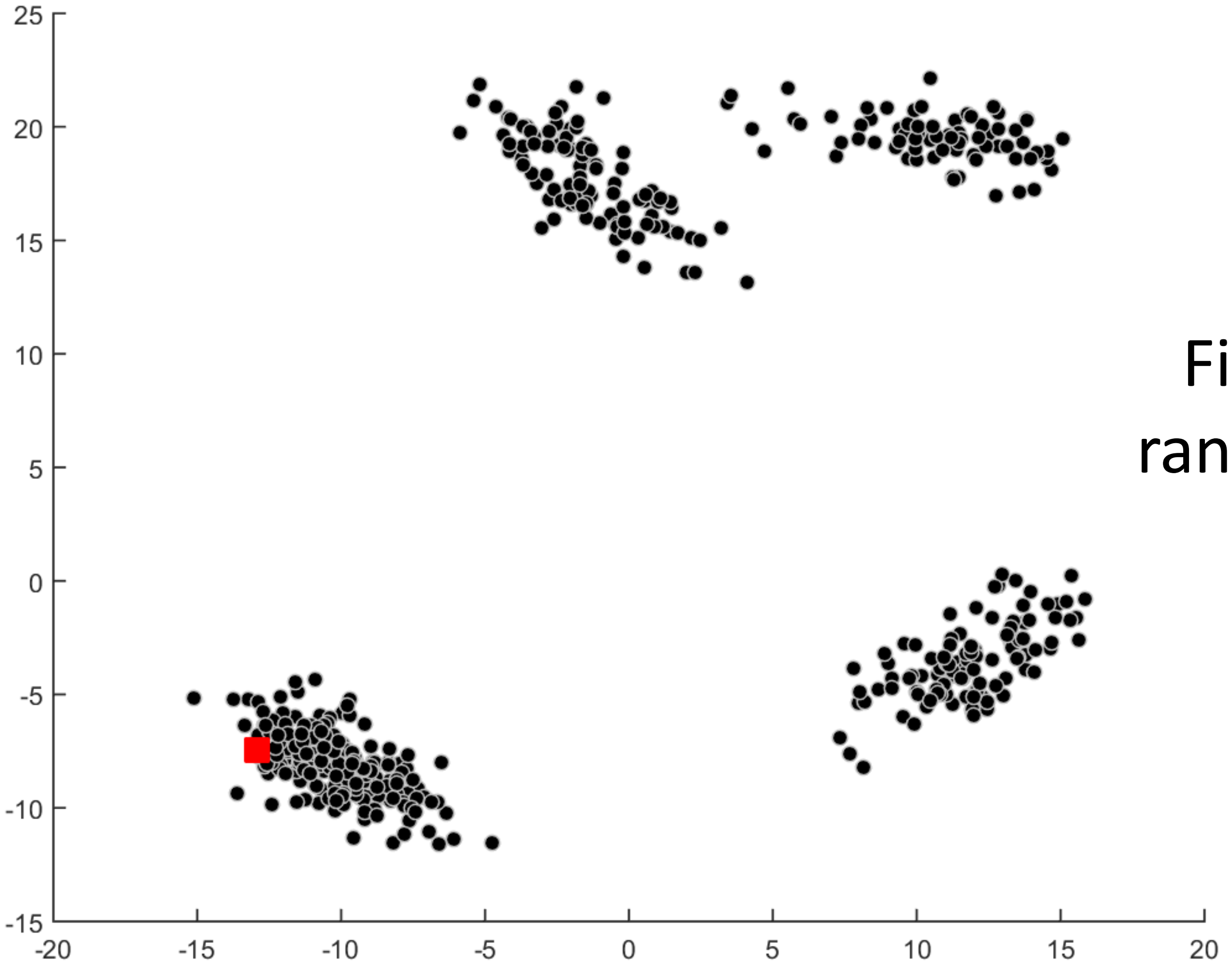
     $$p_i \propto d_i^2 \Longrightarrow p_i = \frac{d_i^2}{\sum_{j=1}^{n} d_j^2}$$

  5. Stop when we have k means, otherwise return to 2.
- Expected approximation ratio is O(log(k)).

K-Means++

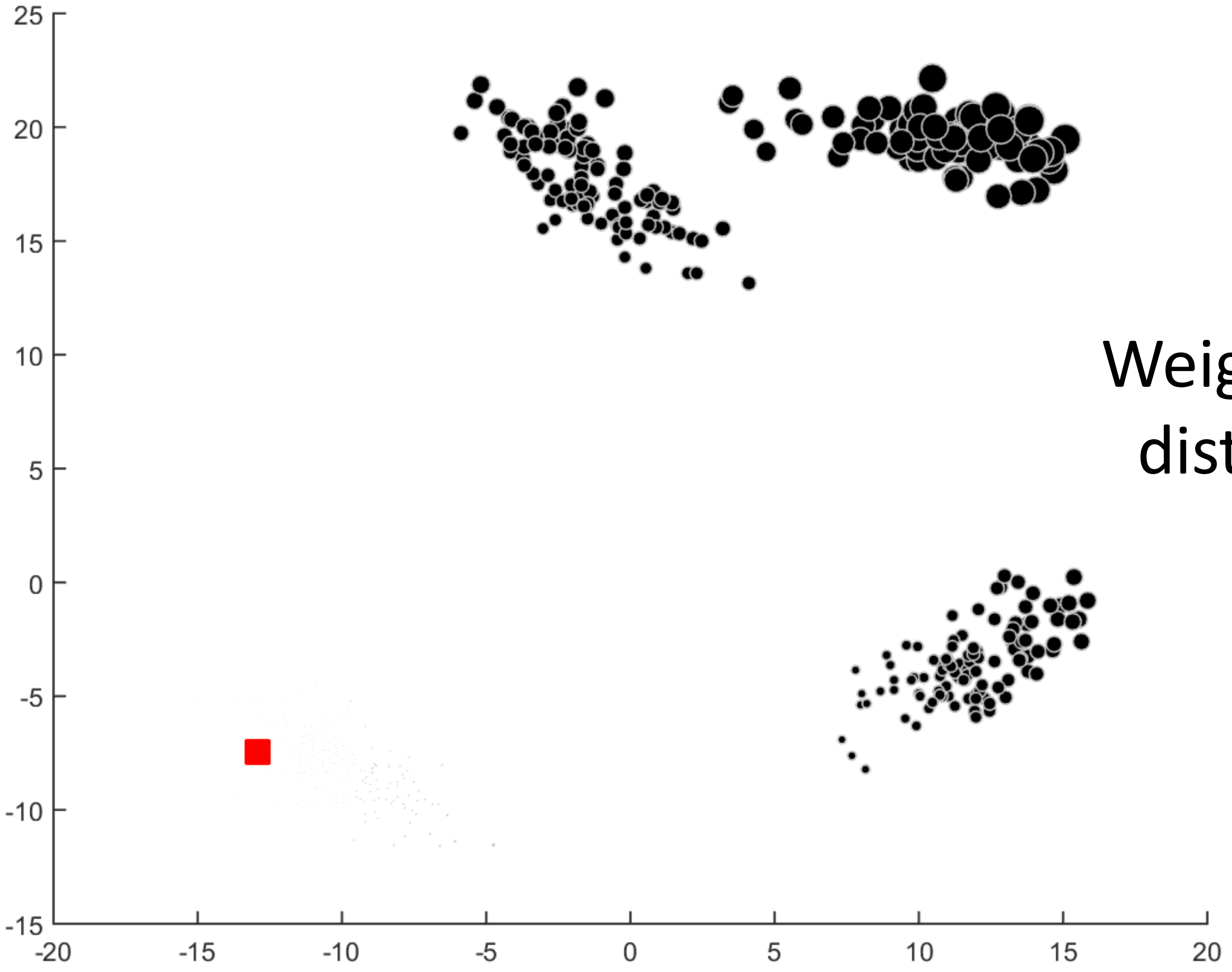


K-Means++



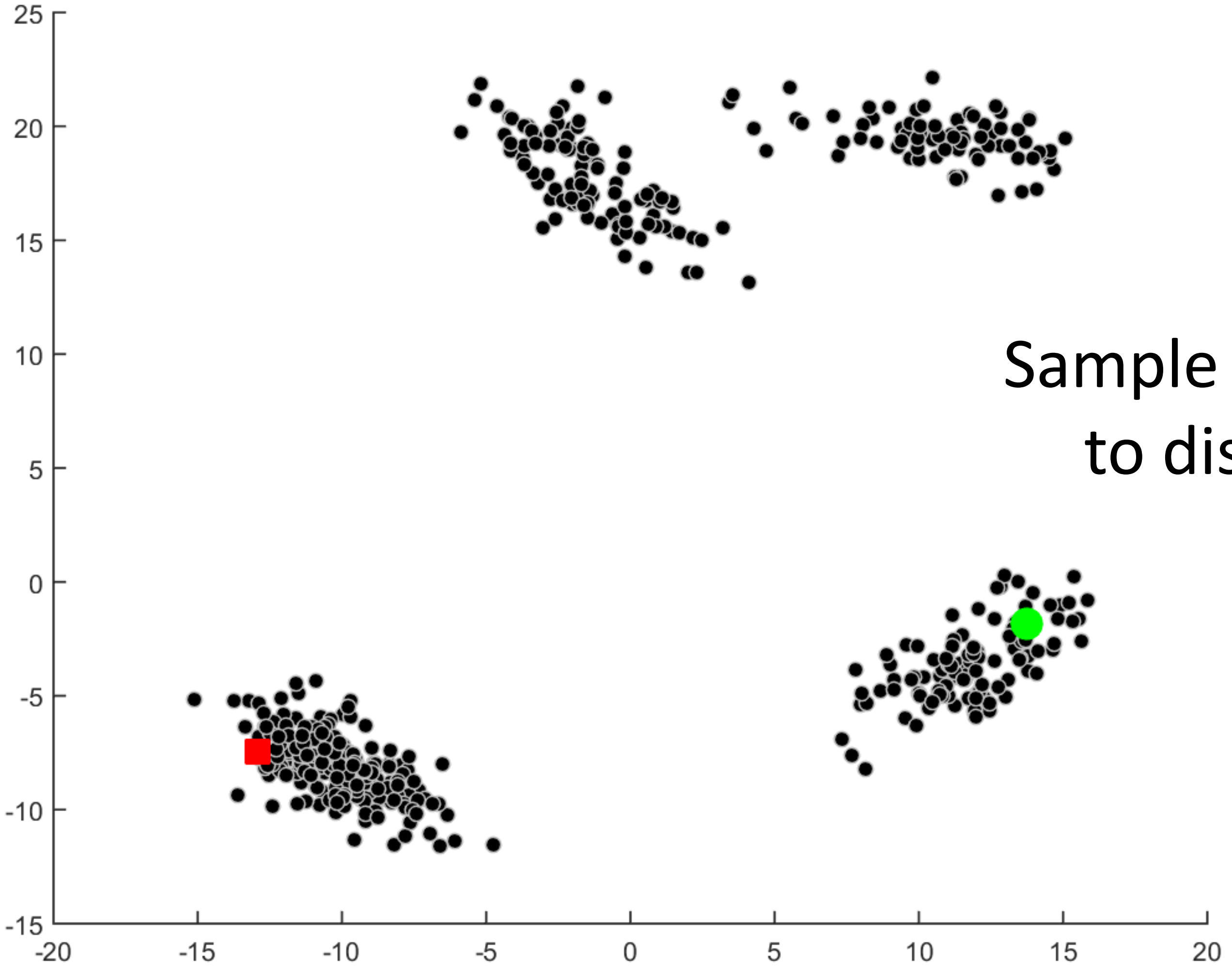
First mean is a random example.

K-Means++



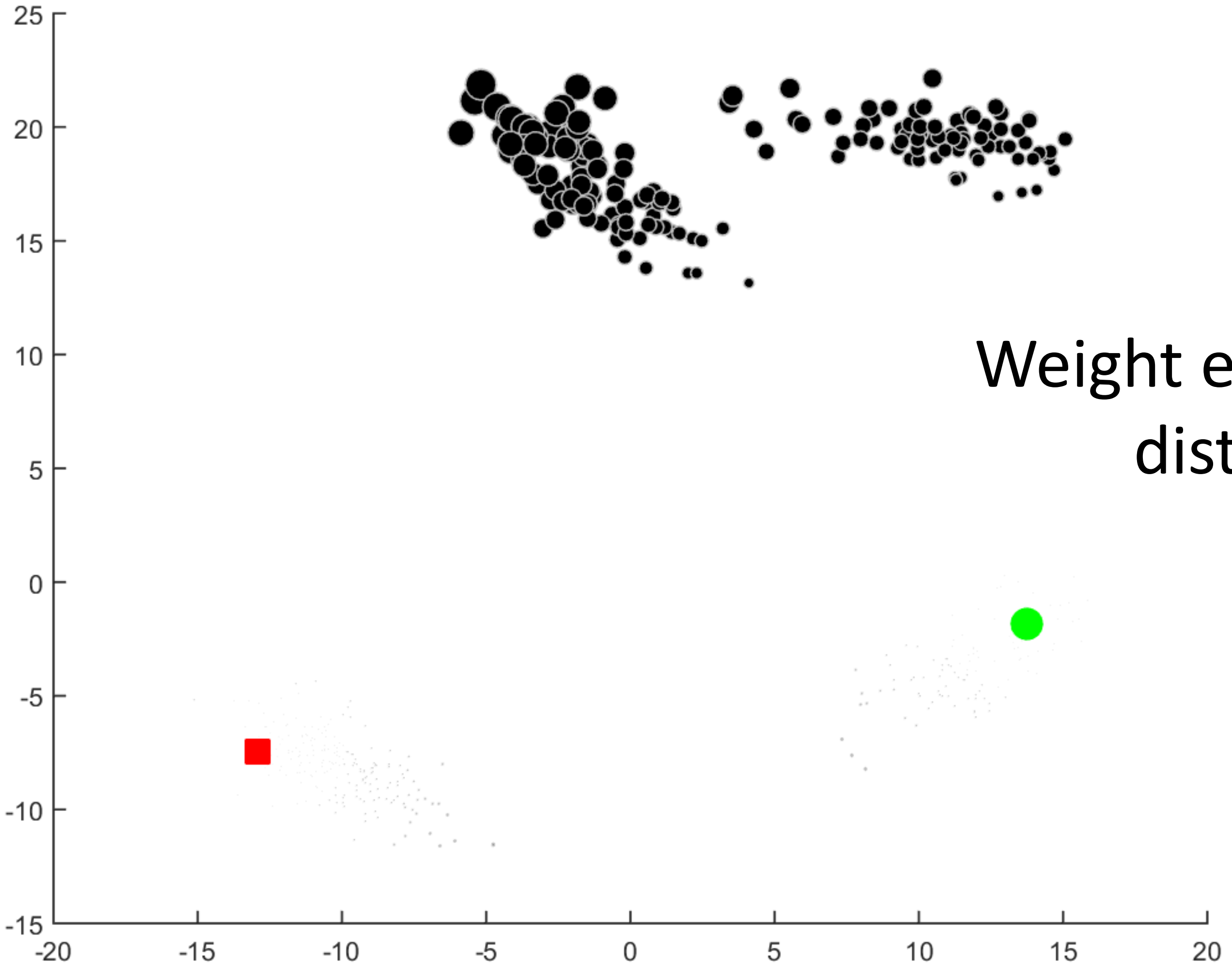
Weight examples by distance squared.

K-Means++



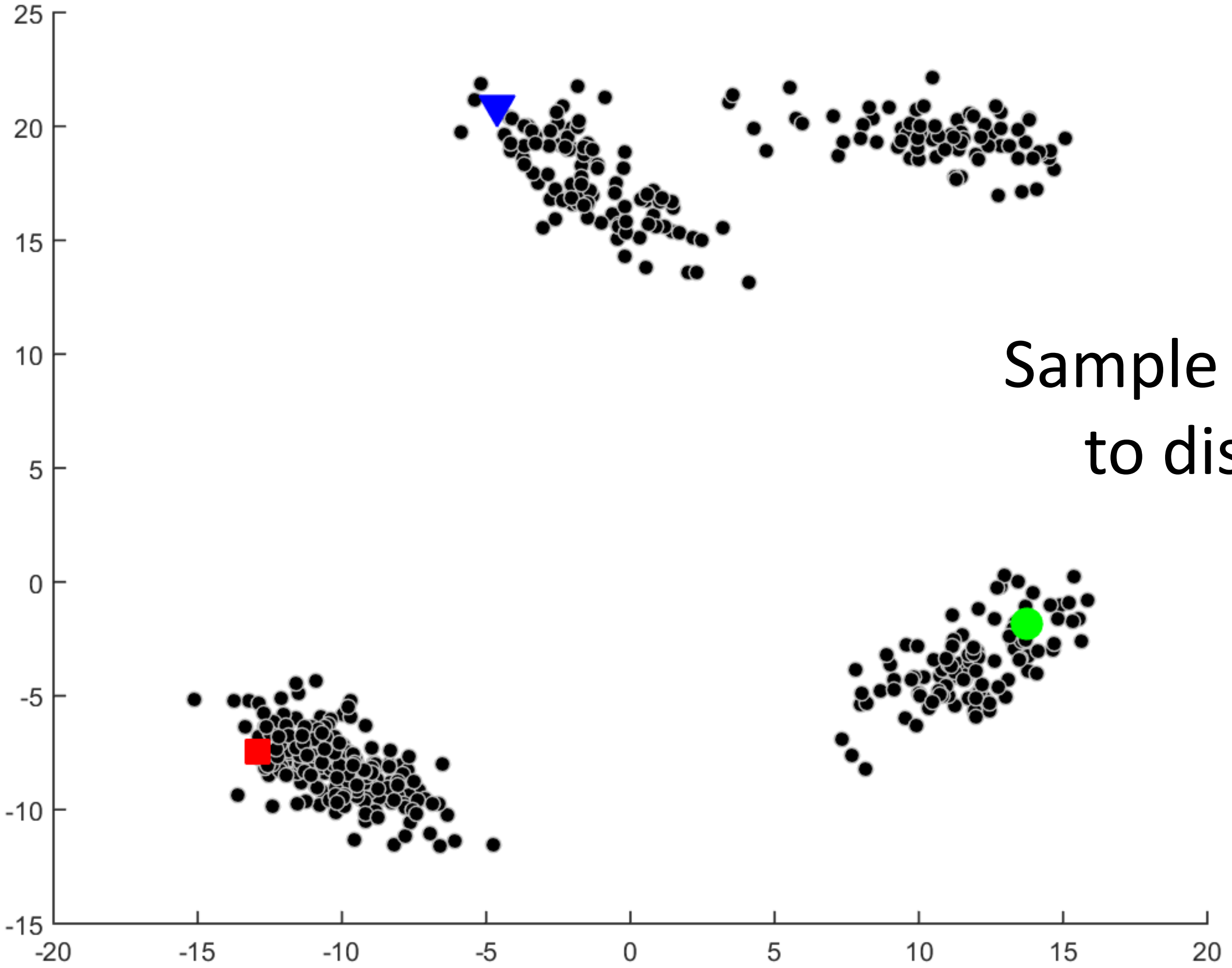
Sample mean proportional to distances squared.

K-Means++



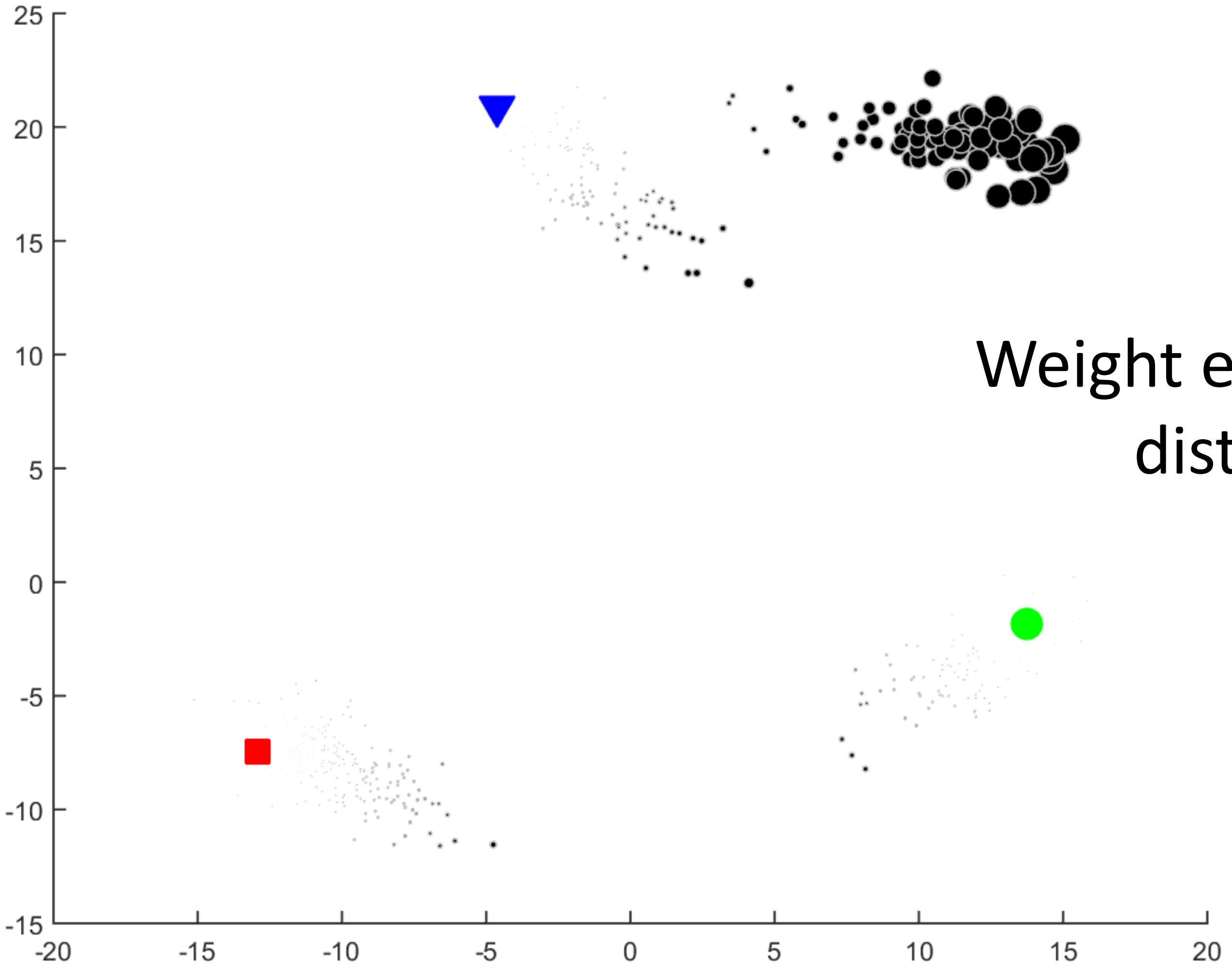
Weight examples by squared distance to mean.

K-Means++



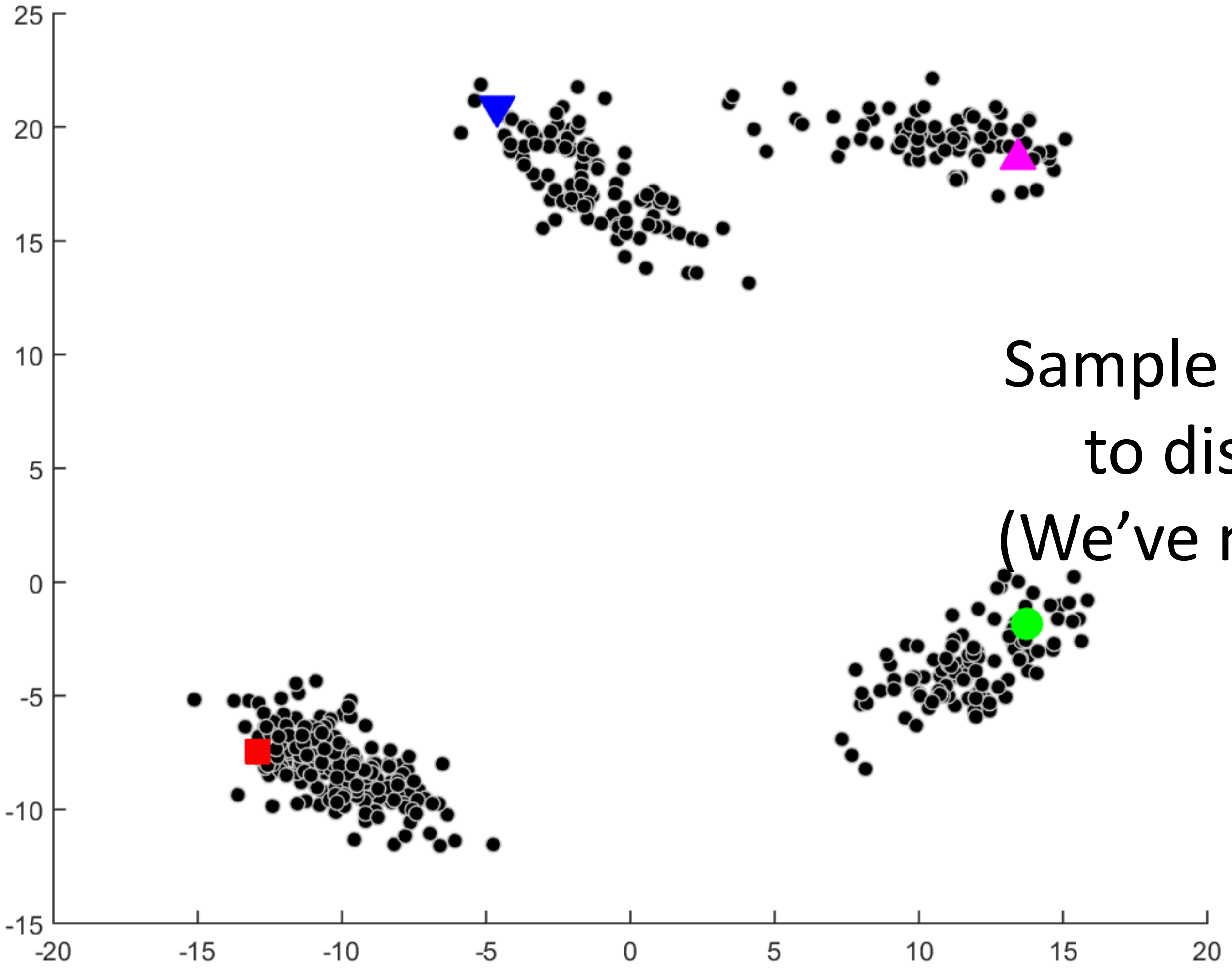
Sample mean proportional to distances squared.

K-Means++



Weight examples by squared distance to mean.

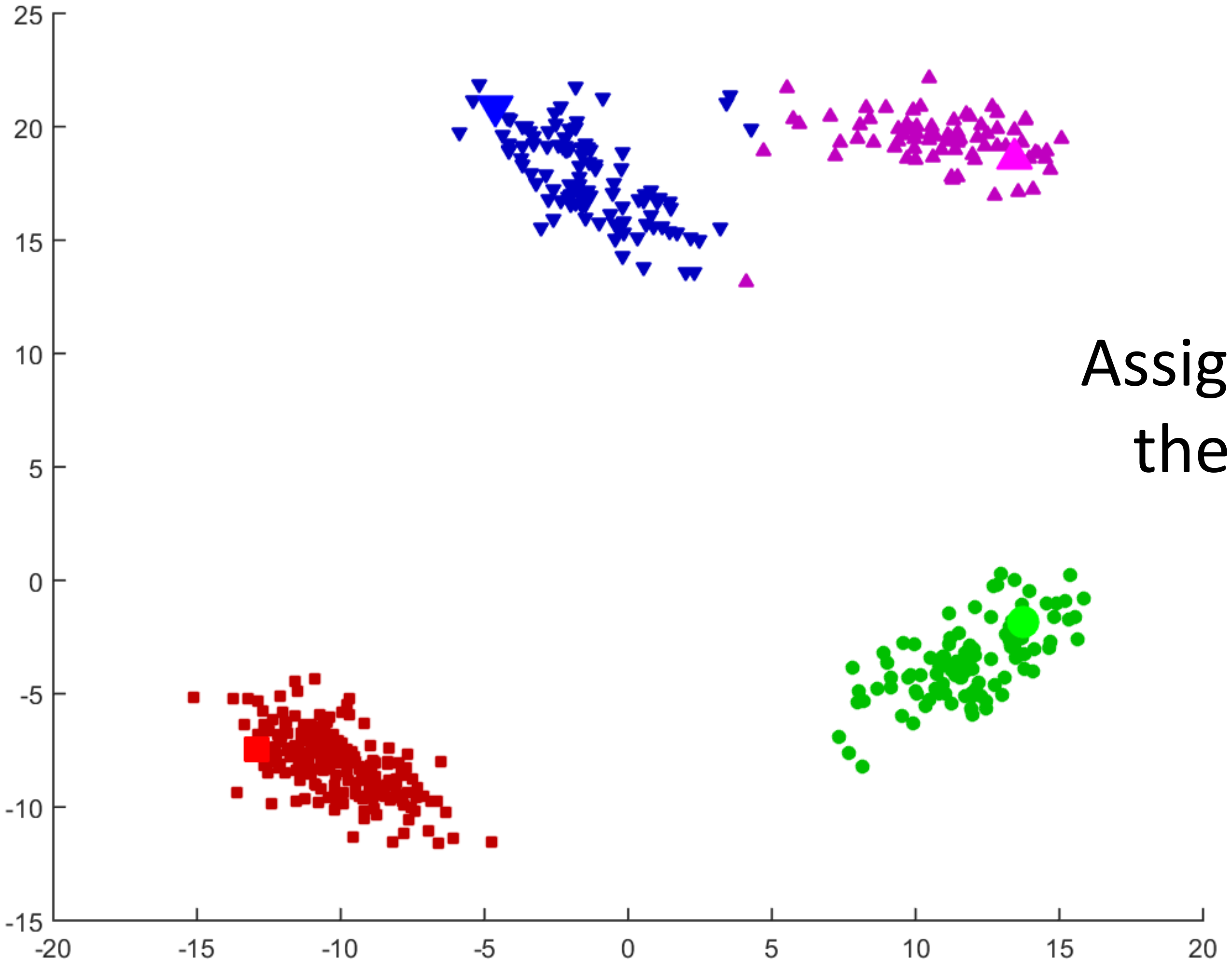
K-Means++



Sample mean proportional to distances squared.

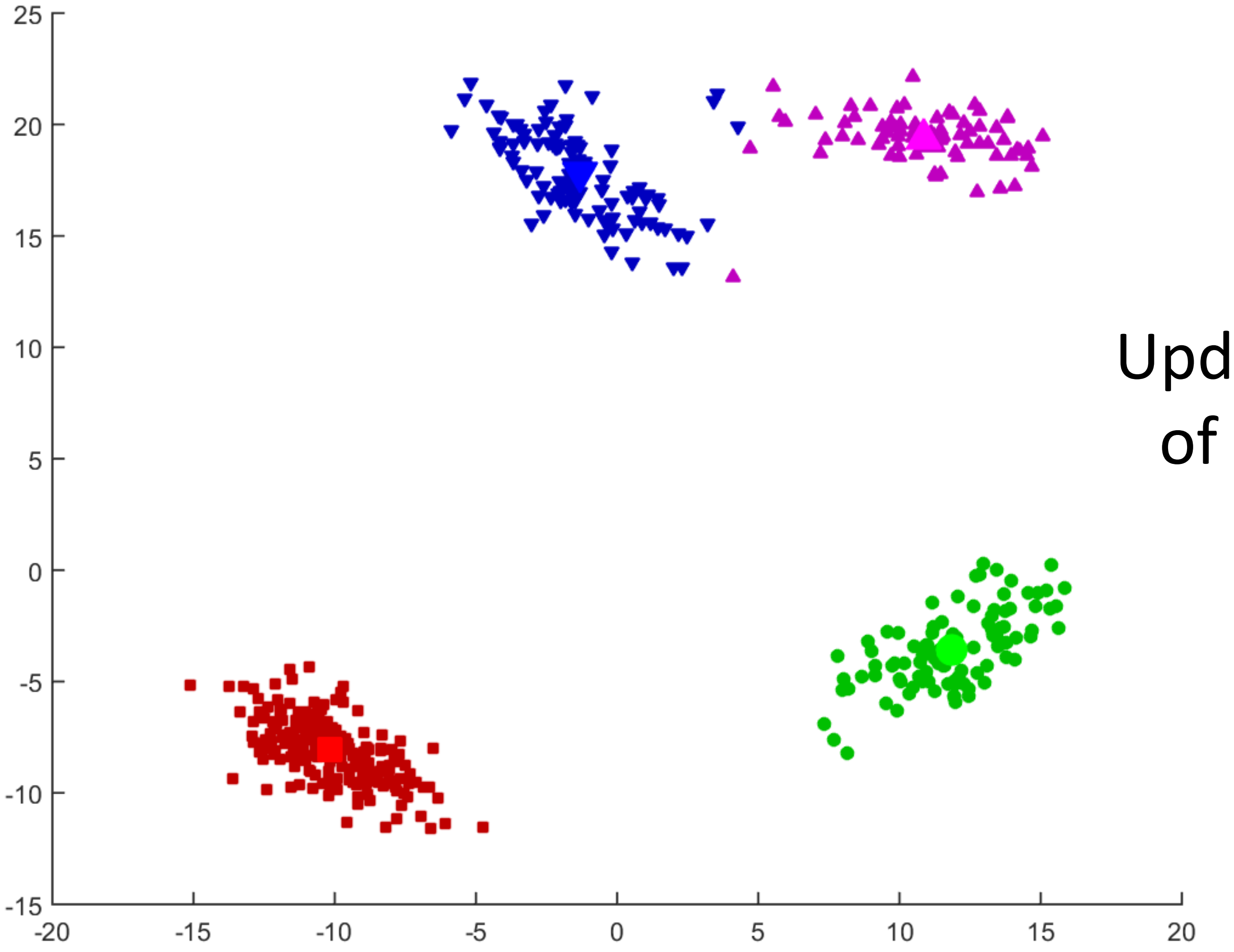
(We’ve now hit target k=4.)

K-Means++



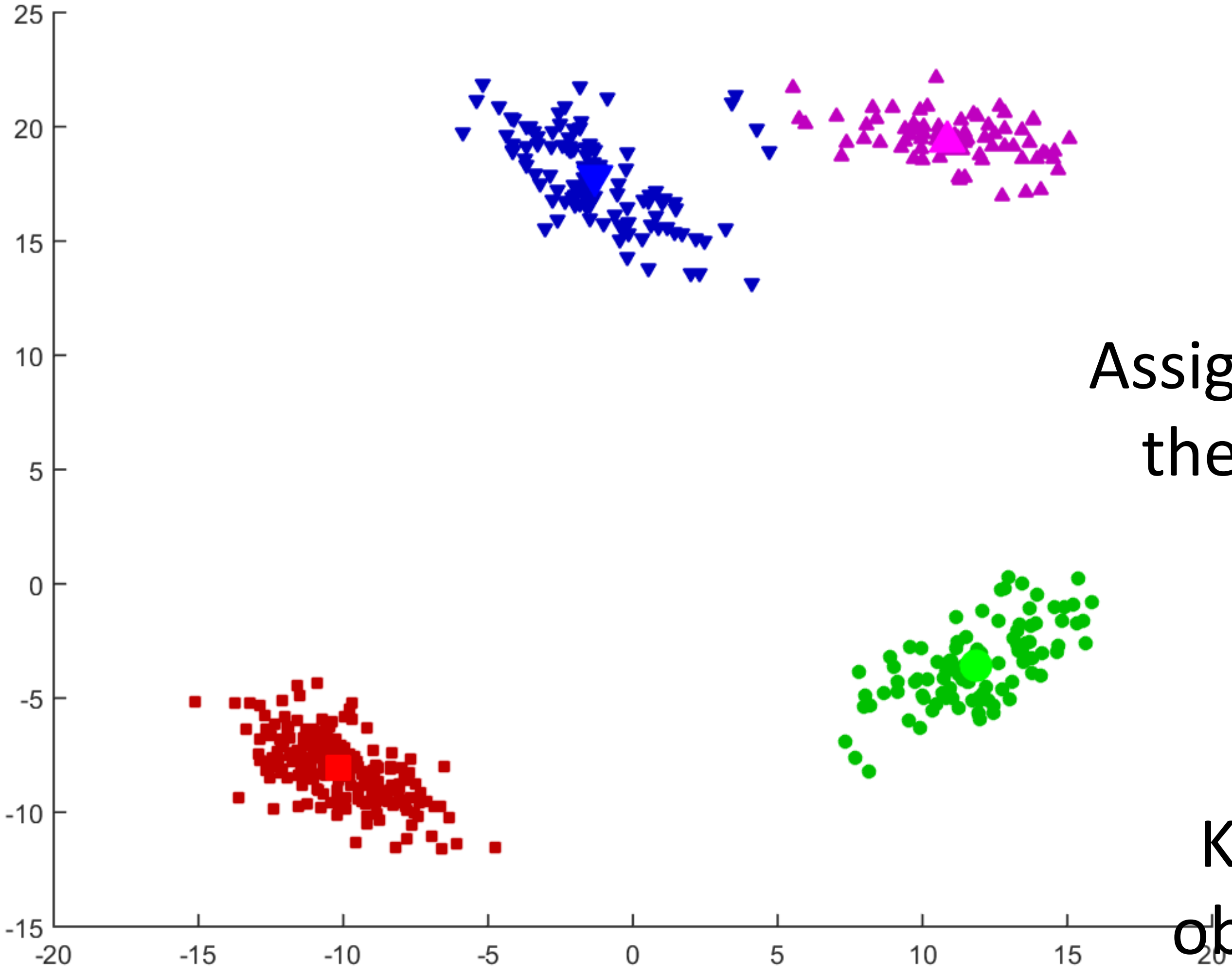
Assign each object to the closest mean.

K-Means++



Update the mean of each group.

K-Means++

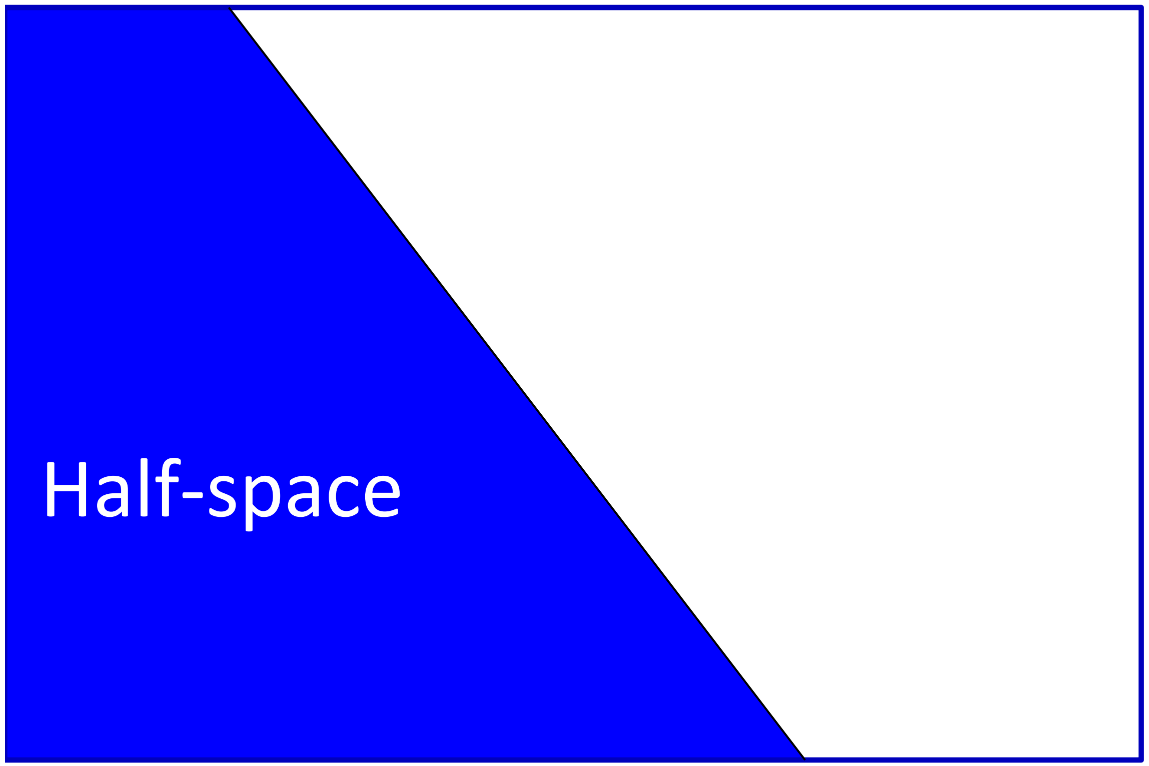


Assign each object to the closest mean.

Keep going until no o objects change groups.

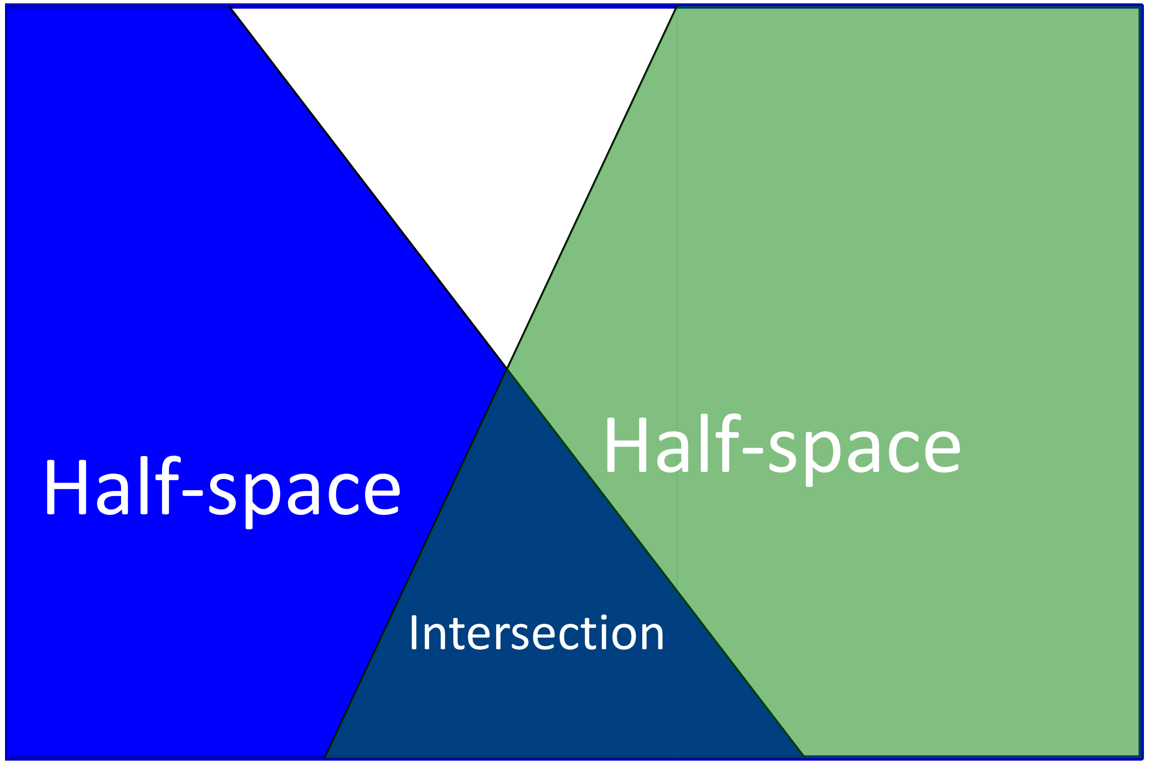
Shape of K-Means Clusters

- K-means clusters are formed by the intersection of half-spaces.

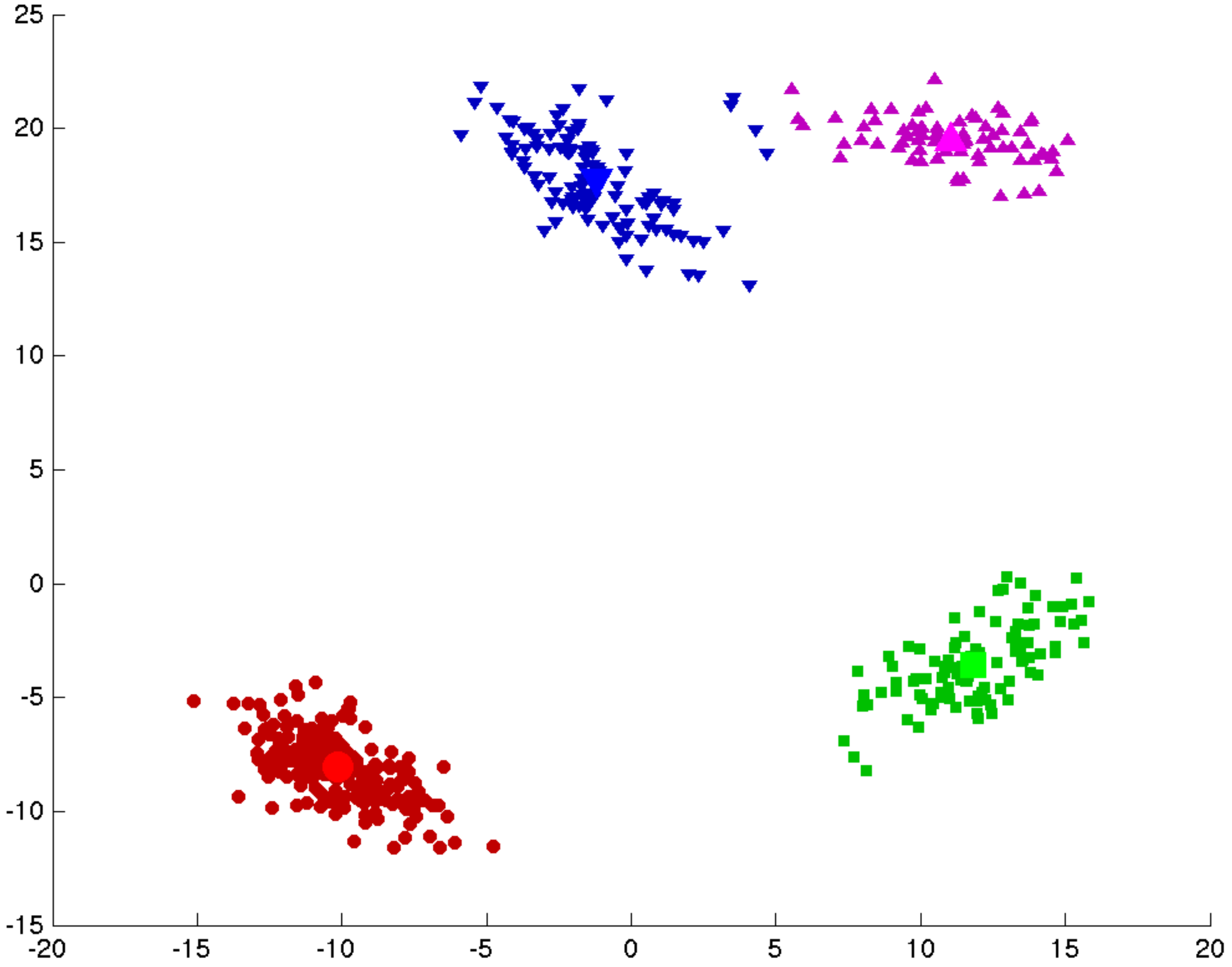


Shape of K-Means Clusters

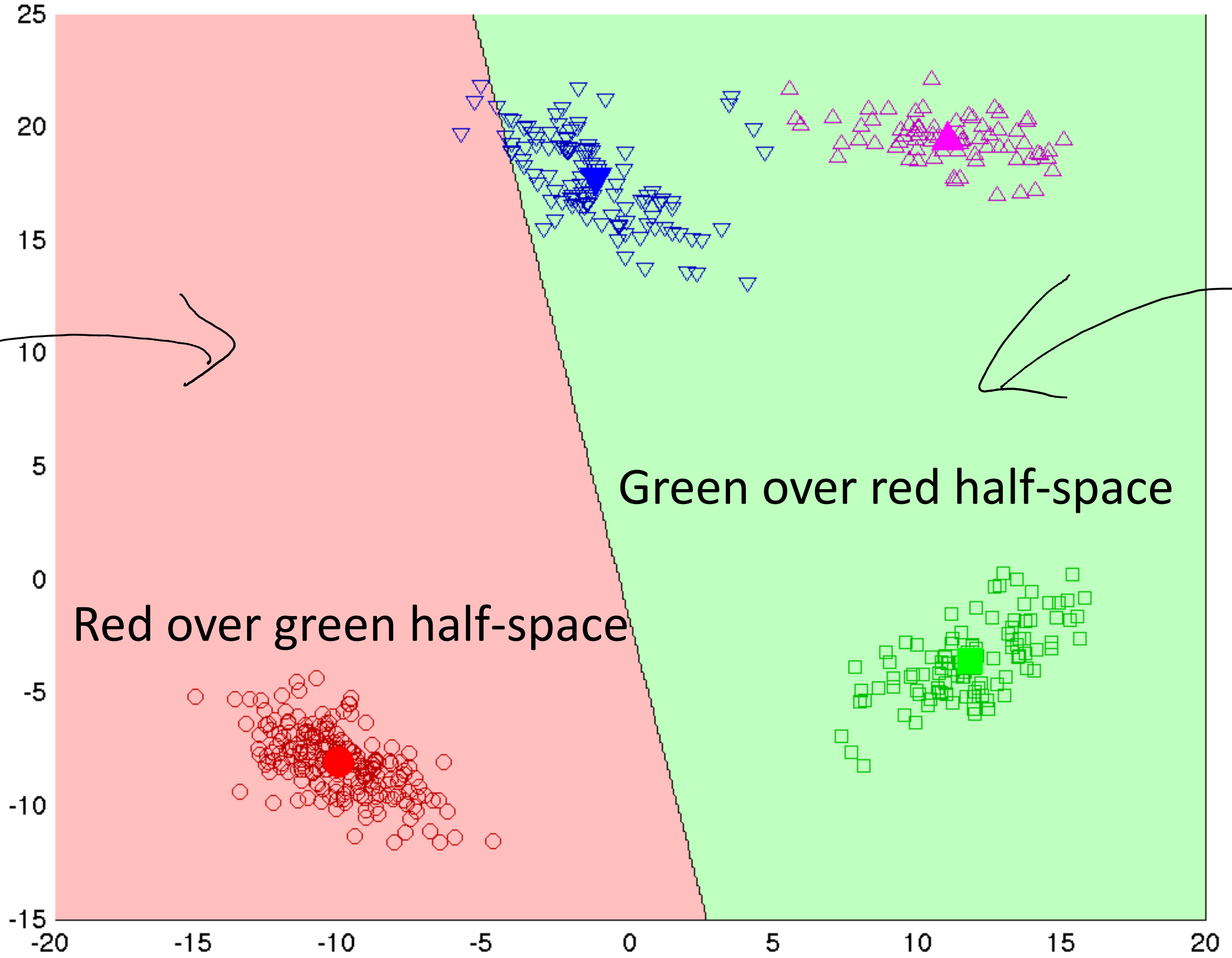
- K-means clusters are formed by the intersection of half-spaces.



Shape of K-Means Clusters



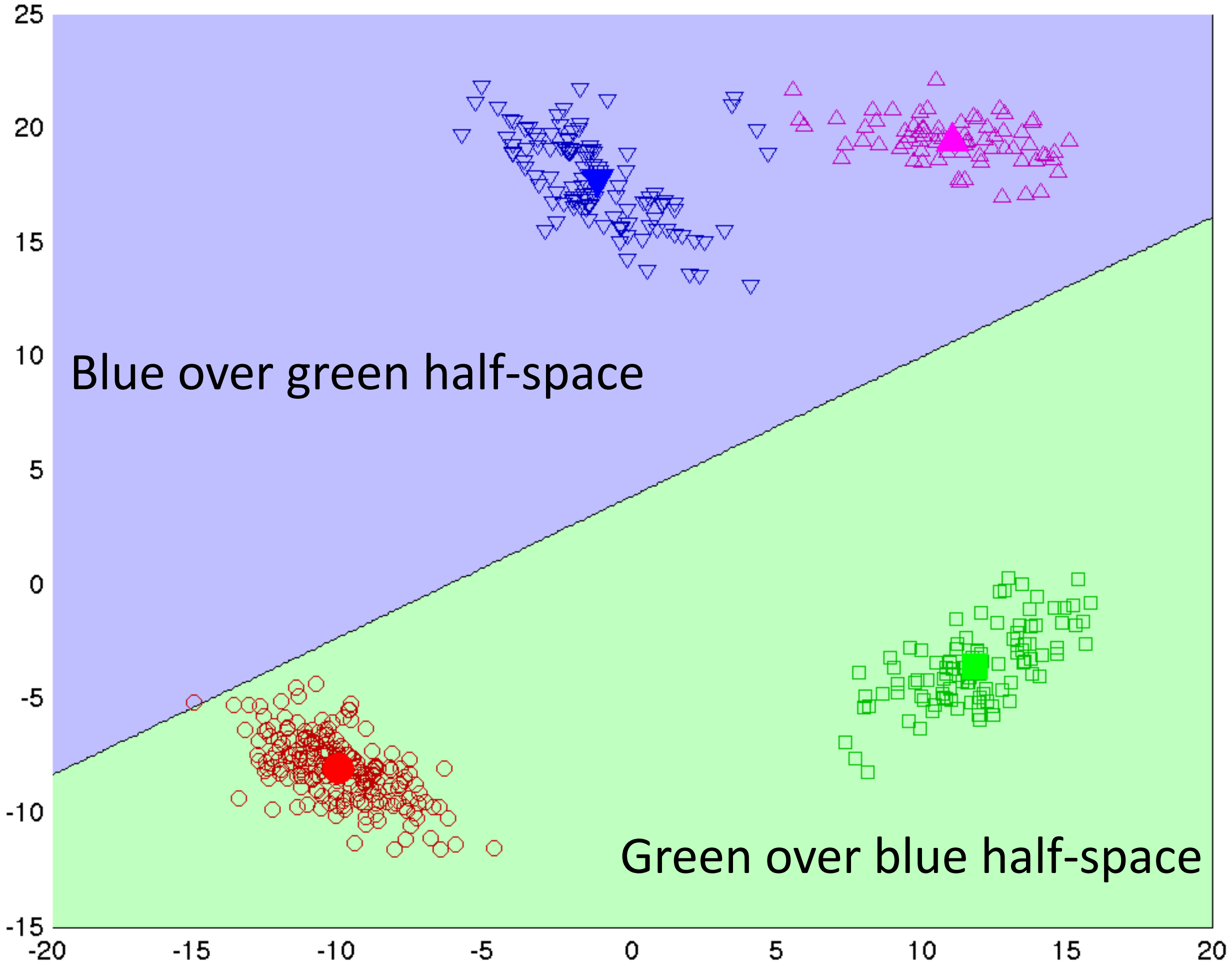
Shape of K-Means Clusters



closer to red mean

closer to green mean

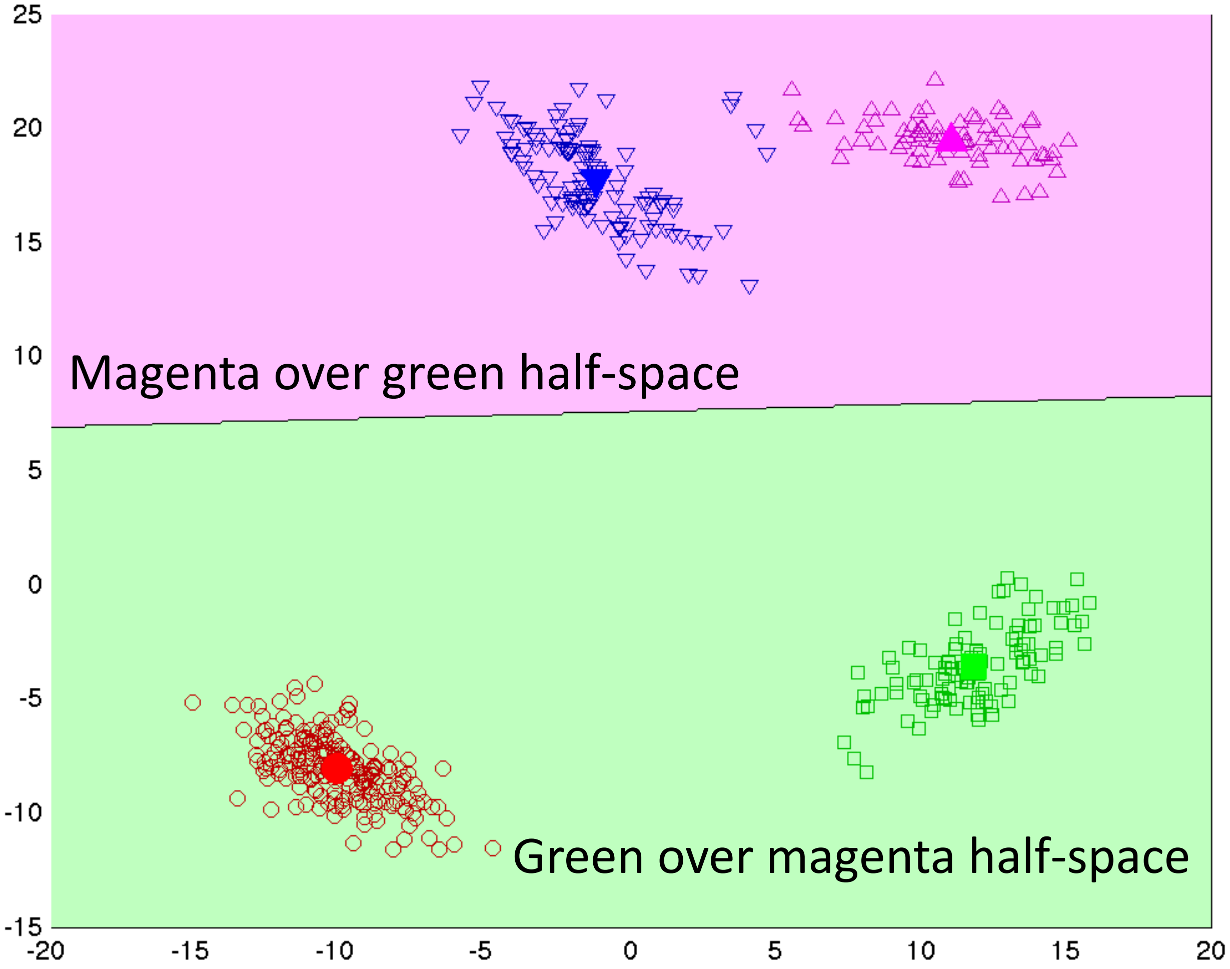
Shape of K-Means Clusters



Blue over green half-space

Green over blue half-space

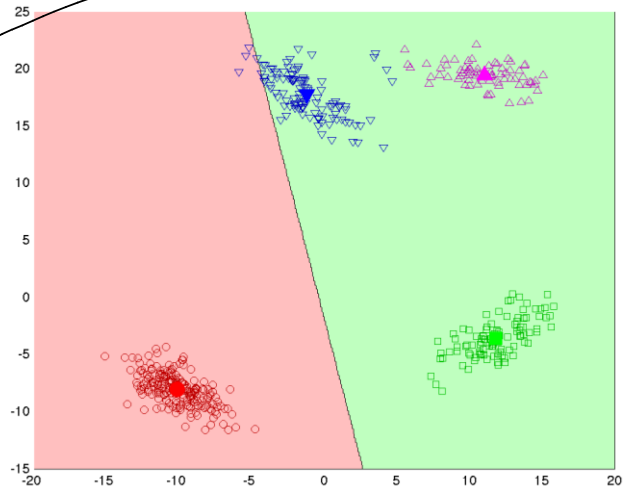
Shape of K-Means Clusters

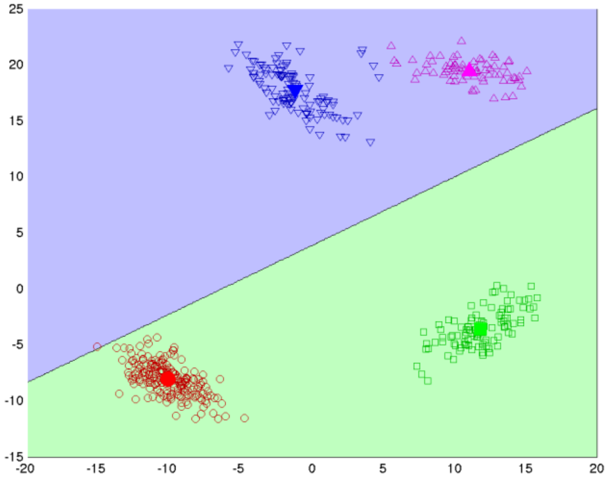


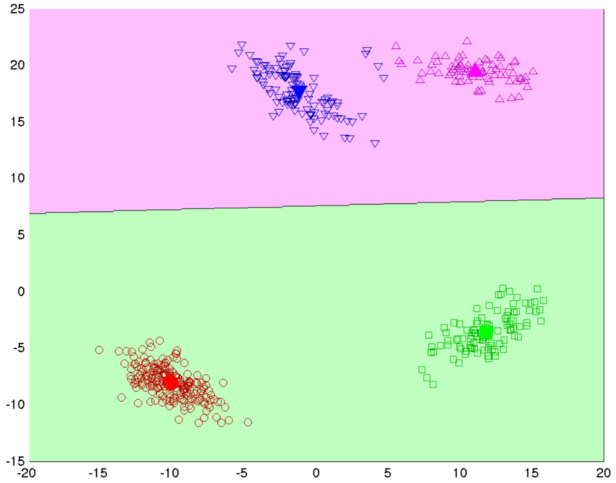
Magenta over green half-space

Green over magenta half-space

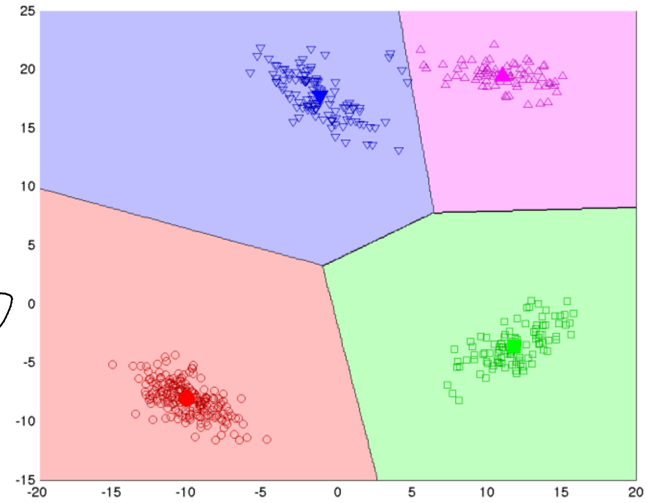
Shape of K-Means Clusters





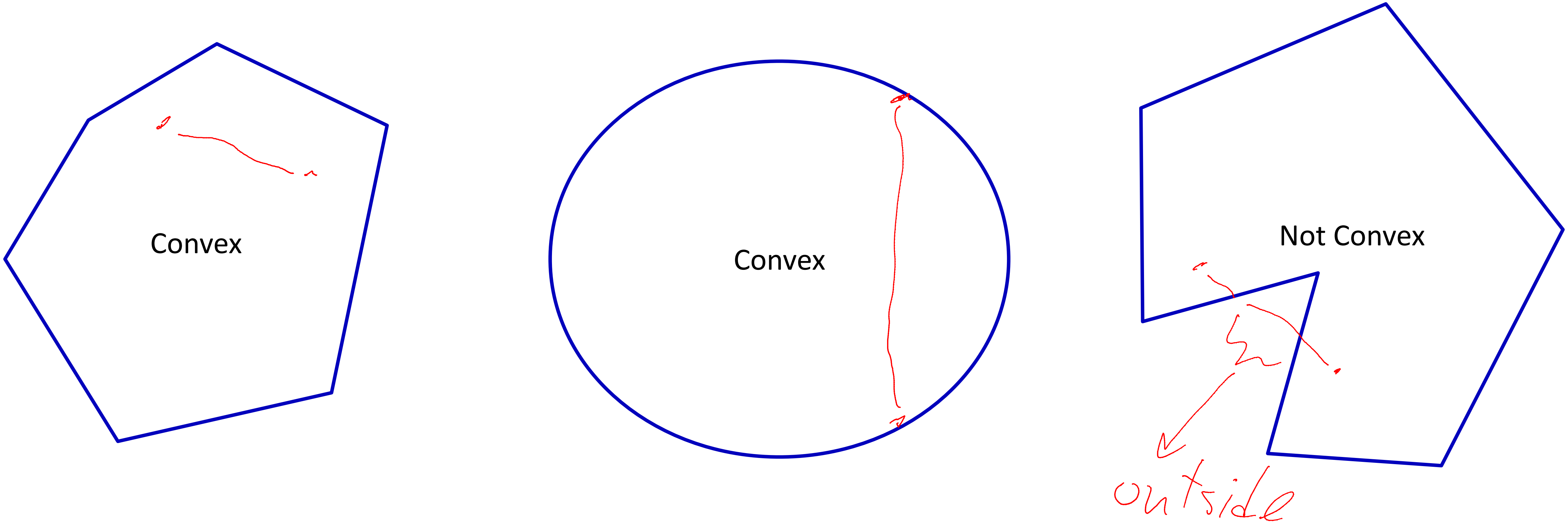


intersect
green
half-spaces

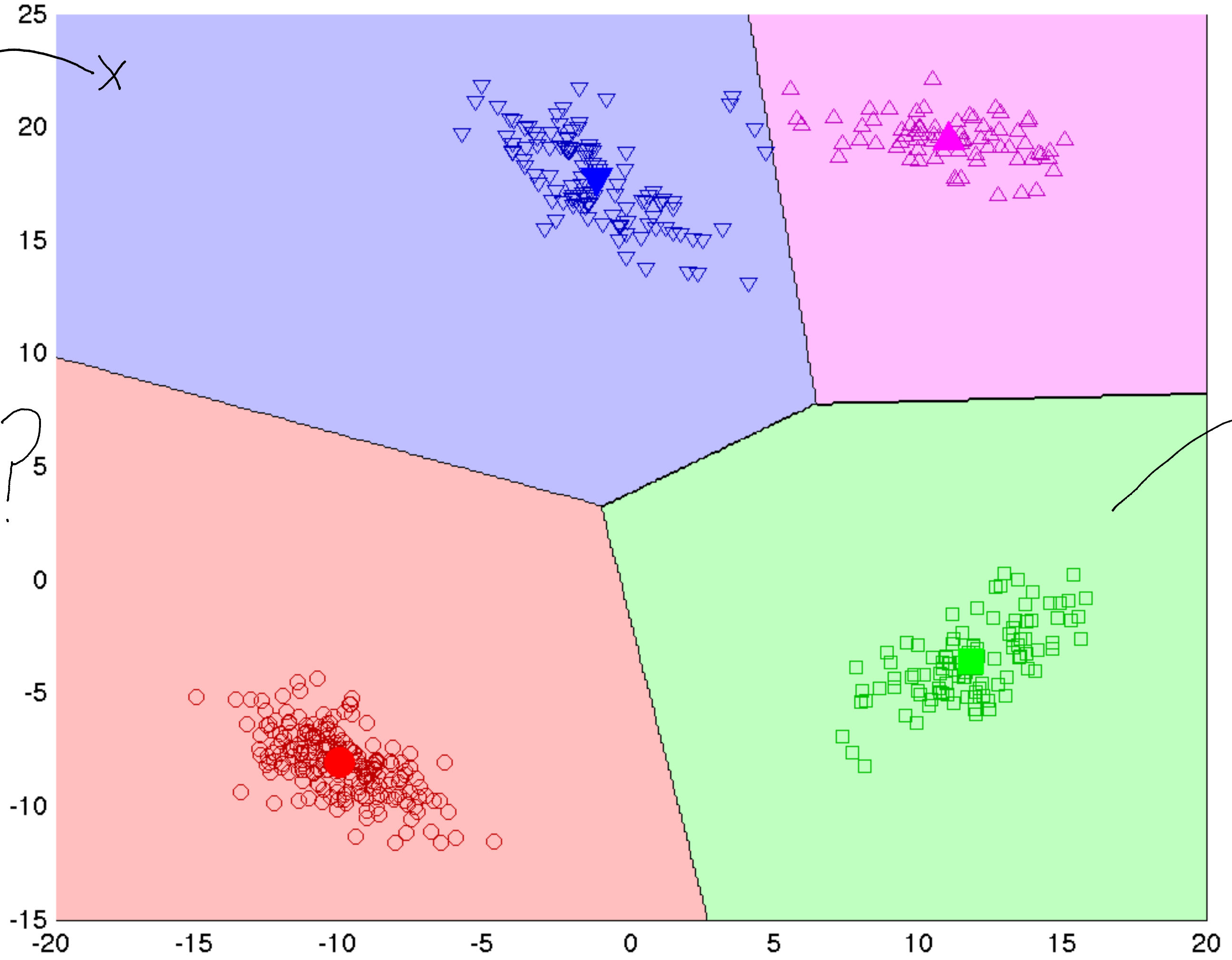


Shape of K-Means Clusters

- Intersection of half-spaces forms a convex set:
  - Line between any two points in the set stays in the set.



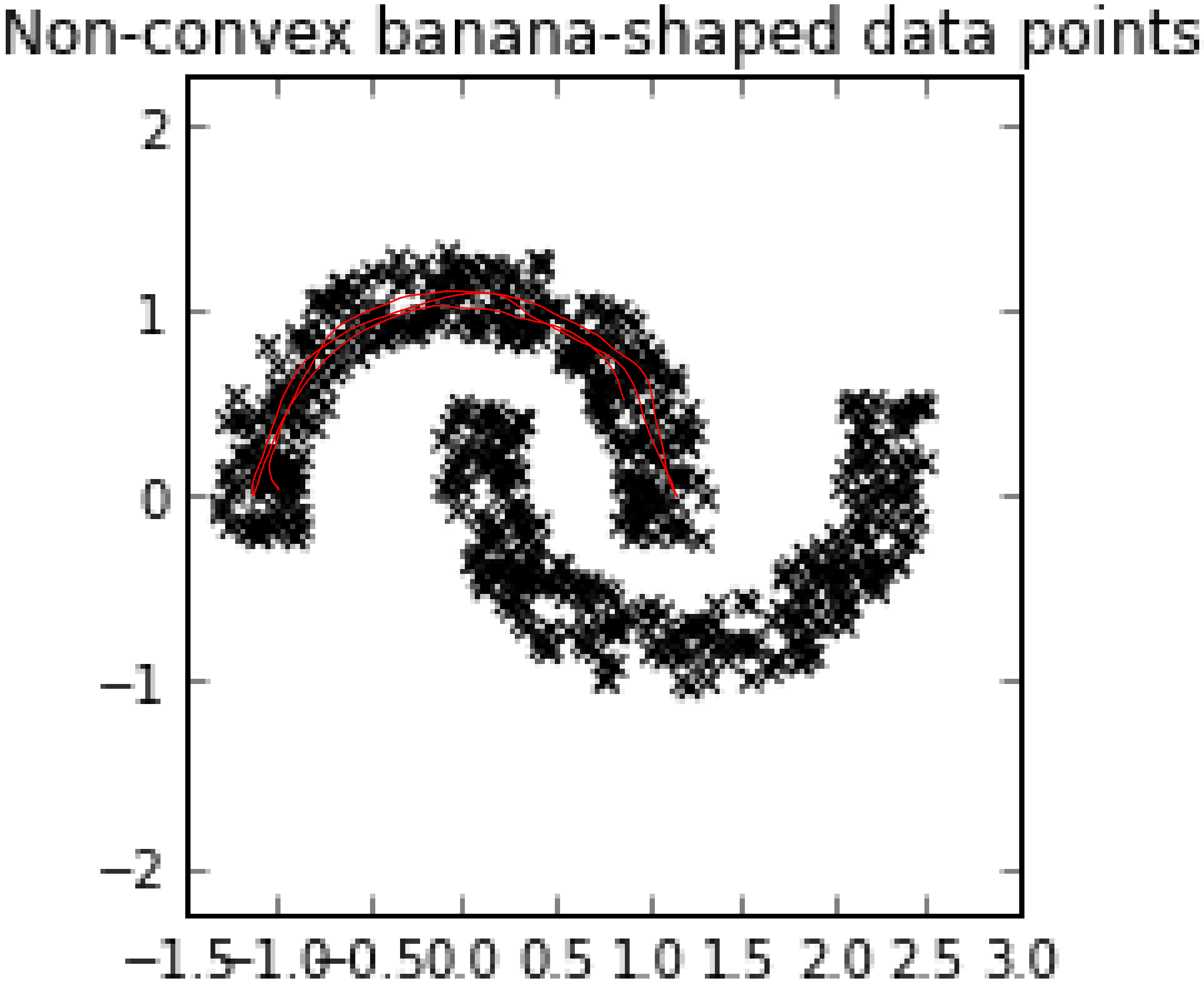
Shape of K-Means Clusters



should we really call this 'blue'?

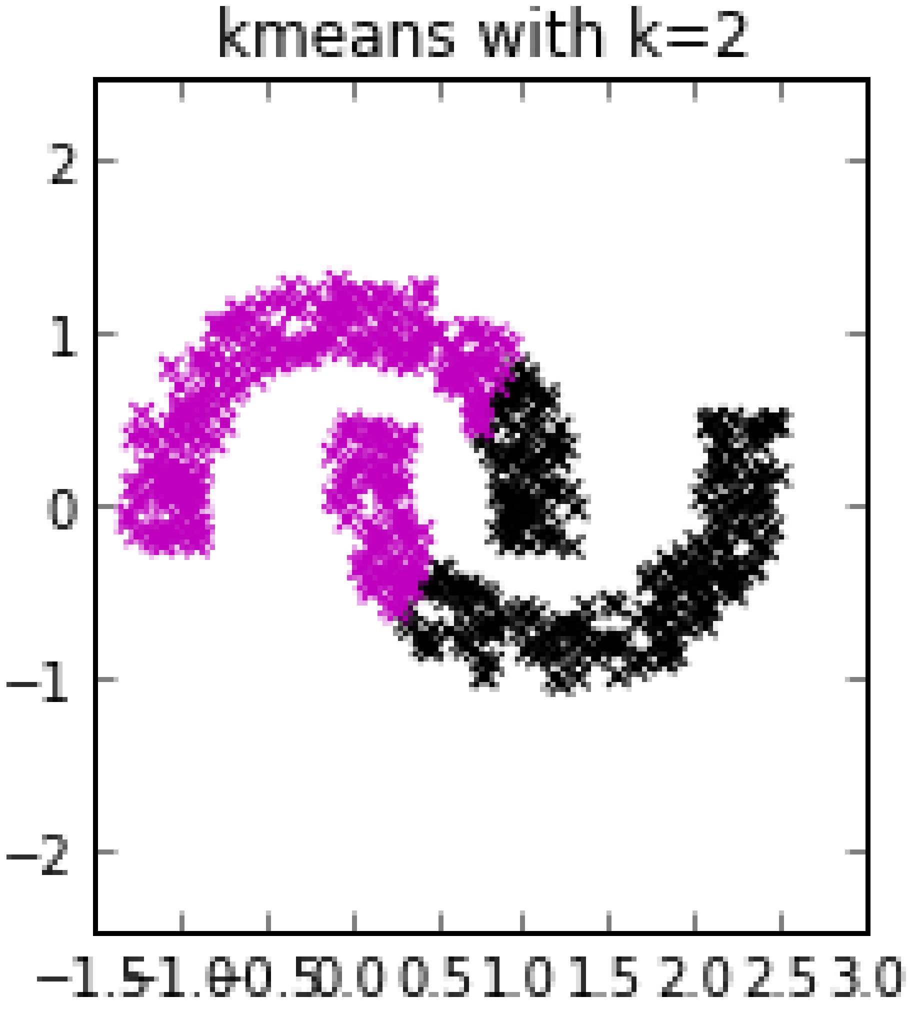
convex

K-Means with Non-Convex Clusters



https://corelifesciences.com/human-long-non-coding-rna-expression-microarray-service.html

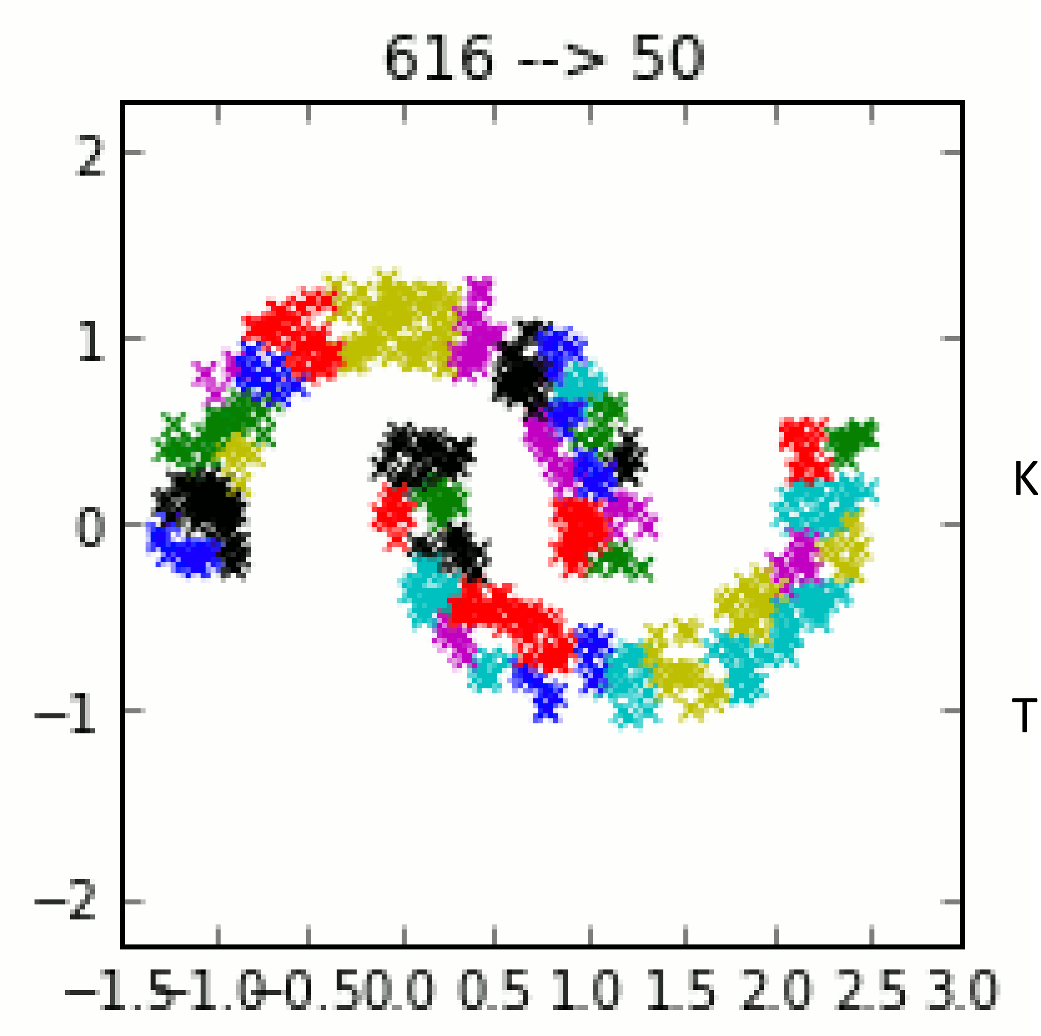
K-Means with Non-Convex Clusters



K-means cannot separate non-convex

https://corelifesciences.com/human-long-non-coding-rna-expression-microarray-service.html

K-Means with Non-Convex Clusters



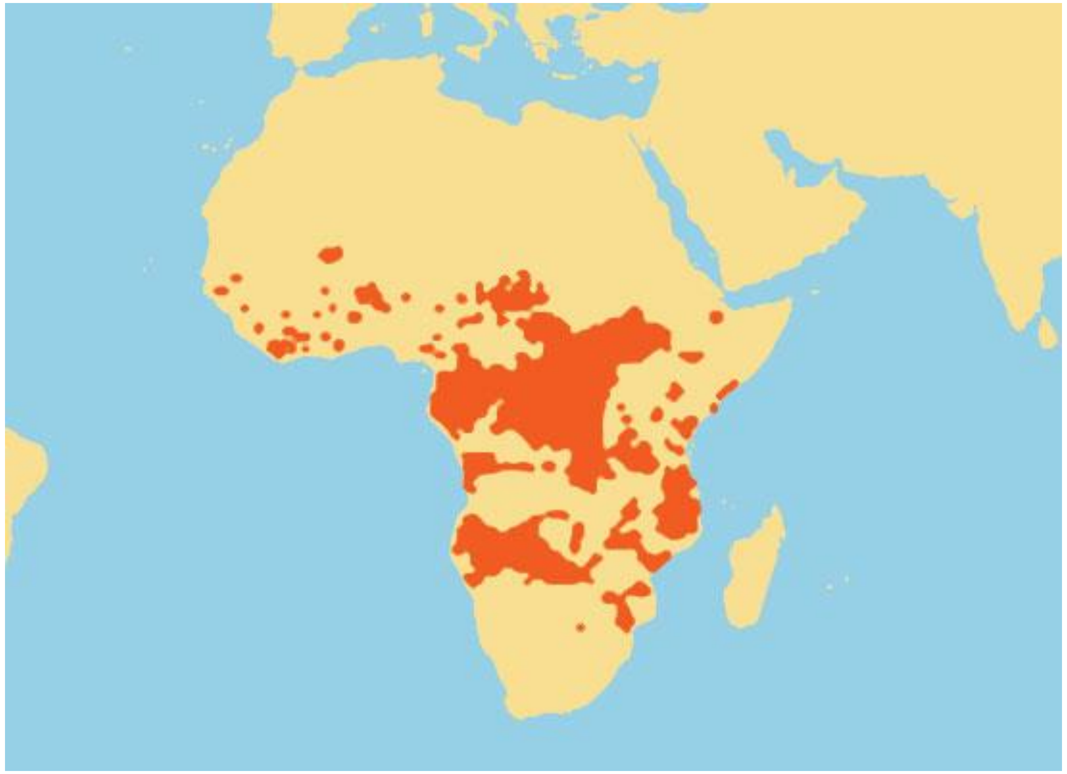
K-means cannot separate non-convex

Though over-clustering can help (next class)

https://corelifesciences.com/human-long-non-coding-rna-expression-microarray-service.html

Application: Elephant Range Map

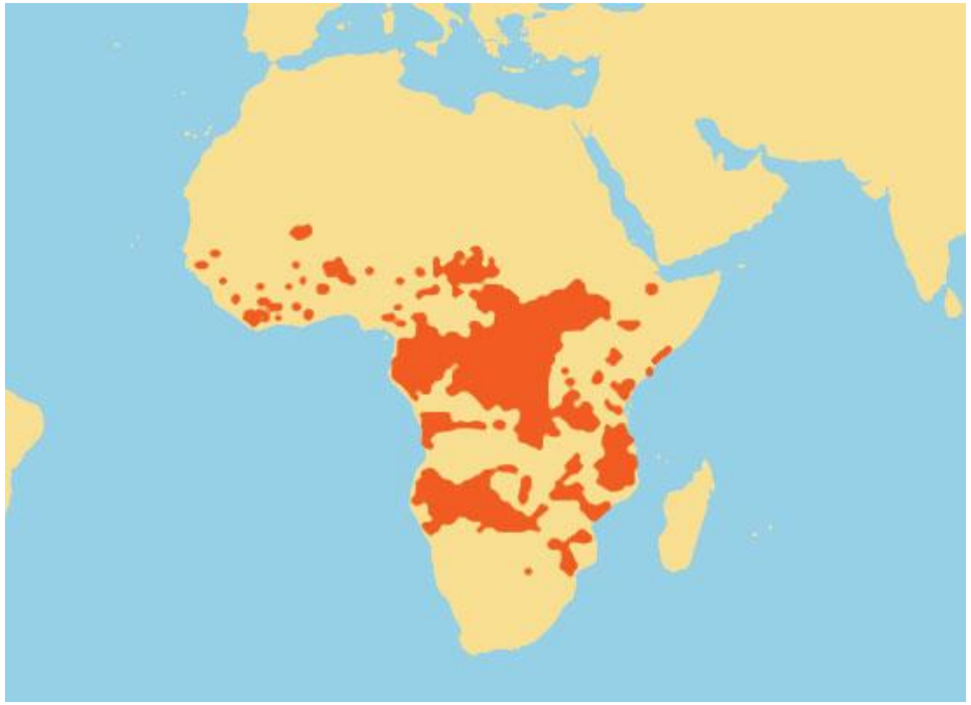
- Find habitat area of African elephants.
  - Useful for assessing/protecting population.
- Build clusters from observations of locations.
- Clusters are non-convex:
  - affected by vegetation, relief, rivers, water access.
- We do not want a partition:
  - Some regions should not have a cluster.



http://www.defenders.org/elephant/basic-facts

Motivation for Density-Based Clustering

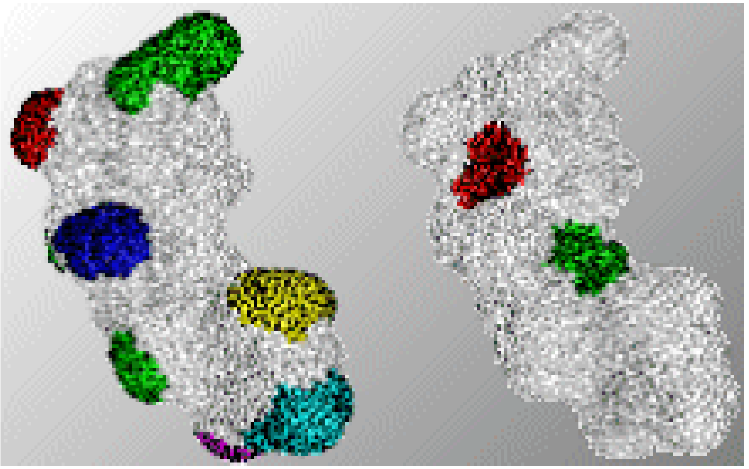
- Density-based clustering is a non-parametric clustering method:
  - Clusters are defined by connected dense regions.
    - Become more complicated the more data we have.
  - Data points in non-dense regions are not assigned a cluster.

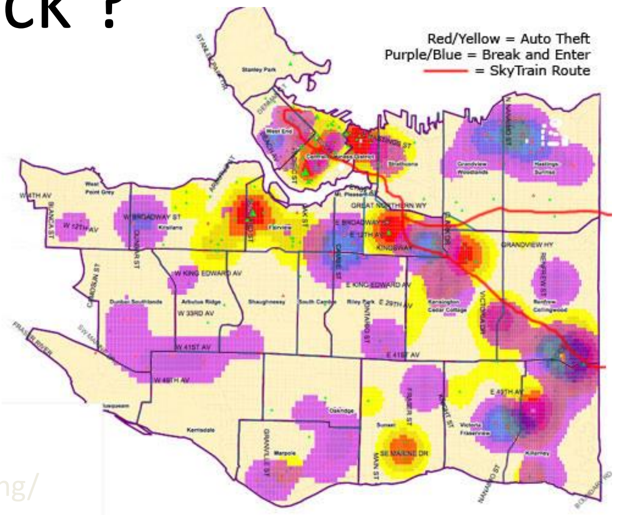


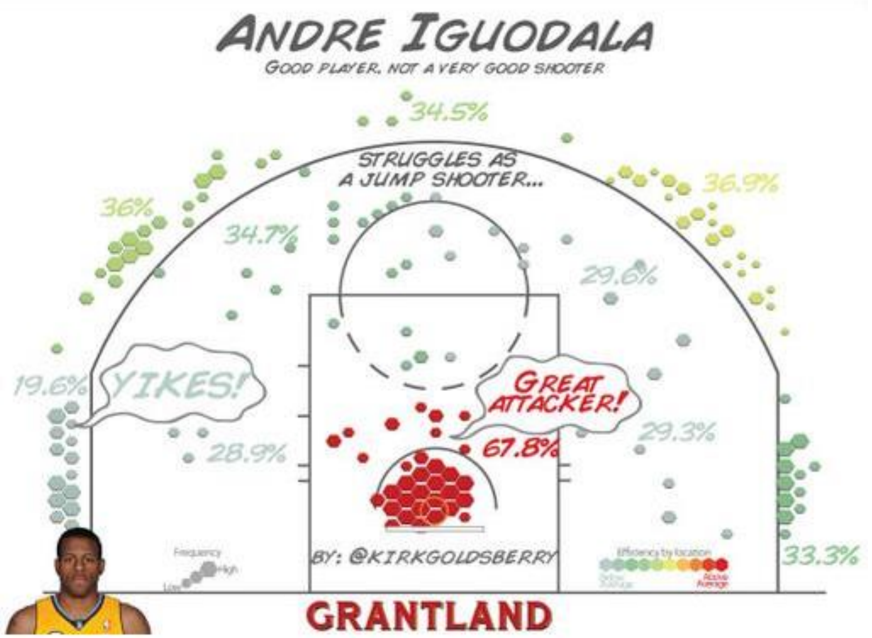
http://www.defenders.org/elephant/basic-facts

Other Potential Applications

- Where are high crime regions of a city?
- Where should taxis patrol?
- Where does Iguodala make/miss shots?
- Which products are similar to this one?
- Which pictures are in the same place?
- Where can protein ‘dock’?



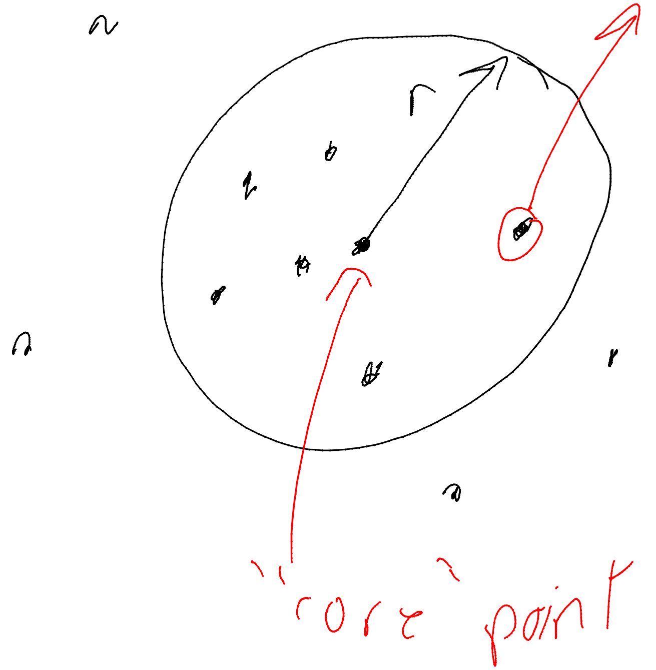




https://en.wikipedia.org/wiki/Cluster_analysis
https://www.flickr.com/photos/dbarefoot/420194128/
http://letsgowarriors.com/replacing-jarrett-jack/2013/10/04/
http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/

Density-Based Clustering

- Density-based clustering algorithm (DBSCAN) has two parameters:
  - Radius: minimum distance between points to be considered 'close'.
  - MinPoints: number of 'close' points needed to define a cluster.

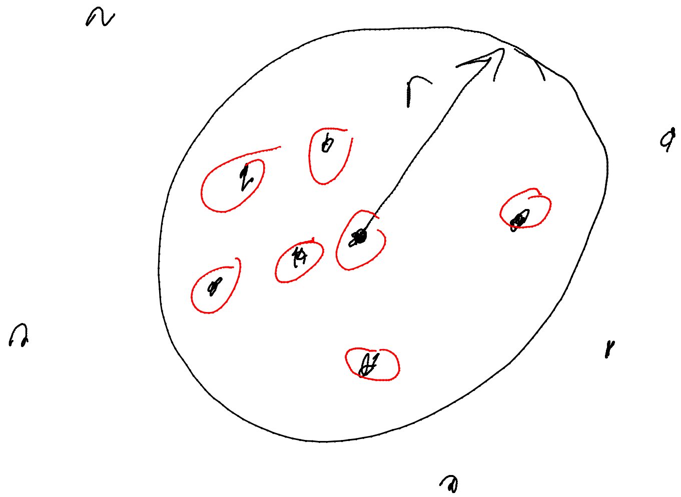


6 points within radius 'r' of center point.

If minPoints ≤ 6, this is called a "core" point of cluster.

Points within radius 'r' of core points are "reachable."

Density-Based Clustering



6 points within radius 'r' of center point.

If minPoints ≤ 6, this is called a "core" point of cluster.

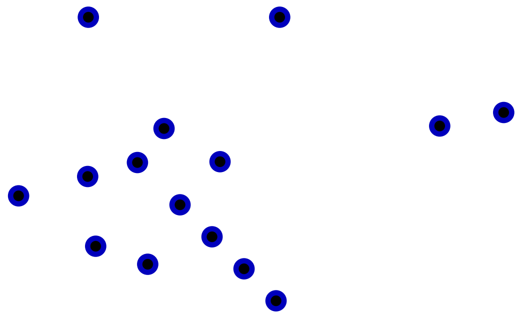
Points within radius 'r' of core points are "reachable."

If core points are reachable from each other, merge clusters.

→ Final cluster is core points, and points reachable from core points.

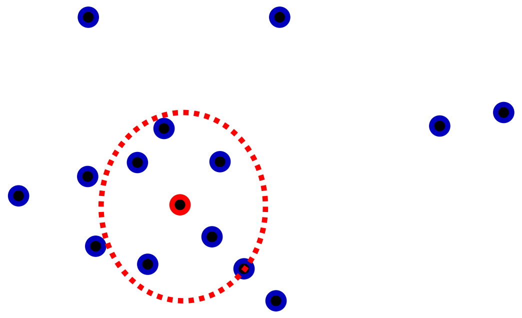
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



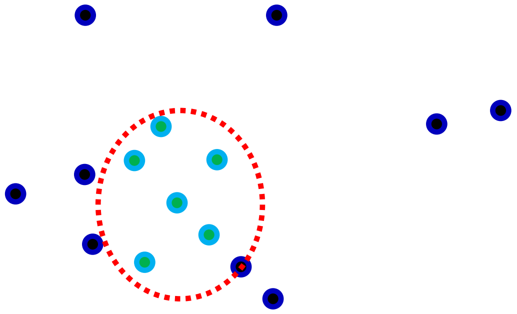
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



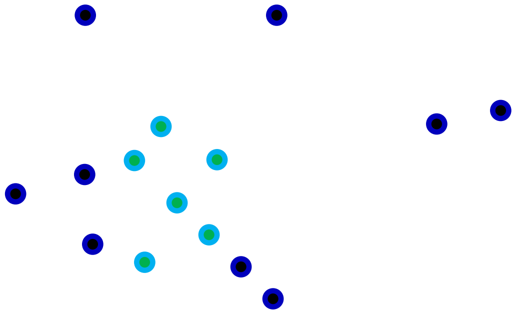
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



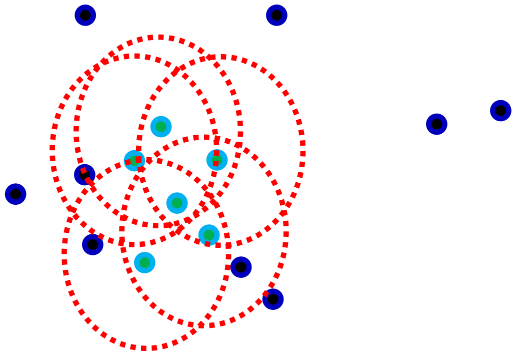
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



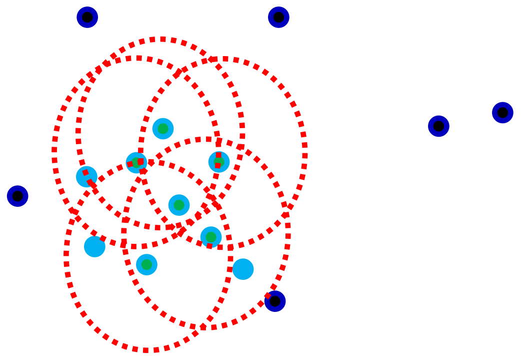
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ 'r'), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance 'r' of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



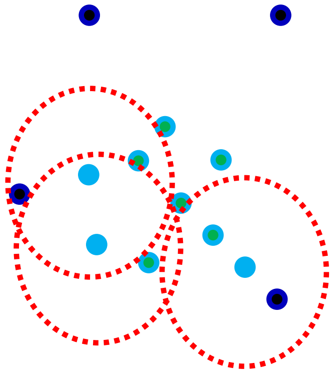
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



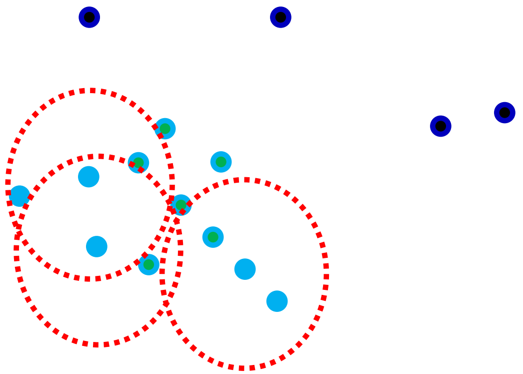
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



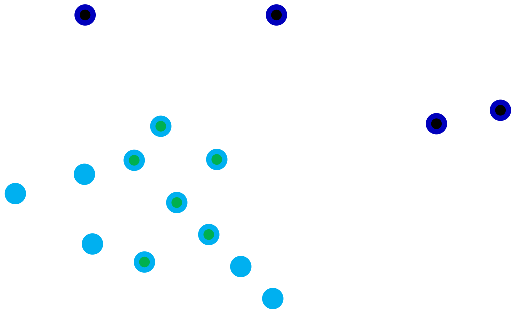
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



Density-Based Clustering

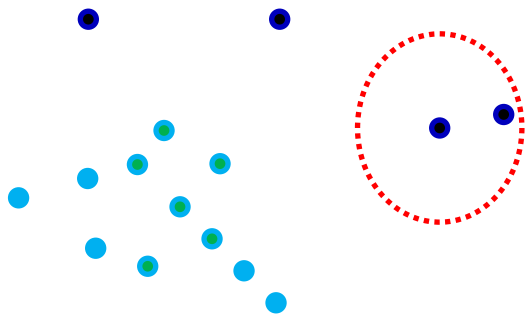
- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ 'r'), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance 'r' of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



boundary point for both clusters.

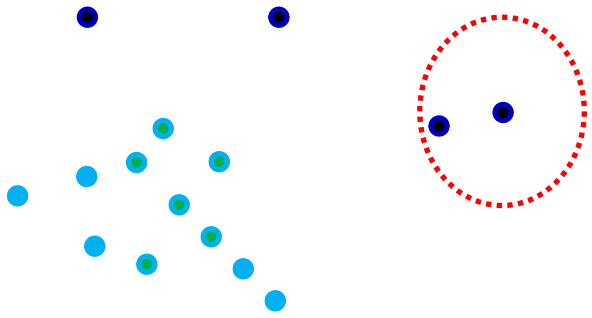
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



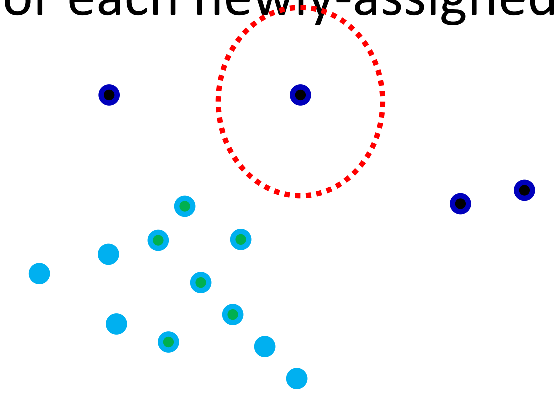
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



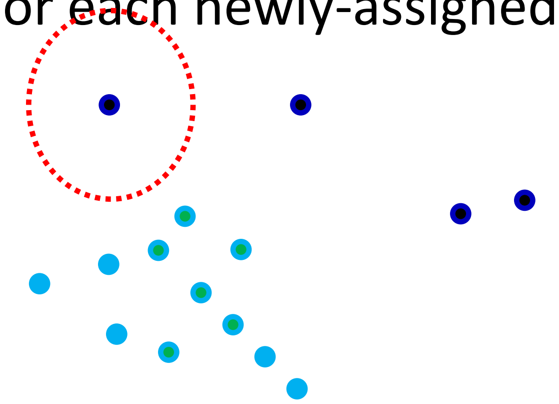
Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.

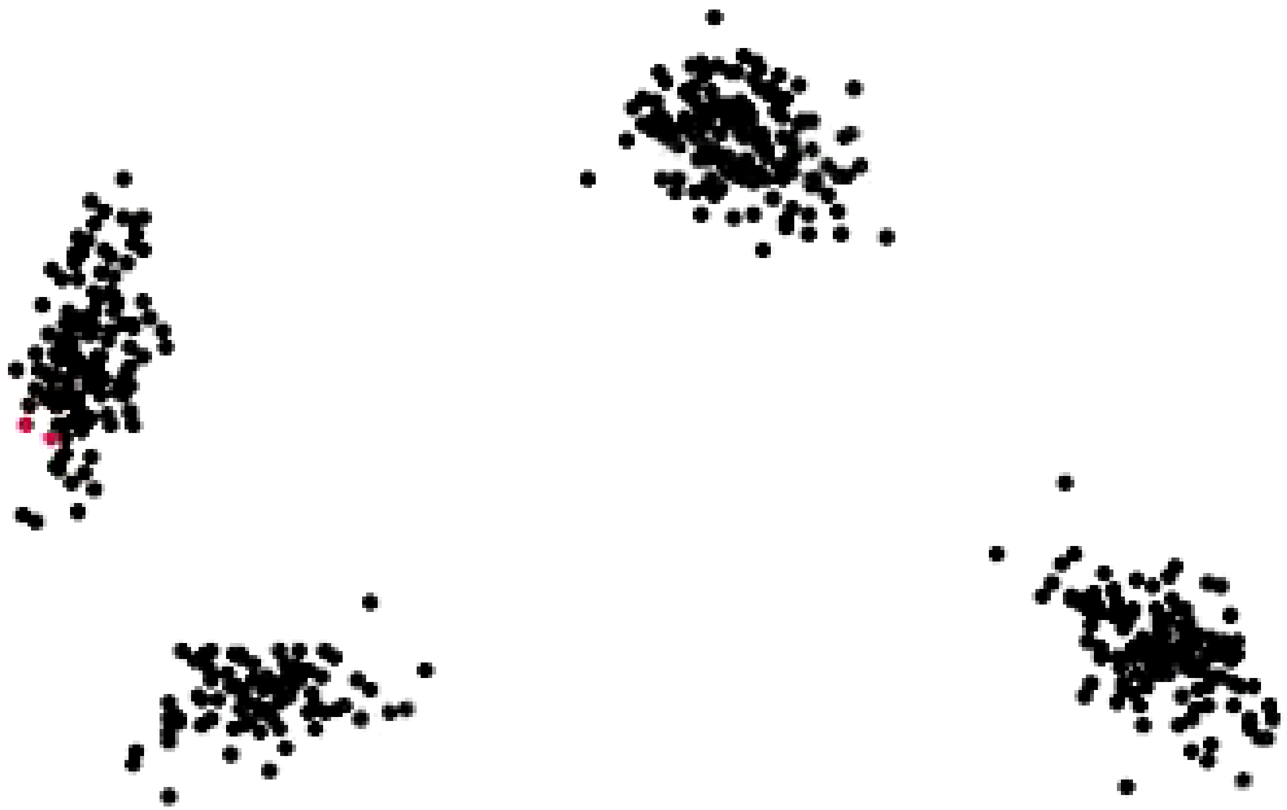


Density-Based Clustering

- Pseudocode for DBSCAN:
  - For each example x_i:
    - If x_i is already assigned to a cluster, do nothing.
    - If x_i is not core point (less than minPoints neighbours with distance ≤ ‘r’), do nothing.
    - If x_i is a core point, expand cluster.
  - Expand cluster function:
    - Assign all x_j within distance ‘r’ of core point x_i to cluster.
    - For each newly-assigned neighbour x_j that is a core point, expand cluster.



Density-Based Clustering



Density-Based Clustering Issues

- Some points are not assigned to a cluster.
  - Good or bad, depending on the application.
- Sensitive to the choice of radius and minPoints.
- Ambiguity of ‘non-core’ (boundary) points:
  - They could be assigned more than once.
- Other than this ambiguity, not sensitive to initialization.
- Assigning new points to clusters is expensive.
- In high-dimensions, need a lot of points to ‘fill’ the space.

Summary

1. K-means++: randomized initialization with good expected performance.
2. Shape of K-means clusters: intersection of half-spaces => convex sets.
3. Density-based clustering: useful for finding non-convex connected clusters.
4. DBSCAN algorithm: assign points in dense regions to same cluster.

- Next time:
  - Dealing with clusters of different densities.