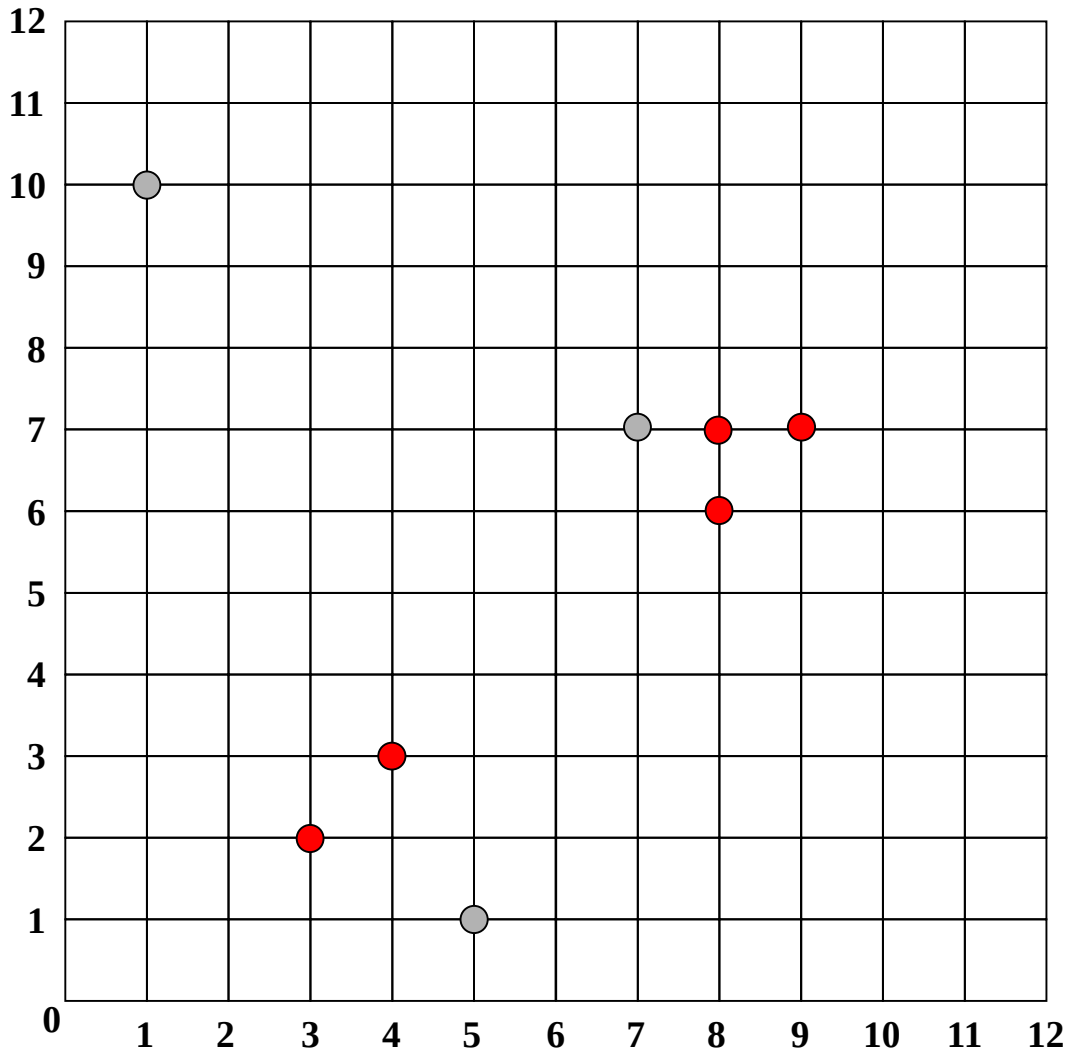
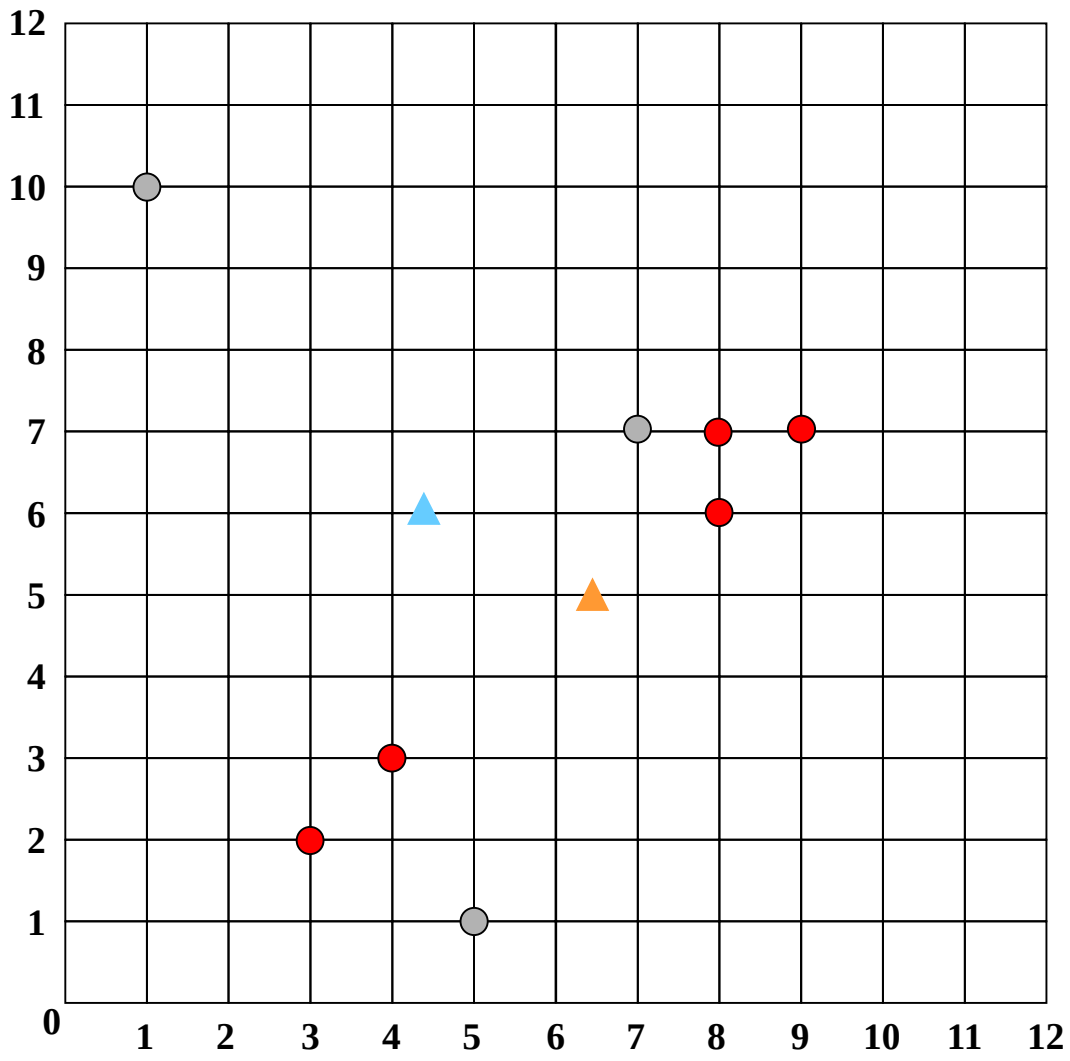
Clustering with variance minimization (k=2)



Initialization

Clustering with variance minimization (k=2)

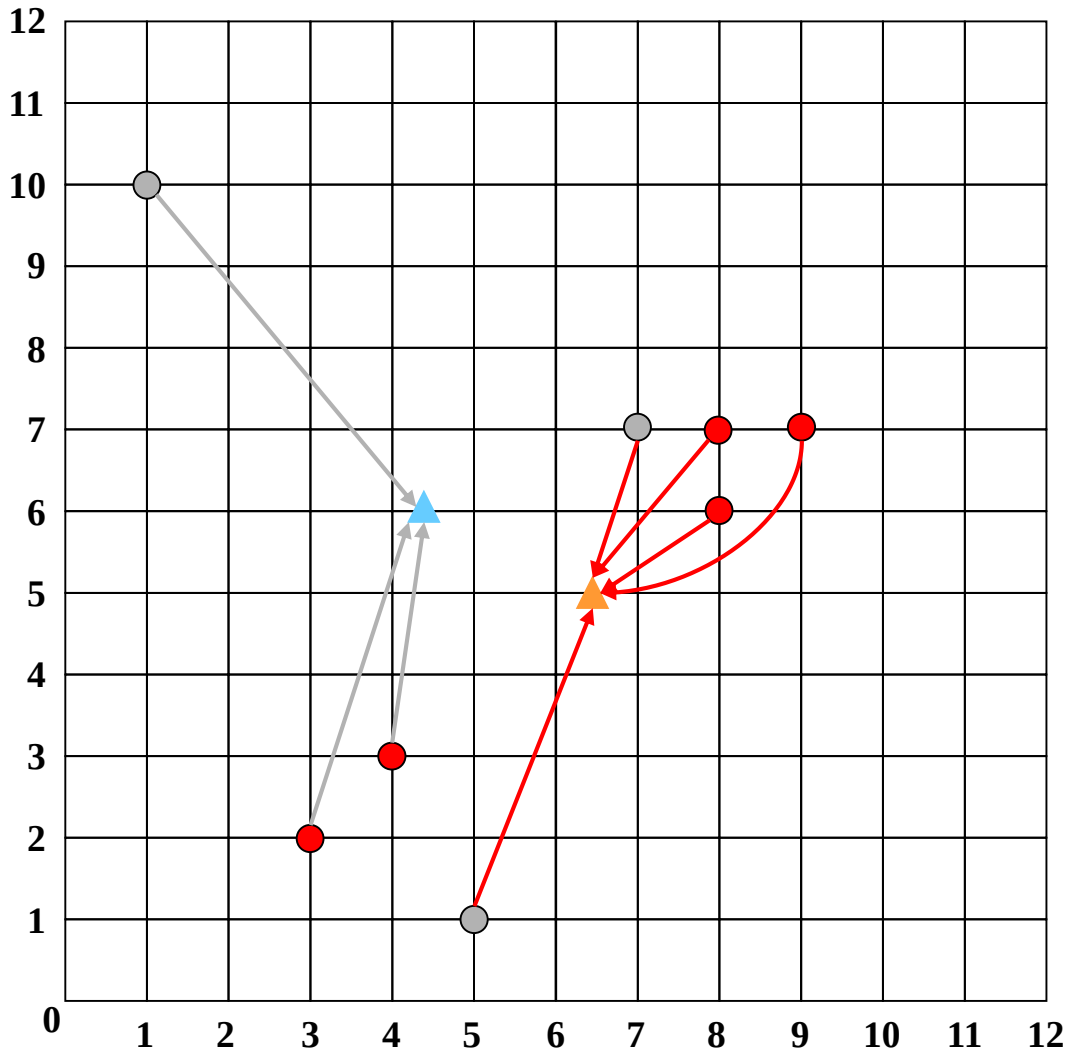


Calculate centroids

$\mu = (4.3 / 6.0)$

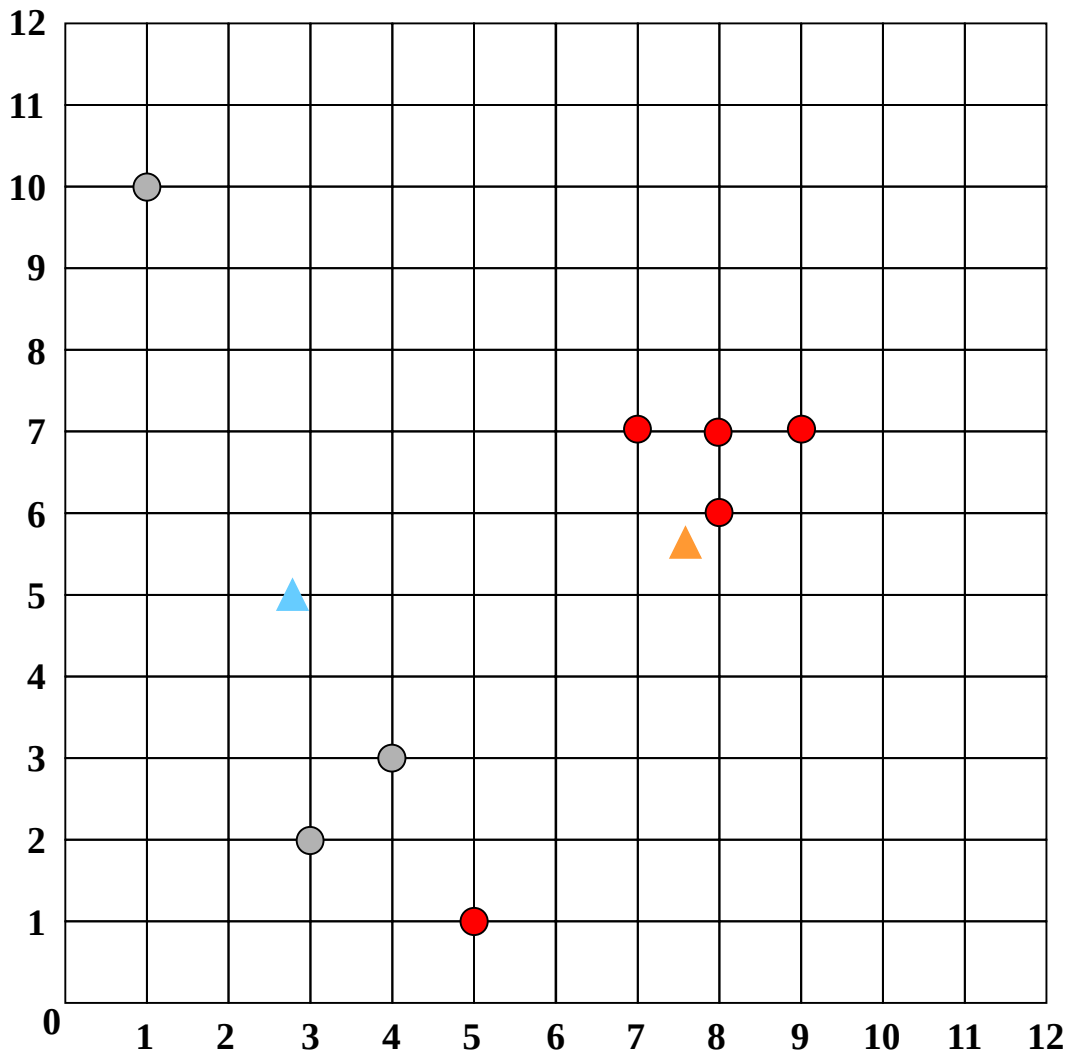
$\mu = (6.4 / 5.0)$

Clustering with variance minimization (k=2)



Reassign points

Clustering with variance minimization (k=2)

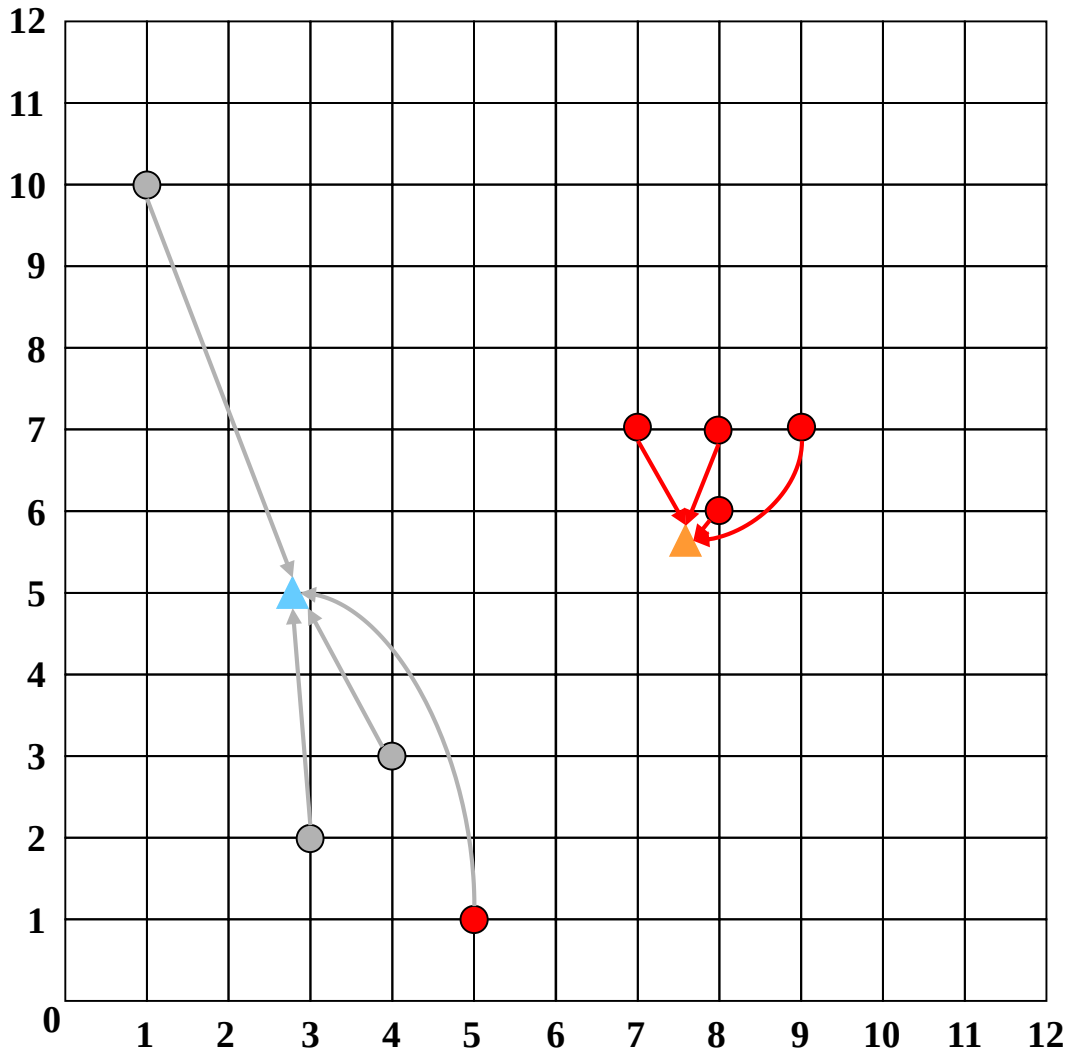


Calculate centroids

$\mu = (2.7 / 5.0)$

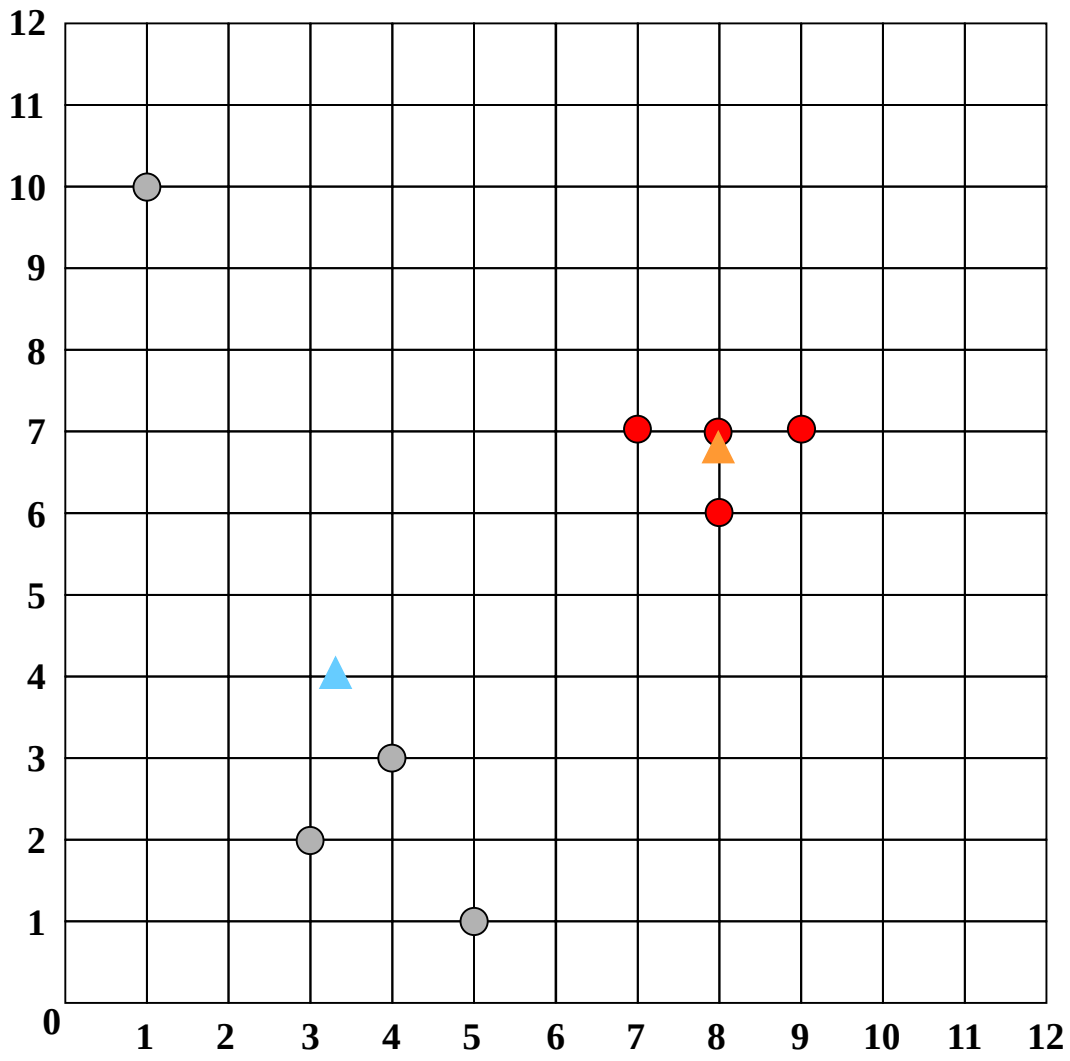
$\mu = (7.4 / 5.6)$

Clustering with variance minimization (k=2)



Reassign points

Clustering with variance minimization (k=2)

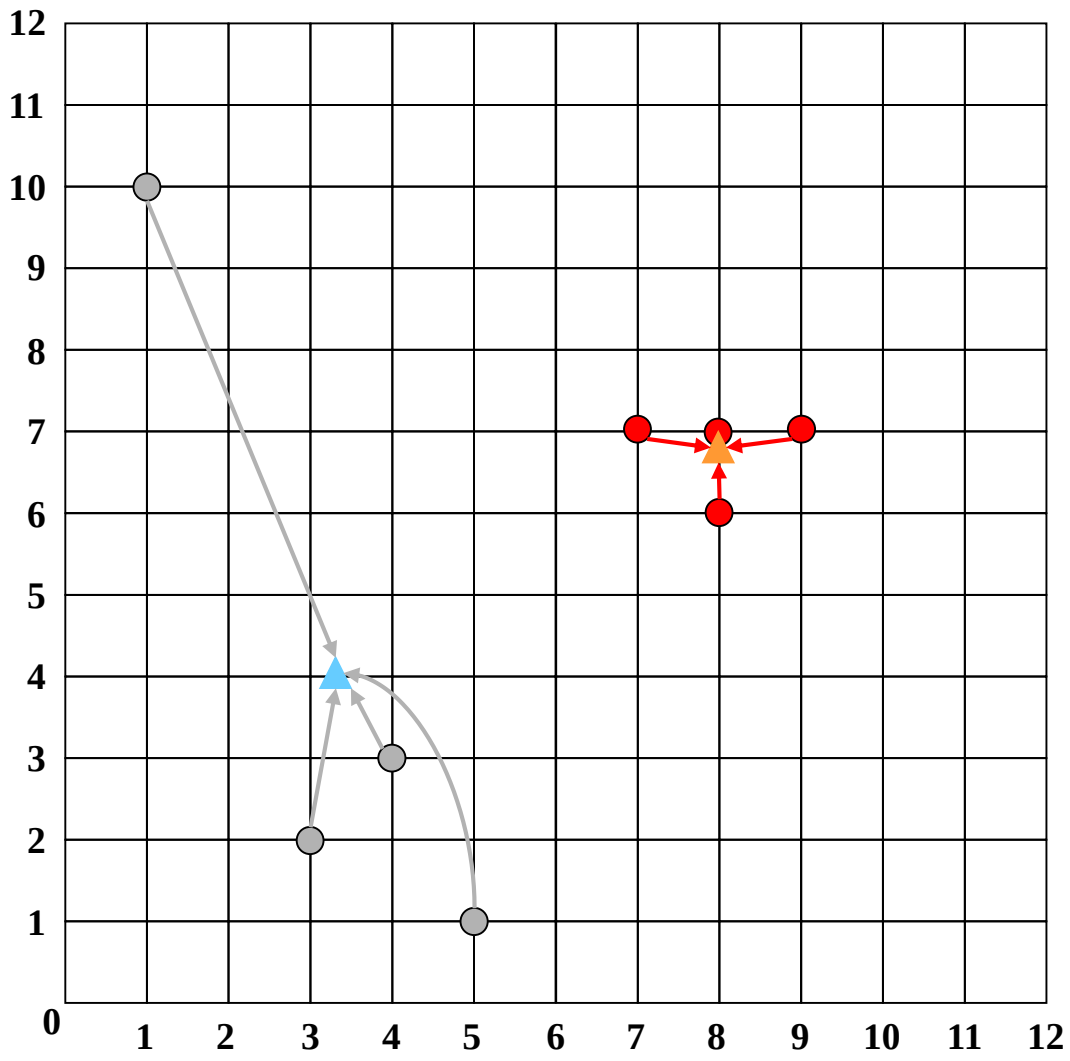


Calculate centroids

$\mu = (3.25 / 4.0)$

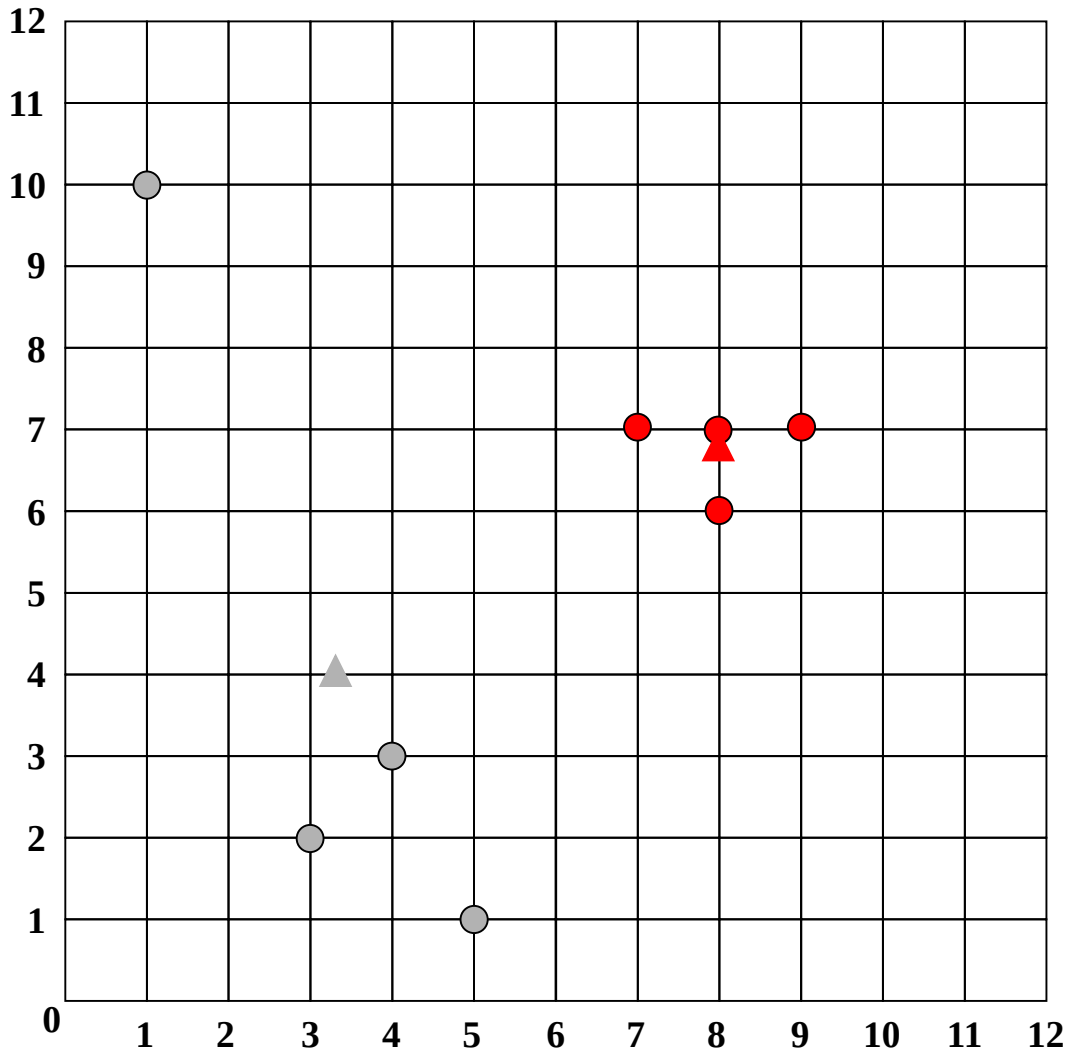
$\mu = (8.0 / 6.75)$

Clustering with variance minimization (k=2)



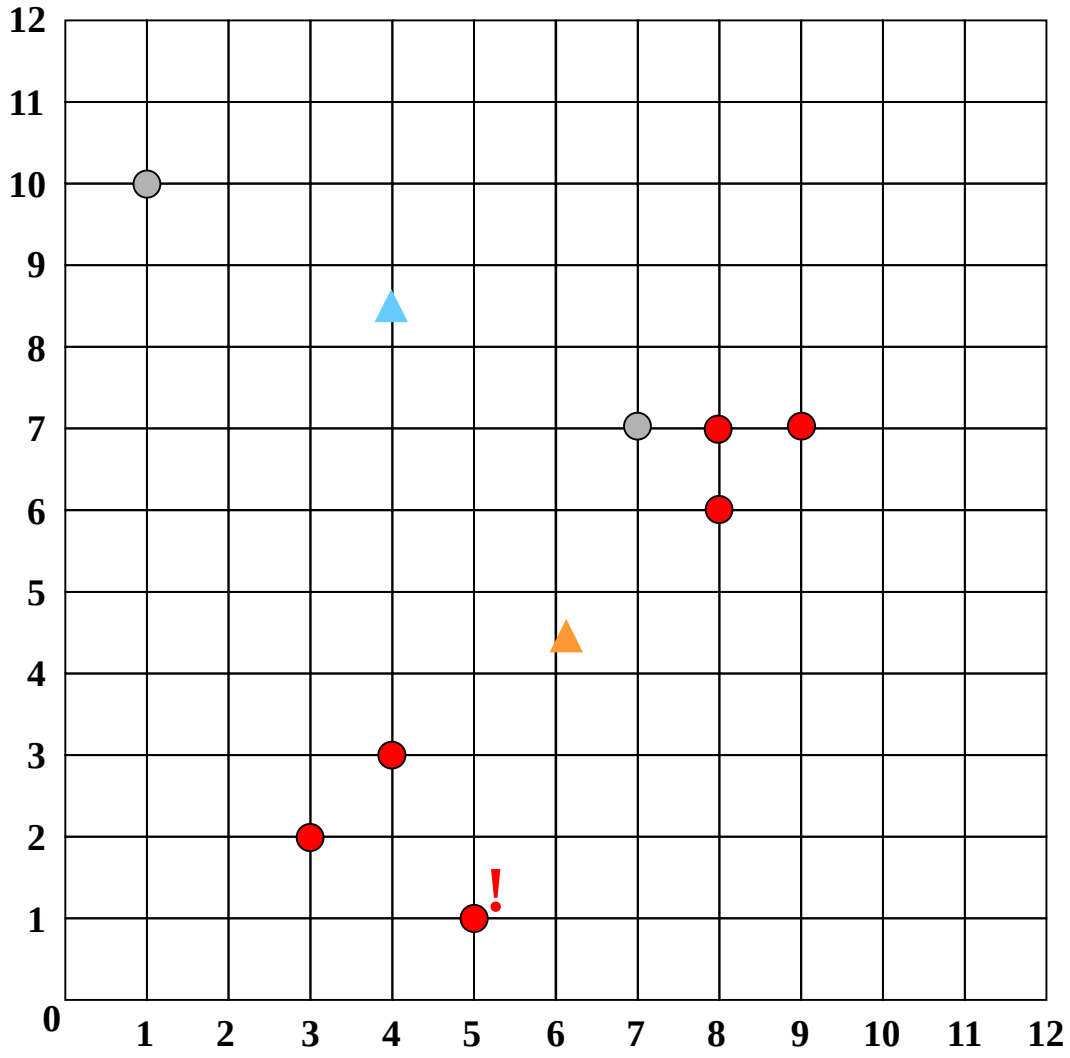
Reassign points

Clustering with variance minimization (k=2)



Convergence

Clustering with variance minimization (k=2): alternative initial clustering



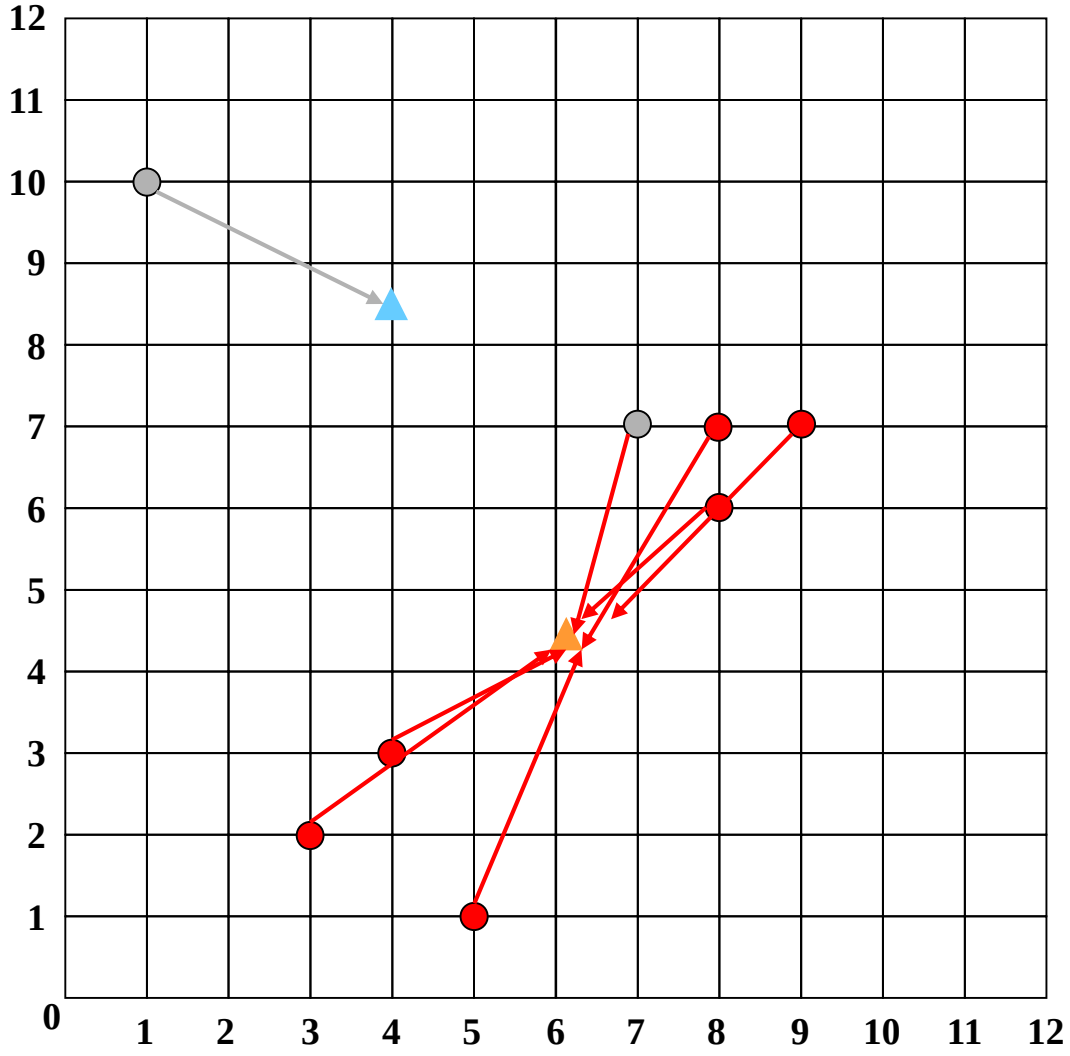
Initialization

Calculate centroids

μ = (4 / 8.5)

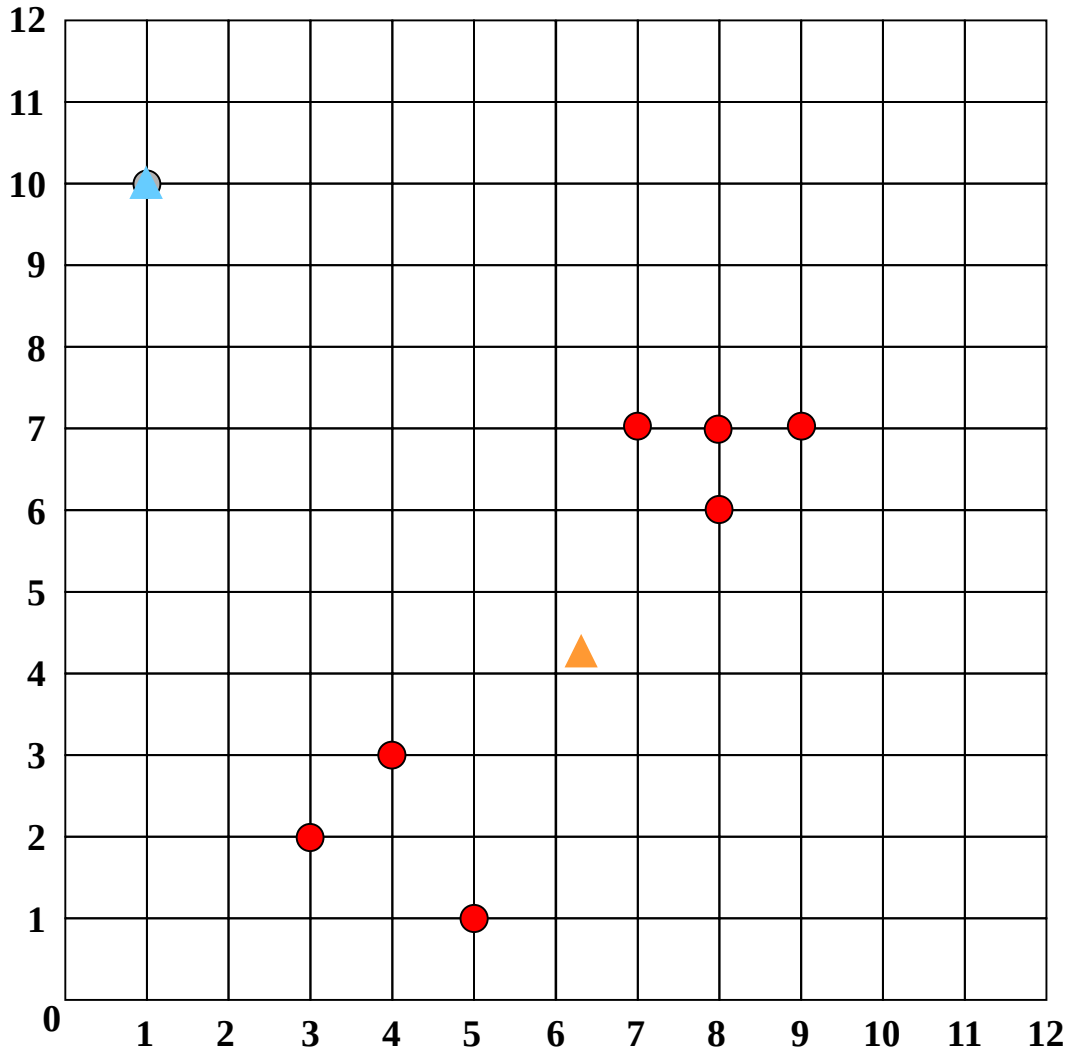
μ = (6.17/ 4.33)

Clustering with variance minimization (k=2): alternative initial clustering



Reassign points

Clustering with variance minimization (k=2): alternative initial clustering

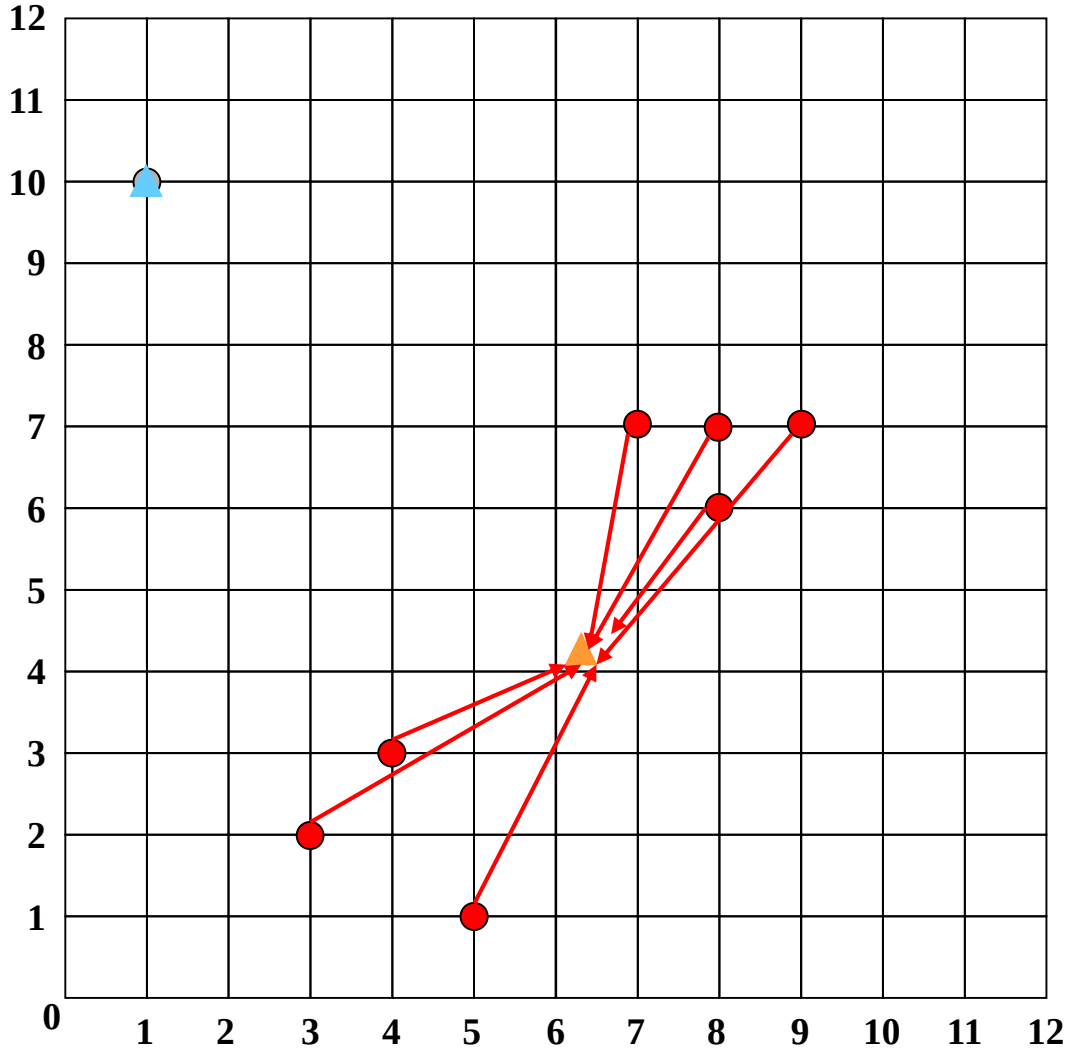


Calculate centroids

$\mu = (1 / 10)$

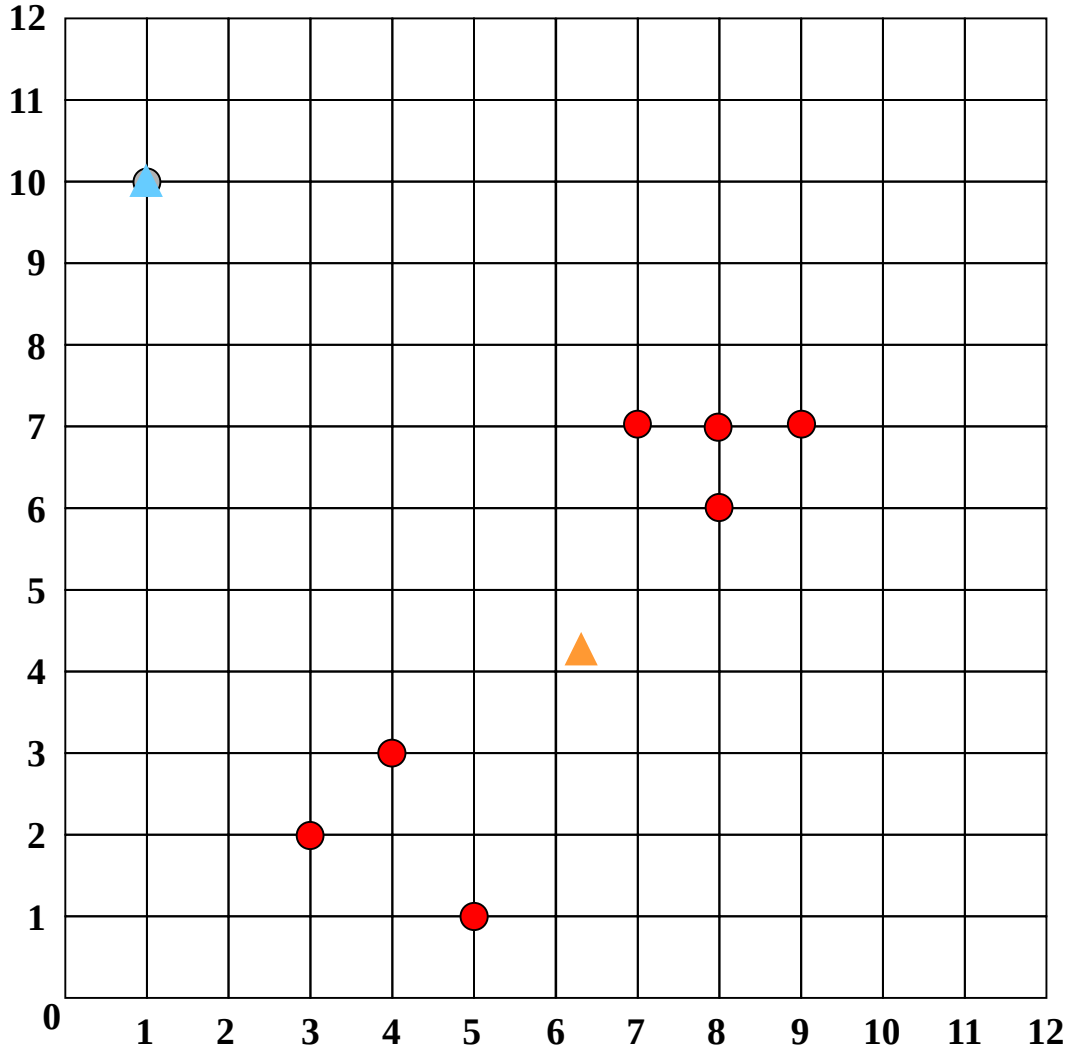
$\mu = (6.3/ 4.7)$

Clustering with variance minimization (k=2): alternative initial clustering



Reassign points

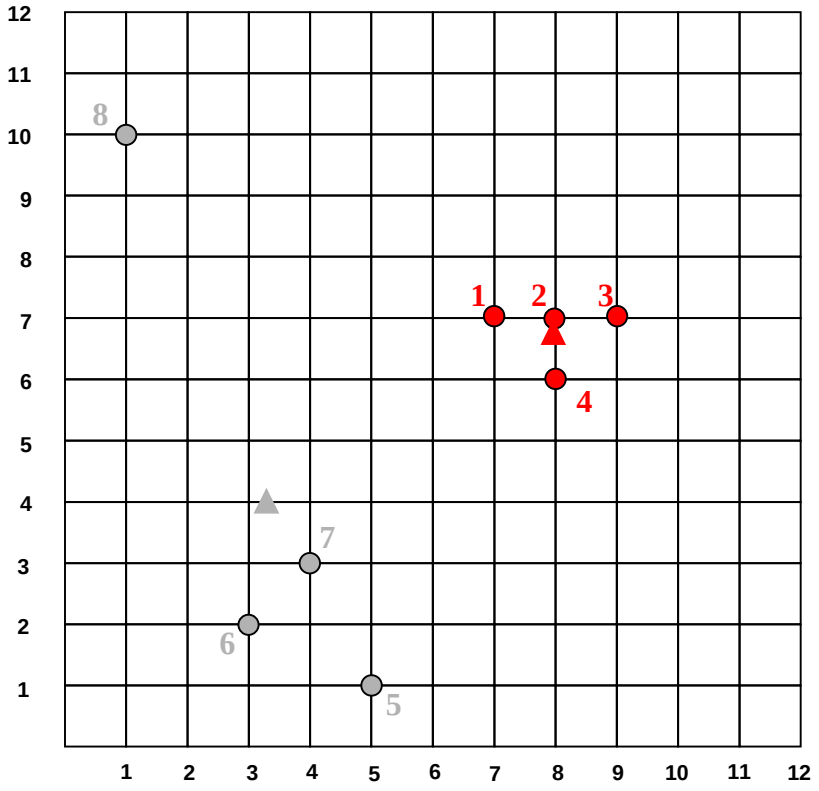
Clustering with variance minimization (k=2): alternative initial clustering



Convergence

Costs?

Obtained clustering



$\mu = (3.25 / 4.0)$

$\mu = (8.0 / 6.75)$

$Dist(\mu,1) = |8.0 - 7.0| + |6.75 - 7.0| = 1 + 0.25 = 1.25$

$Dist(\mu,2) = |8.0 - 8.0| + |6.75 - 7.0| = 0.25$

$Dist(\mu,4) = |8.0 - 9.0| + |6.75 - 7.0| = 1.25$

$Dist(\mu,3) = |8.0 - 8.0| + |6.75 - 6.0| = 0.75$

$TD^2(C1) = 3.75$

$Dist(\mu,5) = |3.25 - 5.0| + |4.0 - 1.0| = 1.75 + 3 = 4.75$

$Dist(\mu,6) = |3.25 - 3.0| + |4.0 - 2.0| = 0.25 + 2 = 2.25$

$Dist(\mu,7) = |3.25 - 4.0| + |4.0 - 3.0| = 0.75 + 1 = 1.75$

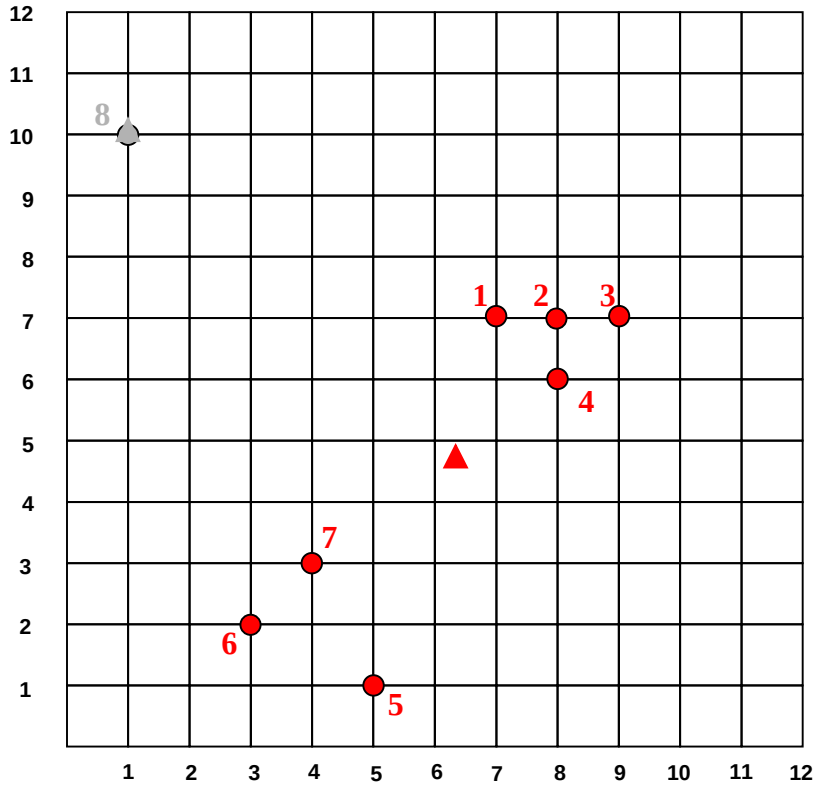
$Dist(\mu,8) = |3.25 - 1.0| + |4.0 - 10.0| = 2.75 + 6 = 8.75$

$TD^2(C2) = 107.25$

$TD^2 = 107.25 + 3.75 = 111$

Costs?

Alternative clustering



$\mu = (1.0 / 10.0)$

$\mu = (6.3 / 4.7)$

$Dist(\mu,1) = |6.3 - 7.0| + |4.7 - 7.0| = 0.7 + 2.3 = 3.0$

$Dist(\mu,2) = |6.3 - 8.0| + |4.7 - 7.0| = 1.7 + 2.3 = 4.0$

$Dist(\mu,3) = |6.3 - 9.0| + |4.7 - 7.0| = 2.7 + 2.3 = 5.0$

$Dist(\mu,4) = |6.3 - 8.0| + |4.7 - 6.0| = 1.7 + 1.3 = 3.0$

$Dist(\mu,5) = |6.3 - 5.0| + |4.7 - 1.0| = 1.3 + 3.7 = 5.0$

$Dist(\mu,6) = |6.3 - 3.0| + |4.7 - 2.0| = 3.3 + 2.7 = 6.0$

$Dist(\mu,7) = |6.3 - 4.0| + |4.7 - 3.0| = 2.3 + 1.7 = 4.0$

$TD^2(C1) = 136$

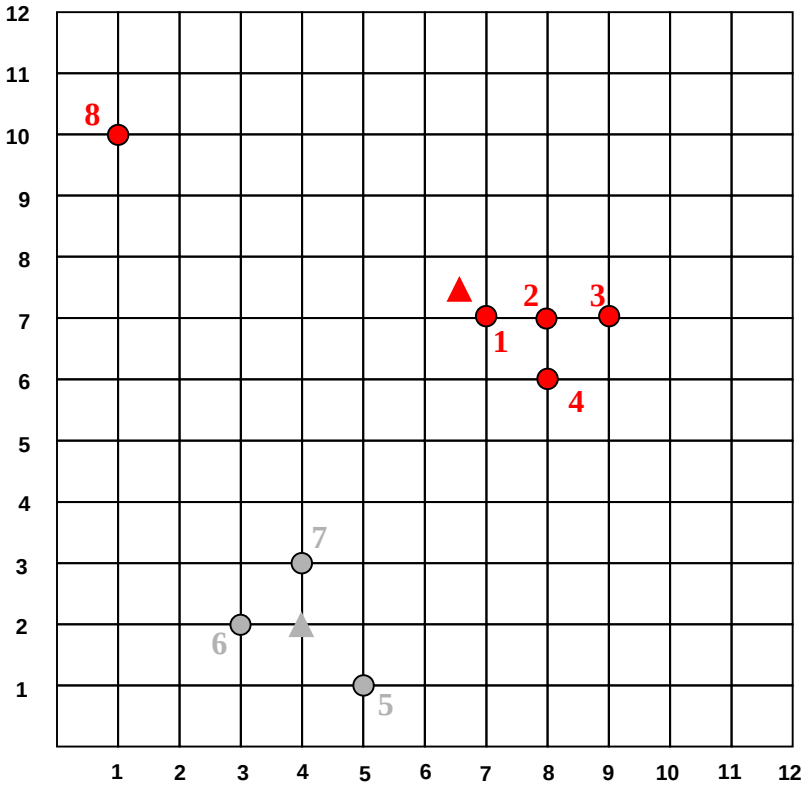
$Dist(\mu,8) = |1.0 - 1.0| + |10.0 - 10.0| = 0$

$TD^2(C2) = 0$

$TD^2 = 136$

Costs?

Optimal Clustering



$\mu = (4.0 / 2.0)$

$\mu = (6.6 / 7.4)$

$Dist(\mu,1) = |6.6 - 7.0| + |7.4 - 7.0| = 0.4 + 0.4 = 0.8$

$Dist(\mu,2) = |6.6 - 8.0| + |7.4 - 7.0| = 1.4 + 0.4 = 1.8$

$Dist(\mu,3) = |6.6 - 8.0| + |7.4 - 6.0| = 1.4 + 1.4 = 2.8$

$Dist(\mu,4) = |6.6 - 9.0| + |7.4 - 7.0| = 2.4 + 0.4 = 2.8$

$Dist(\mu,8) = |6.6 - 1.0| + |7.4 - 10.0| = 5.6 + 2.6 = 8.2$

$TD^2(C1) = 86.8$

$Dist(\mu,5) = |4.0 - 5.0| + |2.0 - 1.0| = 2$

$Dist(\mu,6) = |4.0 - 3.0| + |2.0 - 2.0| = 1$

$Dist(\mu,7) = |4.0 - 4.0| + |2.0 - 3.0| = 1$

$TD^2(C2) = 6$

$TD^2 = 86.8 + 6.0 = 92.8$

Conclusion

- Variant k-Means: k-Means already recalculates the centroids after reassigning one node only, instead of recalculating after regarding all nodes.
- Note, that clustering with variance minimization…:
  - …mostly converges only to a local minimum
  - …depends on the order of the assignment and on the initial partitioning
  - …is prone to noise

  Nevertheless, clustering with variance minimization is at the moment the most popular clustering algorithm in industry and in large parts of science