Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program
605-447/625-438
Dr. Mark Fleischer

Module 13.1: Clustering

In This Module We Will Cover:

- Clustering:
  - The nature of the problem
  - Defining it
  - Modeling it as an optimization problem
- Radial Basis Functions
  - Their connection to clustering

What is the Clustering Problem

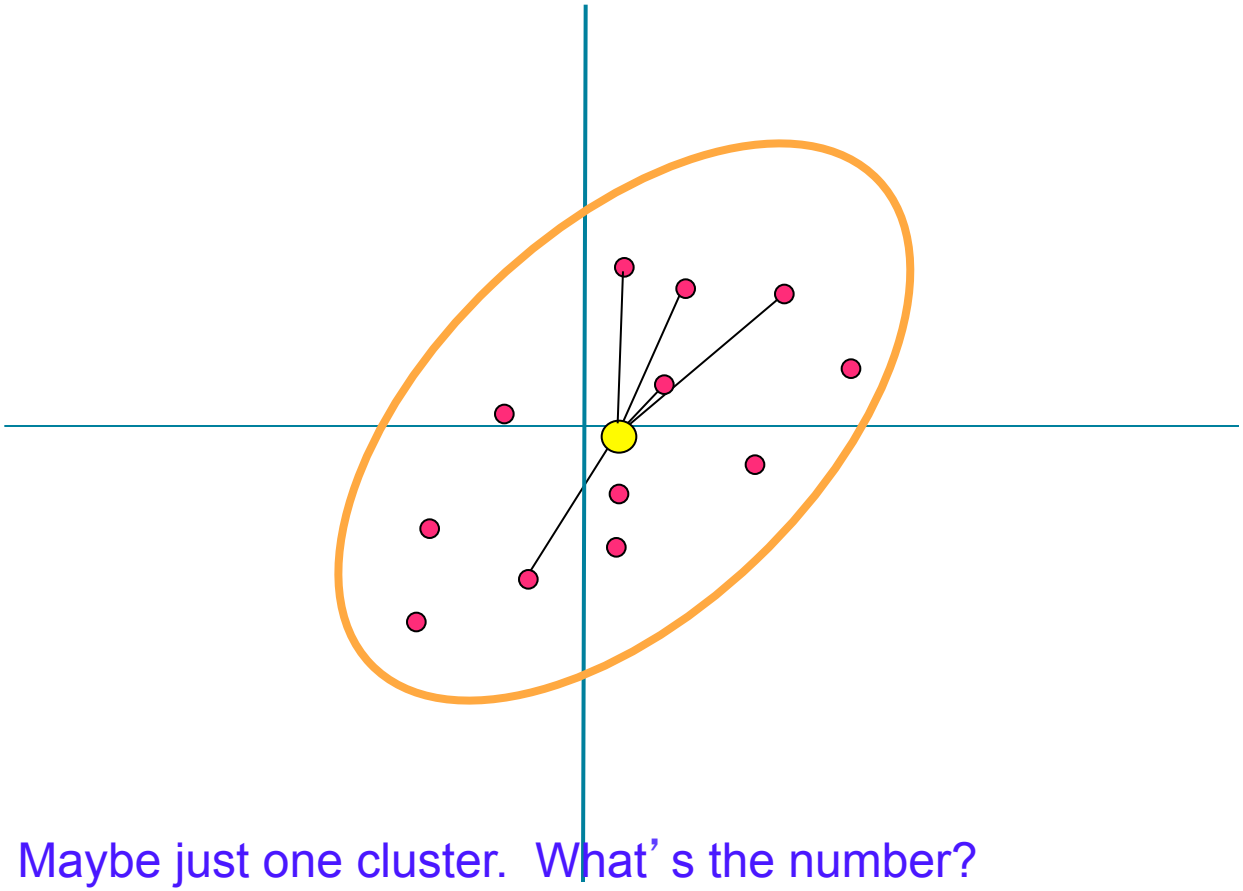
- What is a cluster?
  - One definition: A set of points that are within some defined distance of a 'centroid'.
- Questions:
  - Which 'cluster' does a point belong? How can we classify a point to be a member of a 'cluster'?
  - How many clusters should there be?

Cluster Definitions

- Many possible ways to define a cluster.
- Can require that a cluster ‘centroid’ be an element of the population.
- Or, we can require it to be defined by some function.

How Many Clusters?

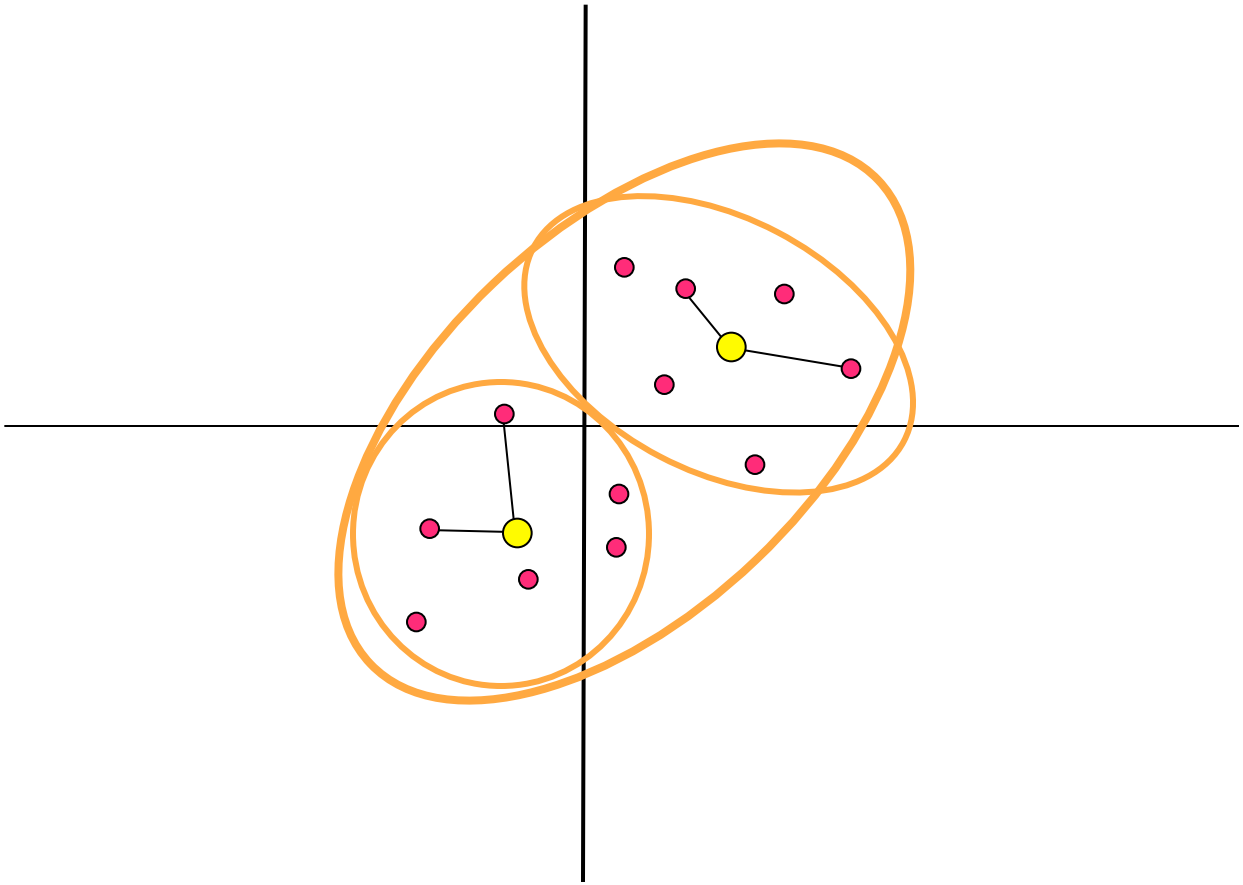
There are 12 points. What are the clusters?



Maybe just one cluster. What's the number?

How Many Clusters?

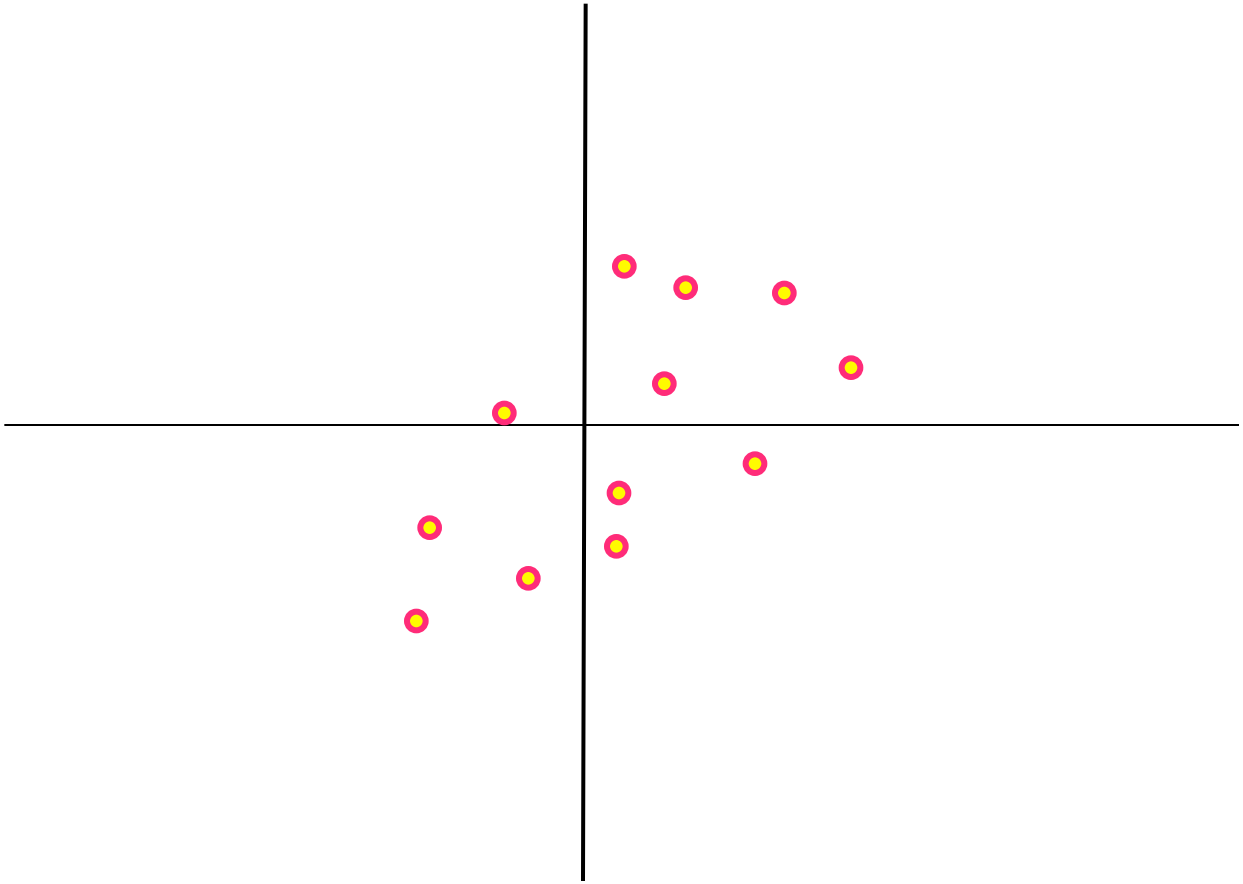
There are 12 points. What are the clusters?



We can add more clusters. How many?

How Many Clusters?

There are 12 points. What are the clusters?



The Problem with Clusters

- Points can appear to have ‘natural’ clustering
  - Not a very scientific or objective way of describing or determining clusters
  - How can we objectively determine the number of cluster centers?
- We need some metric associated with the number of clusters and where the cluster centers are located

Posing an Optimization Problem

- Establishes objective criteria for how to define clusters
- Could capture issues associated with costs and benefits of clusters
- Main idea: A point should be closest to 'its' cluster center
- What is the tradeoff?

The Optimization Problem

The Tradeoff:

- If there are more cluster centers, a point will tend to be closer to its center
- More centers, less distance
- Fewer centers, greater distance
- Our objective function should entail both elements:
  - 1) the number of cluster centers;
  - 2) the ‘distance’ of each cluster member to its center.

The Optimization Model

- Definitions:
  - Let P be the set of all points.
  - Let C be the set of all cluster centers
  - A point p is a member of cluster c_i if its distance to the centroid (or center) of c_i is less than the distance to any other cluster center.
  - The distance metric associated with cluster c_i is some function of the distance from the center of c_i to all of its members

The Optimization Model

A set of points in some space: $P = \{p_1, p_2, \ldots, p_n\}$

A set of clusters/centers: $C = \{c_1, c_2, \ldots, c_M\}$

Membership criteria: $p' \in c_j \text{ iff } \rho(p', c_j) < \rho(p', c_i) \; \forall \, i \neq j$

The Optimization Model

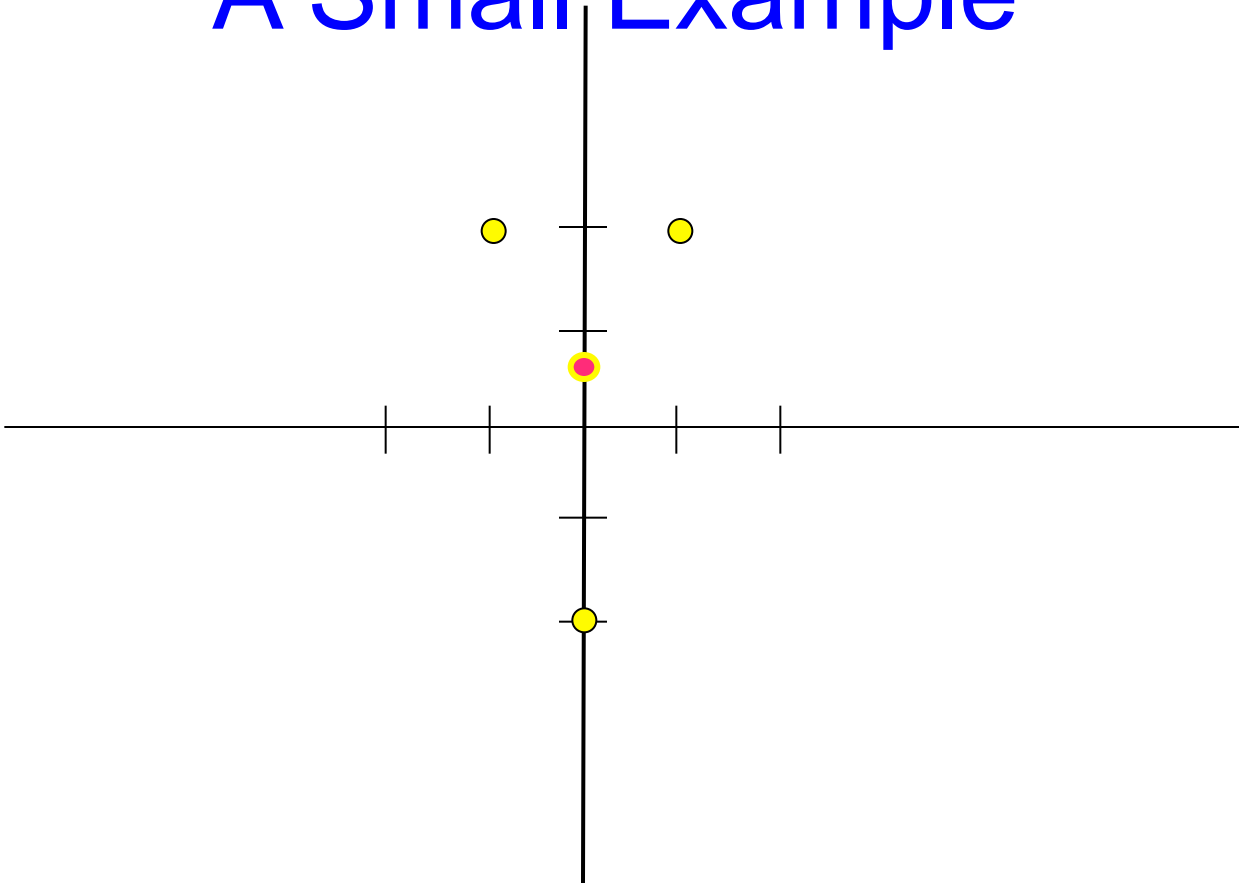
The number of cluster centers = $|C|$

A distance metric for each cluster center:

$$D_{c_j} = \sum_{p_i \in c_j} \rho\left(p_i, c_j\right)$$

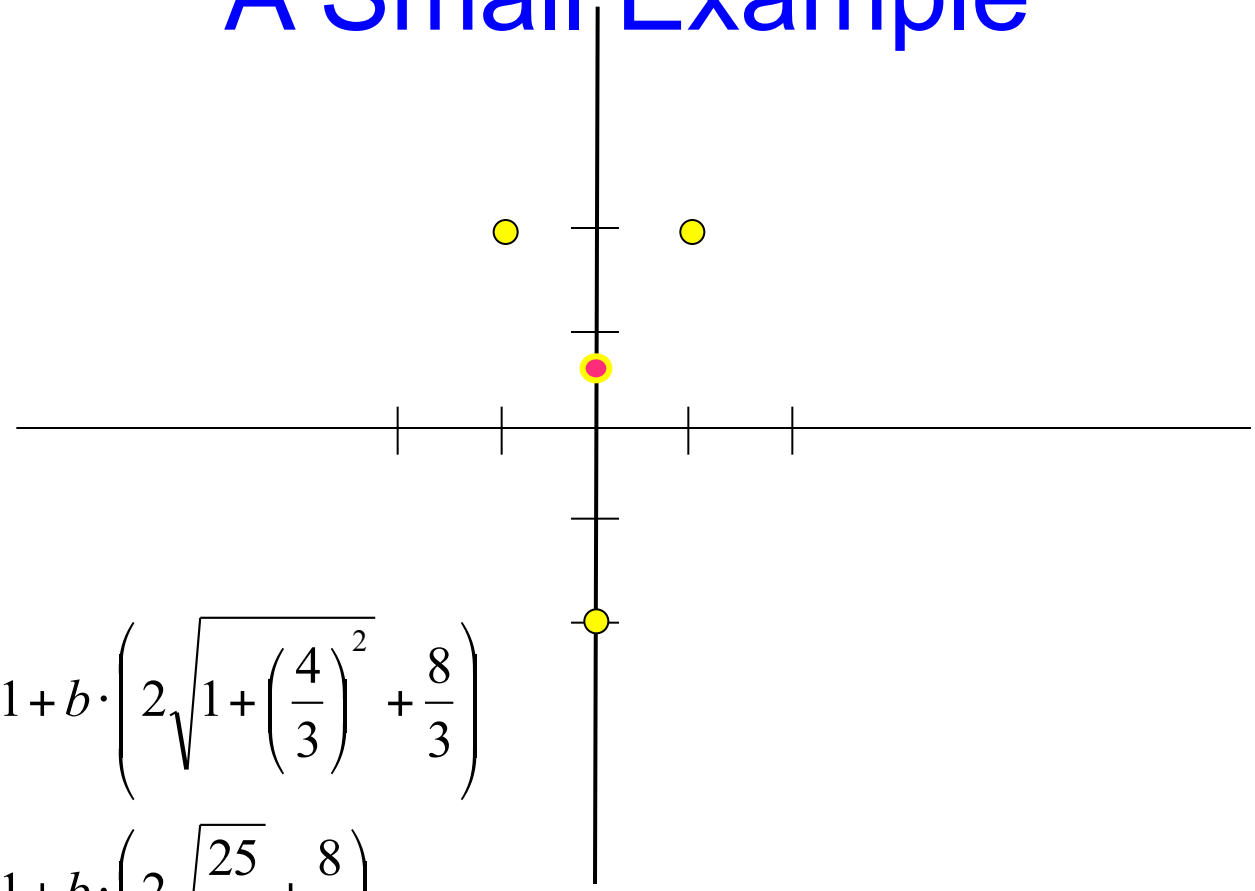
$$f(\rho, C) = a|C| + b\sum_{c_i \in C} D_{c_i}$$

A Small Example



$$f(\rho, C) = a \cdot |C| + b \cdot D_{c_1}$$

A Small Example



$$f(\rho, C) = a \cdot 1 + b \cdot \left(2\sqrt{1 + \left(\frac{4}{3}\right)^2} + \frac{8}{3} \right)$$

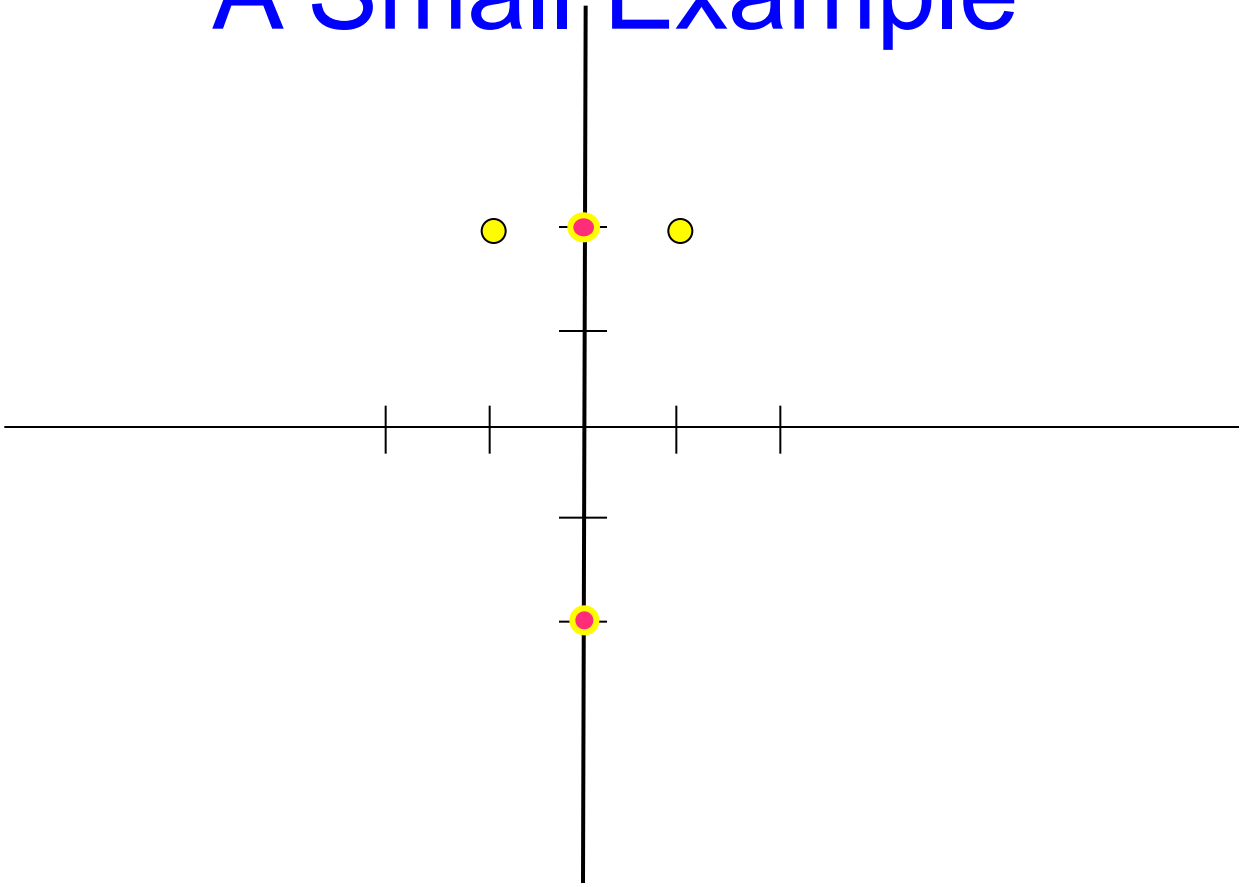
$$= a \cdot 1 + b \cdot \left(2\sqrt{\frac{25}{9}} + \frac{8}{3} \right)$$

$$= a \cdot 1 + b \cdot 6$$

A Small Example

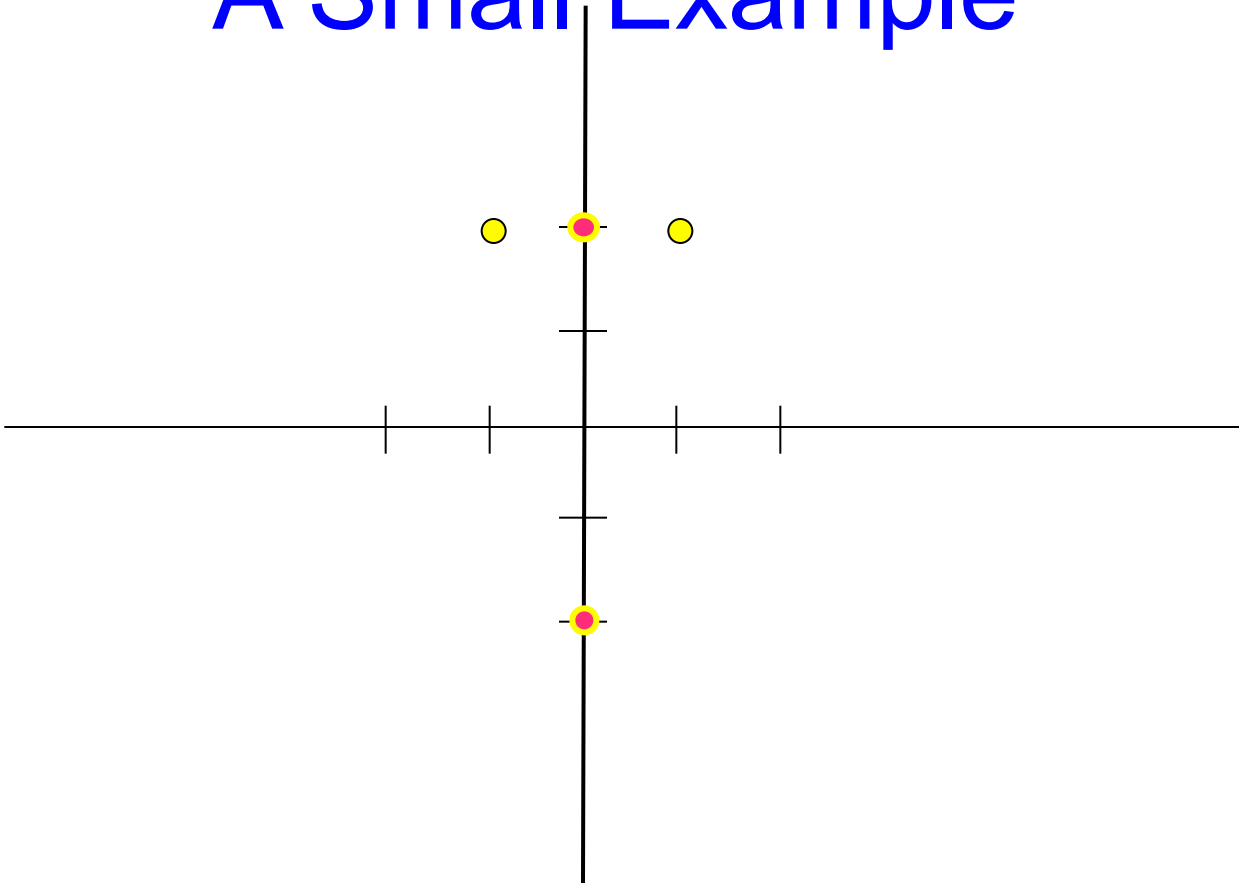
- So if we weight the distance metric as more costly, say b = 10, than the number of cluster centers where a = 1, then the ‘cost’ function is:
- 61

A Small Example



$$f(\rho, C) = a \cdot |C| + b \cdot D_{c_1}$$

A Small Example



$$f(\rho, C) = a \cdot 2 + b \cdot (2 + 0)$$

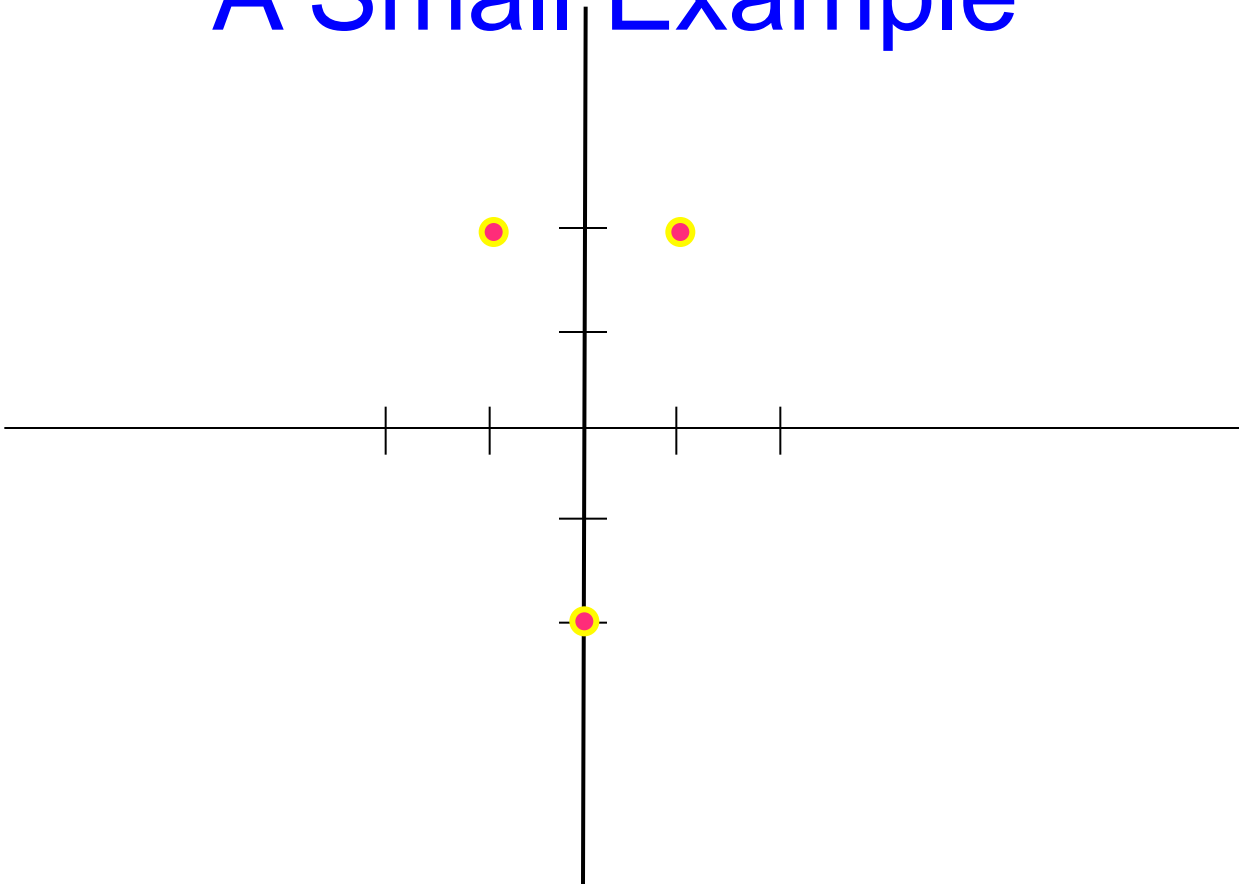
A Small Example

- So with a = 1 and b = 10, the ‘cost’ is: 22.

A Small Example

- What if the relative costs were different? Let a = 10, b = 1:
- Then the first arrangement has value:
- $10 \cdot 1 + 1 \cdot (6) = 16$
- And the second arrangement = 22

A Small Example



$$f(\rho, C) = a \cdot 3 + b \cdot (0)$$

A Small Example

- Now the cost = 3.

If a = 10, b = 1, then cost is 30

Summary of Examples

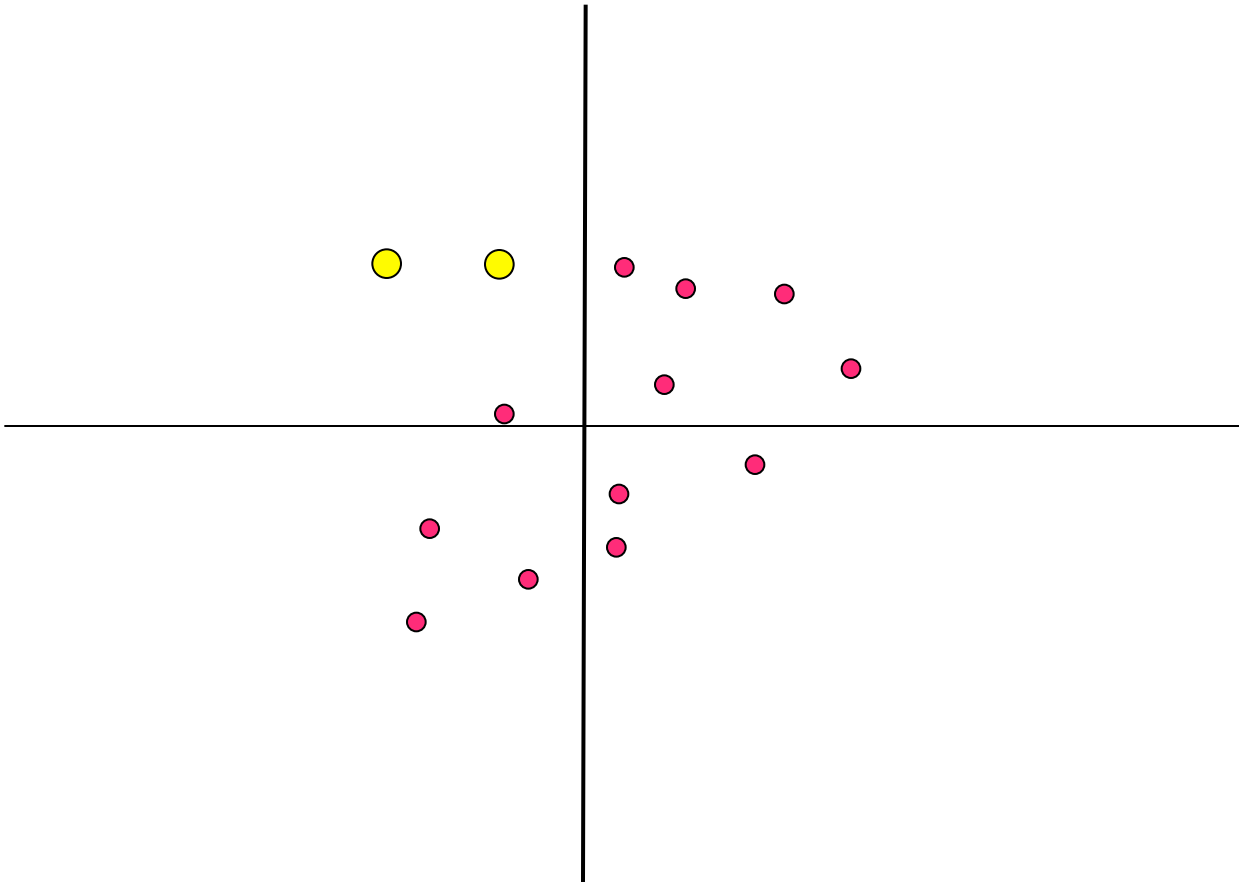
Number of Clusters	a =1, b = 10	a =10, b = 1
1	61	16
2	22	22
3	3	30

The Optimization Model

- Once a metric is chosen, the function *f* is a function of the cluster centers, and the number of them.
- We want to minimize *f.*
- How hard is this problem?
- Really, really hard if we want to be totally objective!

Illustrating the Algorithm

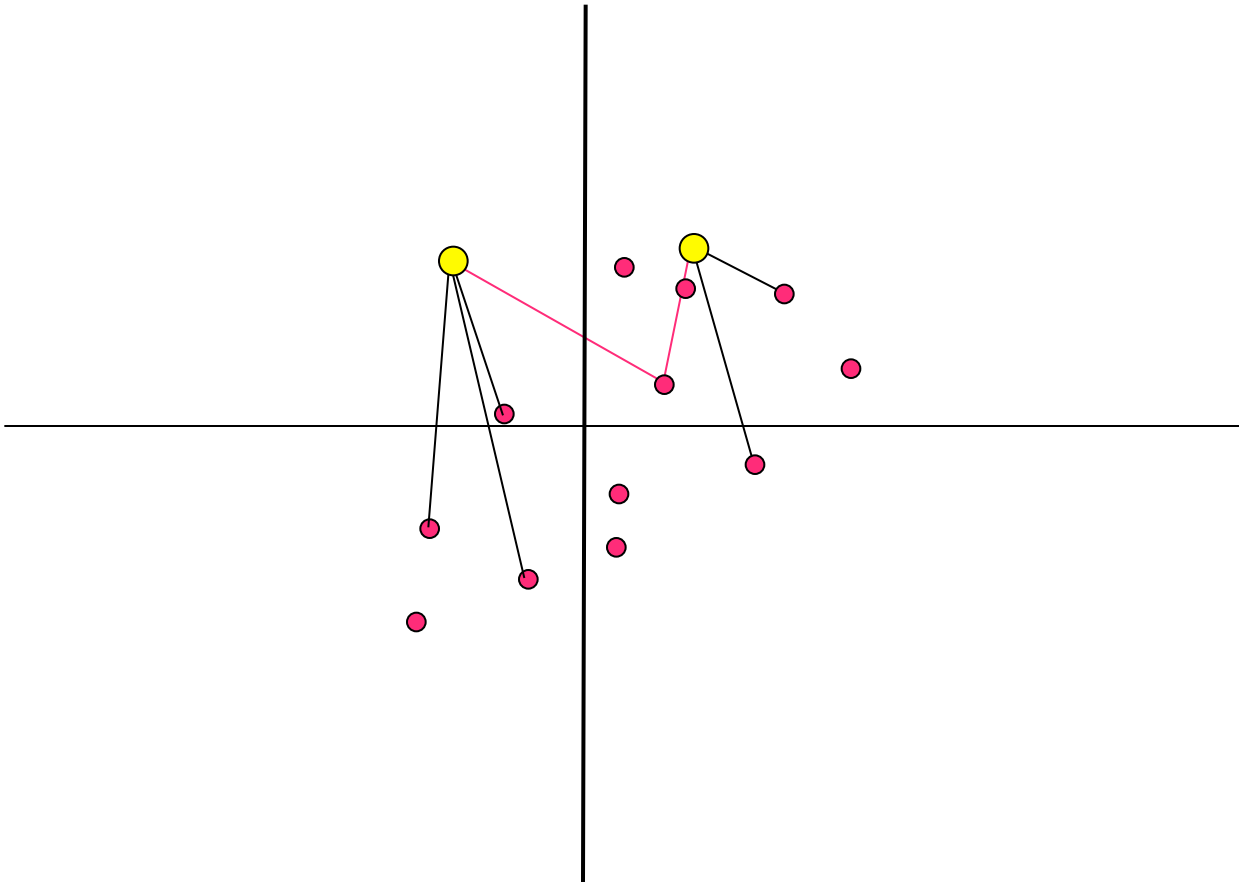
Say we are now considering two cluster centers



We can add more clusters. How many?

Illustrating the Algorithm

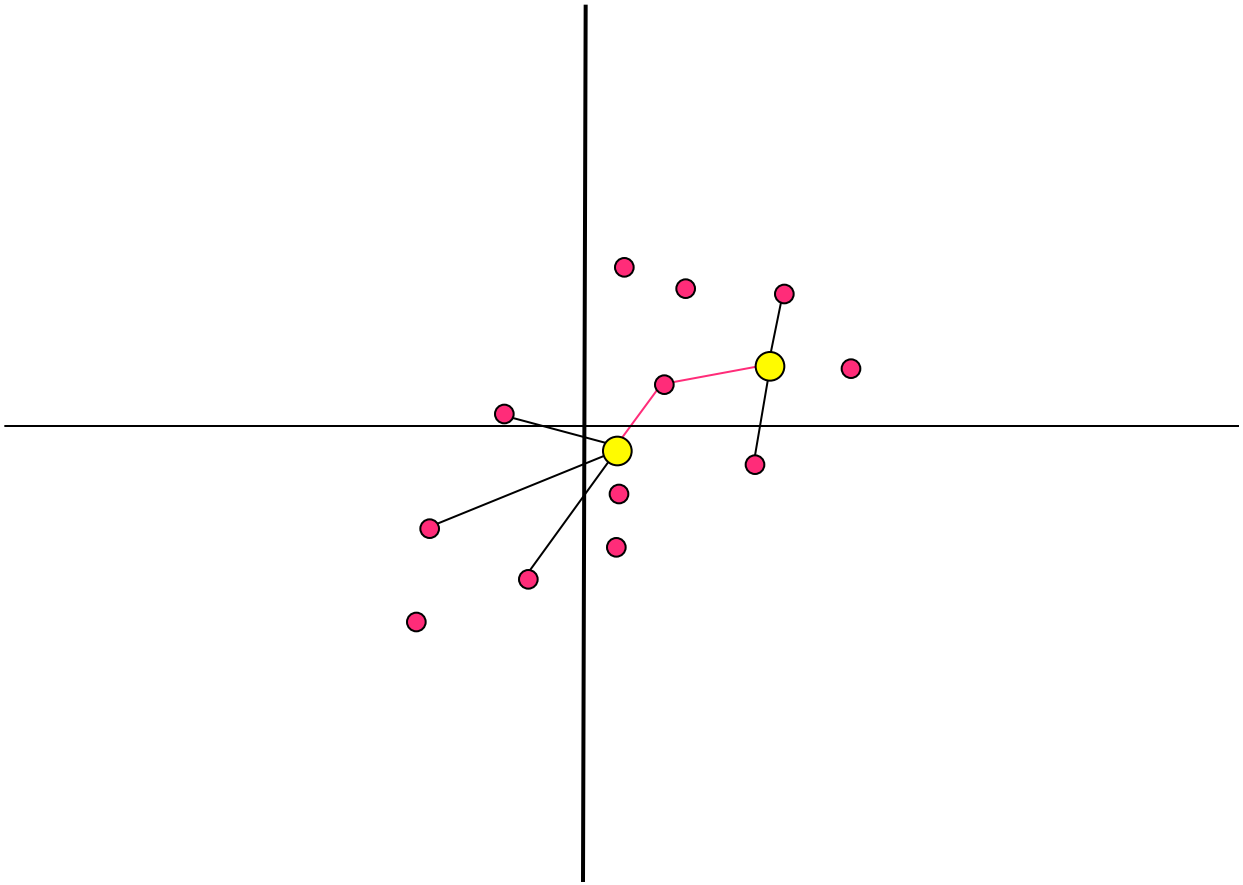
Say we are now considering two cluster centers



Determine cluster location, membership and distances.

Illustrating the Algorithm

Say we are now considering two cluster centers



Determine cluster location, membership and distances.

The Optimization Problem

1. Initialize *a* and *b.*
2. Add another cluster center to the mix.
3. Place the cluster centers.
4. Determine point membership based on cluster center placement.
5. Calculate the function *f.*
6. If *f* is less than the best so far, save it.
7. Goto 3 to reposition cluster centers.
8. If after all cluster center positions have been used, goto step 2.

Heuristics to the Rescue!

- We can make intelligent guesses as to initializing cluster center locations.
- Then add additional cluster centers up to some number.
- How can neural networks be used here?