Chapter 14

- **K-Means Clustering**
  - Goal: find the c mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_c$.
  - Replace the squared Mahalanobis distance

$(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)^t \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)$ by the squared Eculidean distance

$\left\| \mathbf{x}_k - \hat{\boldsymbol{\mu}}_i \right\|^2$.

  - Find the mean $\hat{\boldsymbol{\mu}}_m$ nearest to $\boldsymbol{x}_k$ and approximate

$$\hat{P}(\omega_i \mid \mathbf{x}_k, \hat{\boldsymbol{\theta}}) \quad \text{as} \quad \hat{P}(\omega_i \mid \mathbf{x}_k, \boldsymbol{\theta}) \cong \begin{cases} 1 & \text{if } i = m \\ 0 & \text{otherwise} \end{cases}$$

  - Use the iterative scheme to find $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}}_c$.

– If n is the known number of patterns and c the desired number of clusters, the k-means algorithm is (c samples randomly chosen from the dataset as initial cluster centers):

Begin

initialize n, c, μ_1, μ_2, …, μ_c(randomly selected)

do classify n samples according to nearest μ_i

recompute μ_i

until no change in μ_i

return μ_1, μ_2, …, μ_c

End

Exercise 2 p.594 (Textbook)

Example:

The 25 samples shown in the table were drawn sequentially from the following mixture with $\mu_1 = -2$ and $\mu_2 = 2$.

$$p(x|\mu_1, \mu_2) = \underbrace{\frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu_1)^2\right]}_{\omega_1} + \underbrace{\frac{2}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu_2)^2\right]}_{\omega_2},$$

k	x_k	ω_1	ω_2
1	0.608		×
2	-1.590	×	
3	0.235		×
4	3.949		×
5	-2.249	×	
6	2.704		×
7	-2.473	×	
8	0.672		×

k	x_k	ω_1	ω_2
9	0.262		×
10	1.072		×
11	-1.773	×	
12	0.537		×
13	3.240		×
14	2.400		×
15	-2.499	×	
16	2.608		×

k	x_k	ω_1	ω_2
17	-3.458	×	
18	0.257		×
19	2.569		×
20	1.415		×
21	1.410		×
22	-2.653	×	
23	1.396		×
24	3.286		×
25	-0.712	×	

- Considering the example in the previous figure

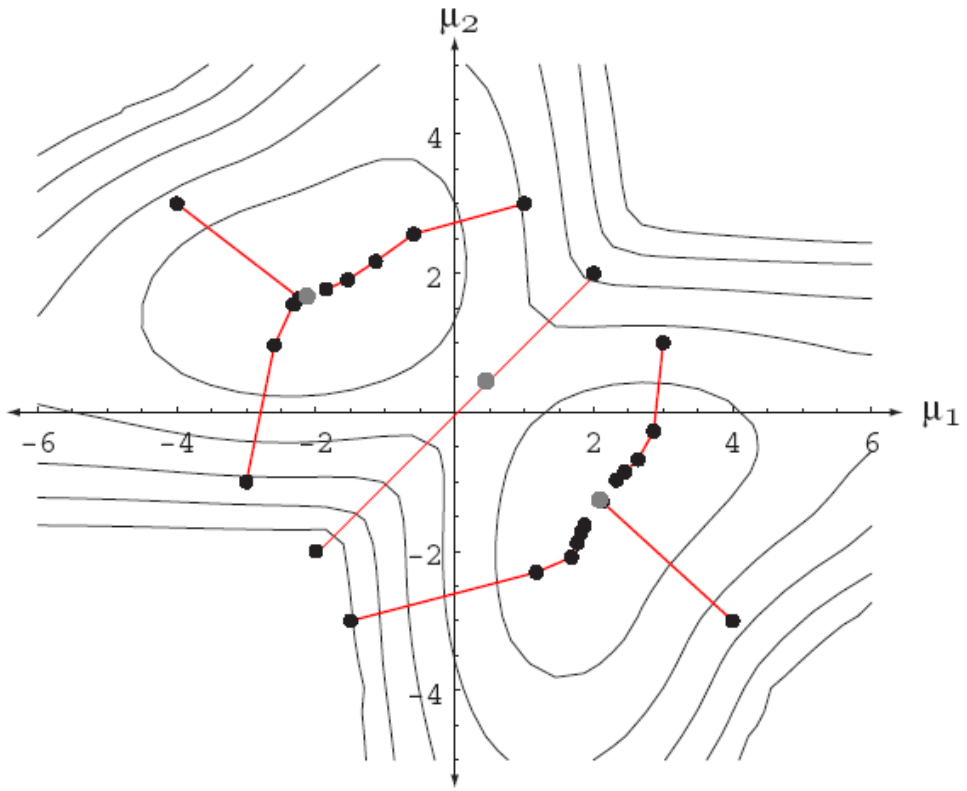
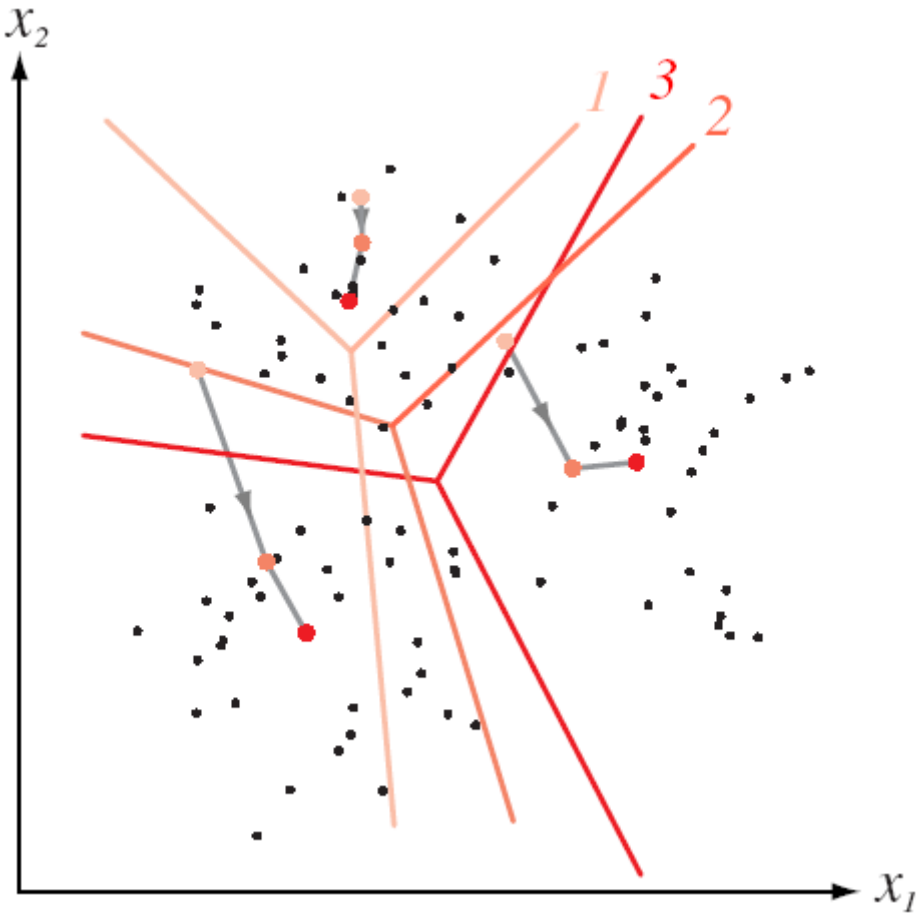


Figure 10.1: The *k*-means clustering procedure is a form of stochastic hill climbing in the log-likelihood function. The contours represent equal log-likelihood values for the one-dimensional data in Example 1. The dots indicate parameter values after different iterations of the *k*-means algorithm. Six of the starting points shown lead to local maxima, whereas two (i.e., $\mu_1(0) = \mu_2(0)$) *lead to a saddle point near* $\boldsymbol{\mu} = \mathbf{0}$.

- Figure 10.1 shows the sequence of values for $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ obtained for several different starting points. Since interchanging $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ merely interchanges the labels assigned to the data, the trajectories are symmetric about the line $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_2$.The trajectories lead either to the point $\hat{\boldsymbol{\mu}}_1 = -2.176$, $\hat{\boldsymbol{\mu}}_2 = 1.684$ or to its symmetric image. This is close to the solution found by the maximum-likelihood method (viz., $\hat{\boldsymbol{\mu}}_2 = -2.130$ and $\hat{\boldsymbol{\mu}}_1 = 1.688$), and the trajectories show a general resemblance to those shown in Example 1.



C=3

The three initial cluster centers, chosen randomly from the training points.

FIGURE 10.3. Trajectories for the means of the k-means clustering procedure applied to two-dimensional data. The final Voronoi tesselation (for classification) is also shown— the means correspond to the "centers" of the Voronoi cells. In this case, convergence is obtained in three iterations.

Fuzzy *k*-means clustering

- In every iteration of the classical k-means procedure, each data point is assumed to be in exactly one cluster
- We can relax this condition and assume that each sample $\mathbf{x}_j$ has some graded or "fuzzy" cluster membership $\mu_i(\mathbf{x}_j)$ in cluster ω_i, where $0 \leq \mu_i(\mathbf{x}_j) \leq 1$.
- At root, these "memberships" are equivalent to the probabilities $\hat{P}(\omega_i \mid \mathbf{x}_j, \hat{\boldsymbol{\theta}})$
- In the resulting fuzzy k-means clustering algorithm we seek a minimum of a global cost function

$$L = \sum_{i=1}^{c} \sum_{j=1}^{n} [\hat{P}(\omega_i | \mathbf{x}_j, \hat{\theta})]^b ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2,$$

where *b* is a free parameter chosen to adjust the "blending" of different clusters. If *b* is set to 0, this criterion function is merely a sum-of-squared errors criterion we shall see again in Eq. 49. If *b*>1, criterion allows each pattern to belong to multiple clusters.

The probabilities of cluster membership for each point are normalized as

$$\sum_{i=1}^{c} \hat{P}(\omega_i | \mathbf{x}_j) = 1, \qquad j = 1, \ldots, n. \tag{25}$$

At the solution, i.e., the minimum of *L*, we have

$$\partial L / \partial \boldsymbol{\mu}_i = 0 \quad \text{and} \quad \partial L / \partial \hat{P}_j = 0,$$

Then we have

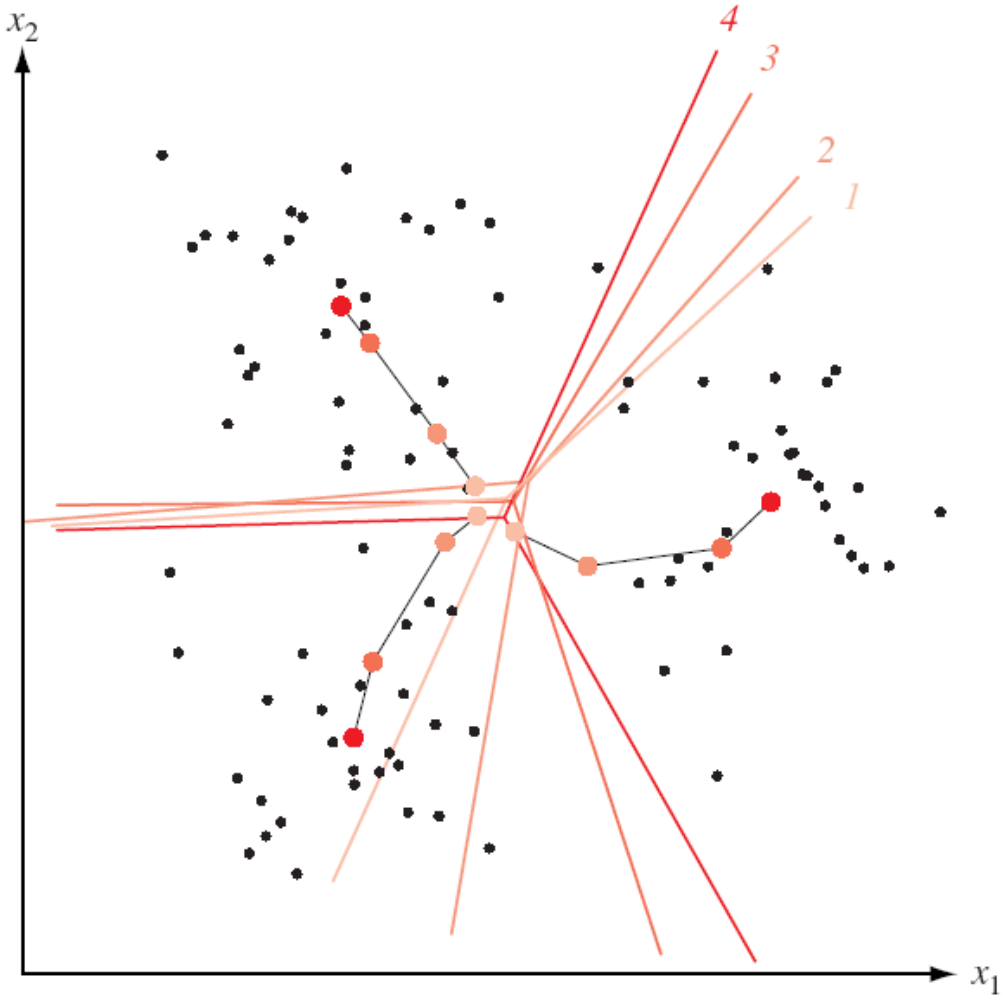
$$\boldsymbol{\mu}_j = \frac{\sum_{j=1}^{n} [P(\omega_i|\mathbf{x}_j)]^b \mathbf{x}_j}{\sum_{j=1}^{n} [P(\omega_i|\mathbf{x}_j)]^b} \qquad (27)$$

and

$$P(\omega_i|\mathbf{x}_j) = \frac{(1/d_{ij})^{1/(b-1)}}{\sum_{r=1}^{c} (1/d_{rj})^{1/(b-1)}}, \qquad d_{ij} = ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2. \qquad (28)$$

Algorithm 2 (Fuzzy k-means clustering)

1 **begin initialize** $n, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, P(\omega_i \mid \mathbf{x}_j), i = 1 \ldots, c;\ j = 1, \ldots, n$
2 normalize proabilities of cluster memberships by Eq. 25
3 **do** classify n samples according to nearest $\boldsymbol{\mu}_i$
4 recompute $\boldsymbol{\mu}_i$ by Eq. 27
5 recompute $P(\omega_i \mid \mathbf{x}_j)$ by Eq. 28
6 **until** no change in $\boldsymbol{\mu}_i$ and $P(\omega_i \mid \mathbf{x}_j)$
7 **return** $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_c$
8 **end**



At early iterations the means lie near the center of the full data set because each point has a non-negligible "membership" (i.e., probability) in each cluster. At later iterations the means separate and each membership tends toward the value 1.0 or 0.0.

The classical *k*-means algorithm is just of special case where the memberships for all points obey

$$P(\omega_i|\mathbf{x}_j) = \begin{cases} 1 & \text{if } \|\mathbf{x}_j - \boldsymbol{\mu}_i\| < \|\mathbf{x}_j - \boldsymbol{\mu}_{i'}\| \text{ for all } i' \neq i \\ 0 & \text{otherwise,} \end{cases}$$