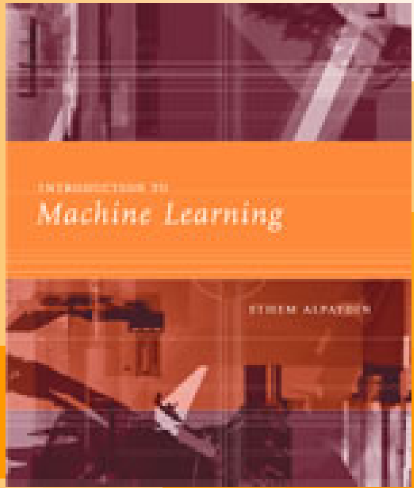


Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml

CHAPTER 7:

Clustering

Semiparametric Density Estimation

- Parametric: Assume a single model for $p(\mathbf{x} \mid C_i)$ (Chapter 4 and 5)
- Semiparametric: $p(\mathbf{x} \mid C_i)$ is a mixture of densities

  Multiple possible explanations/prototypes:

  Different handwriting styles, accents in speech
- Nonparametric: No model; data speaks for itself (Chapter 8)

Mixture Densities

$$p(\mathbf{x}) = \sum_{i=1}^{k} p(\mathbf{x} \mid \mathcal{G}_i) P(\mathcal{G}_i)$$

where $\mathcal{G}_i$ the components/groups/clusters,

$P(\mathcal{G}_i)$ mixture proportions (priors),

$p(\mathbf{x} \mid \mathcal{G}_i)$ component densities

Gaussian mixture where $p(\mathbf{x}|\mathcal{G}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

parameters $\Phi = \{P(\mathcal{G}_i), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{k}$

unlabeled sample $\mathcal{X}=\{\mathbf{x}^t\}_t$ (unsupervised learning)

Classes vs. Clusters

- Supervised: $\mathcal{X} = \{ \mathbf{x}^t, \mathbf{r}^t \}_t$
- Classes C_i $i=1,...,K$

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x} \mid C_i) P(C_i)$$

where $p(\mathbf{x} \mid C_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

- $\Phi = \{P(C_i), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{K}$

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

- Unsupervised : $\mathcal{X} = \{ \mathbf{x}^t \}_t$
- Clusters $\mathcal{G}_i$ $i=1,...,k$

$$p(\mathbf{x}) = \sum_{i=1}^{k} p(\mathbf{x} \mid \mathcal{G}_i) P(\mathcal{G}_i)$$

where $p(\mathbf{x} \mid \mathcal{G}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

- $\Phi = \{P(\mathcal{G}_i), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{k}$

Labels, $\mathbf{r}_i^t$?

k-Means Clustering

- Find k reference vectors (prototypes/codebook vectors/codewords) which best represent data
- Reference vectors, $\boldsymbol{m}_j$, $j = 1,\ldots,k$
- Use nearest (most similar) reference:

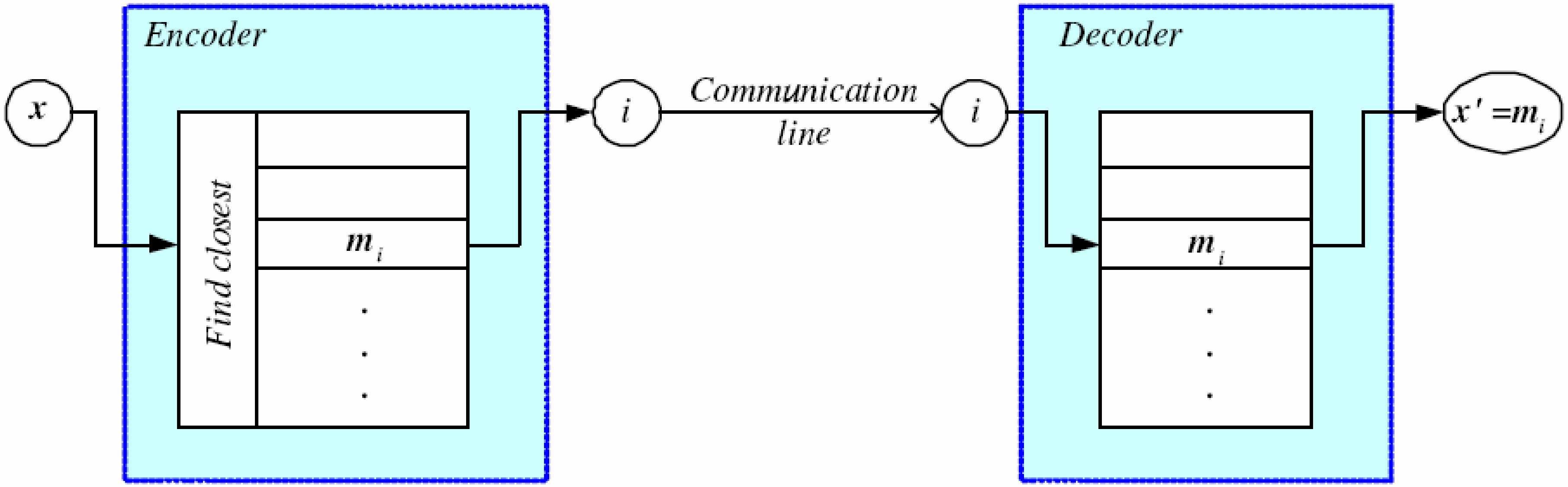
$$\left\| \boldsymbol{x}^t - \boldsymbol{m}_i \right\| = \min_j \left\| \boldsymbol{x}^t - \boldsymbol{m}_j \right\|$$

- Reconstruction error

$$E\left(\{\boldsymbol{m}_i\}_{i=1}^k \middle| \mathcal{X} \right) = \sum_t \sum_i b_i^t \left\| \boldsymbol{x}^t - \boldsymbol{m}_i \right\|$$

$$b_i^t = \begin{cases} 1 & \text{if } \left\| \boldsymbol{x}^t - \boldsymbol{m}_i \right\| = \min_j \left\| \boldsymbol{x}^t - \boldsymbol{m}_j \right\| \\ 0 & \text{otherwise} \end{cases}$$

Encoding/Decoding



$$b_i^t = \begin{cases} 1 & \text{if } \left\| \boldsymbol{x}^t - \boldsymbol{m}_i \right\| = \min_j \left\| \boldsymbol{x}^t - \boldsymbol{m}_j \right\| \\ 0 & \text{otherwise} \end{cases}$$

k-means Clustering

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$

Repeat

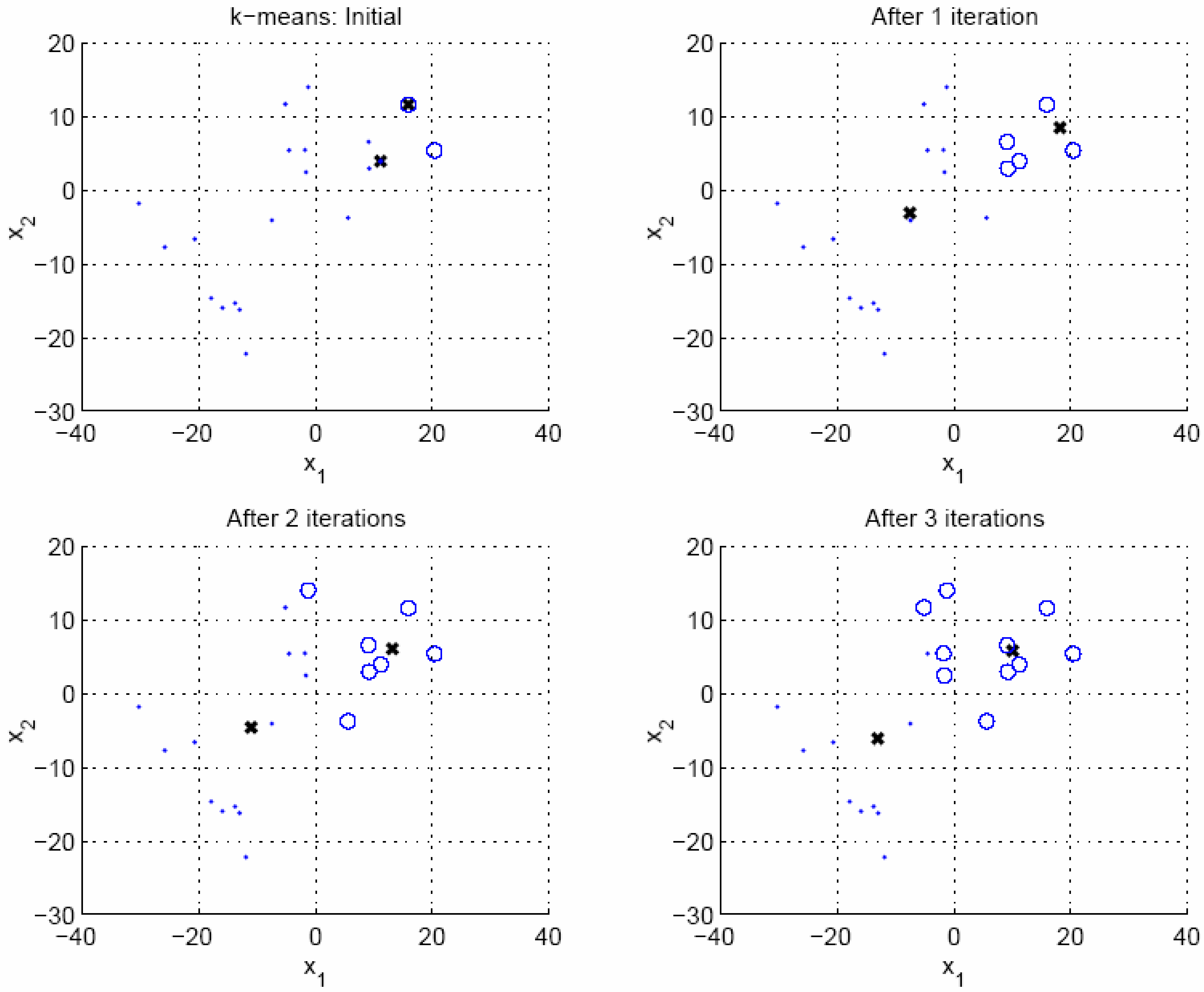
For all $\boldsymbol{x}^t \in \mathcal{X}$

$$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

For all $\boldsymbol{m}_i, i = 1, \ldots, k$

$$\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$$

Until $\boldsymbol{m}_i$ converge


k−means: Initial
After 1 iteration
After 2 iterations
After 3 iterations
x1
x2

Expectation-Maximization (EM)

- Log likelihood with a mixture model

$$\mathcal{L}(\Phi \mid \mathcal{X}) = \log \prod_t p(\boldsymbol{x}^t \mid \Phi)$$

$$= \sum_t \log \sum_{i=1}^{k} p(\boldsymbol{x}^t \mid \mathcal{G}_i) P(\mathcal{G}_i)$$

- Assume hidden variables z, which when known, make optimization much simpler
- Complete likelihood, $\mathcal{L}_c(\Phi \mid \mathcal{X}, \mathcal{Z})$, in terms of $\boldsymbol{x}$ and $\boldsymbol{z}$
- Incomplete likelihood, $\mathcal{L}(\Phi \mid \mathcal{X})$, in terms of $\boldsymbol{x}$

E- and M-steps

- Iterate the two steps

1. E-step: Estimate z given $\mathcal{X}$ and current Φ
2. M-step: Find new Φ' given z, $\mathcal{X}$, and old Φ.

$$\text{E-step}: \mathcal{Q}\left(\Phi \mid \Phi^l\right) = E\left[\mathcal{L}_C\left(\Phi \mid \mathcal{X}, \mathcal{Z}\right) \mid \mathcal{X}, \Phi^l\right]$$

$$\text{M-step}: \Phi^{l+1} = \arg\max_{\Phi} \mathcal{Q}\left(\Phi \mid \Phi^l\right)$$

An increase in $\mathcal{Q}$ increases incomplete likelihood

$$\mathcal{L}\left(\Phi^{l+1} \mid \mathcal{X}\right) \geq \mathcal{L}\left(\Phi^l \mid \mathcal{X}\right)$$

EM in Gaussian Mixtures

- $z^t_i = 1$ if $\boldsymbol{x}^t$ belongs to $\mathcal{G}_i$, 0 otherwise (labels $\boldsymbol{r}^t_i$ of supervised learning); assume $p(\boldsymbol{x}|\mathcal{G}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
- E-step:

$$E\left[z_i^t \middle| \mathcal{X}, \Phi^l\right] = \frac{p\left(\boldsymbol{x}^t \mid \mathcal{G}_i, \Phi^l\right) P\left(\mathcal{G}_i\right)}{\sum_j p\left(\boldsymbol{x}^t \mid \mathcal{G}_j, \Phi^l\right) P\left(\mathcal{G}_j\right)}$$

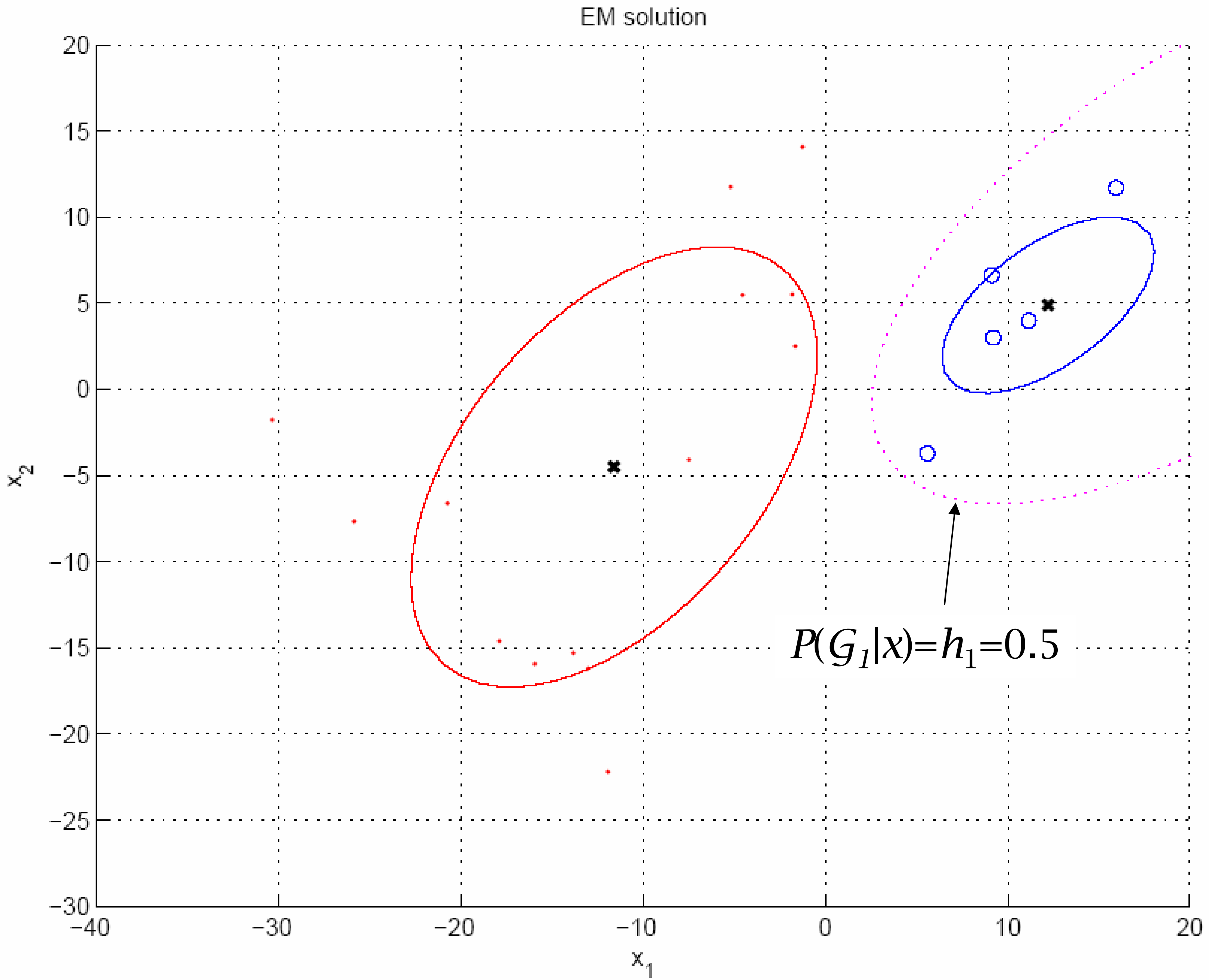
$$= P\left(\mathcal{G}_i \mid \boldsymbol{x}^t, \Phi^l\right) \equiv h_i^t$$

- M-step:

$$P\left(\mathcal{G}_i\right) = \frac{\sum_t h_i^t}{N} \qquad \boldsymbol{m}_i^{l+1} = \frac{\sum_t h_i^t \boldsymbol{x}^t}{\sum_t h_i^t}$$

Use estimated labels in place of unknown labels

$$\mathbf{S}_i^{l+1} = \frac{\sum_t h_i^t \left(\boldsymbol{x}^t - \boldsymbol{m}_i^{l+1}\right)\left(\boldsymbol{x}^t - \boldsymbol{m}_i^{l+1}\right)^T}{\sum_t h_i^t}$$


EM solution
20
15
10
5
0
−5
−10
−15
−20
−25
−30
−40
−30
−20
−10
0
10
20
x2
x1
P(G1|x)=h1=0.5

Mixtures of Latent Variable Models

- Regularize clusters

1. Assume shared/diagonal covariance matrices
2. Use PCA/FA to decrease dimensionality: Mixtures of PCA/FA

$$p(\boldsymbol{x}_t \mid \mathcal{G}_i) = \mathcal{N}\left(\boldsymbol{m}_i, \mathbf{V}_i \mathbf{V}_i^T + \boldsymbol{\psi}_i\right)$$

Can use EM to learn $\mathbf{V}_i$ (Ghahramani and Hinton, 1997; Tipping and Bishop, 1999)

After Clustering

- Dimensionality reduction methods find correlations between features and group features
- Clustering methods find similarities between instances and group instances
- Allows knowledge extraction through

  number of clusters,

  prior probabilities,

  cluster parameters, i.e., center, range of features.

  Example: CRM, customer segmentation

Clustering as Preprocessing

- Estimated group labels h_j (soft) or b_j (hard) may be seen as the dimensions of a new k dimensional space, where we can then learn our discriminant or regressor.
- Local representation (only one b_j is 1, all others are 0; only few h_j are nonzero) vs

  Distributed representation (After PCA; all z_j are nonzero)

Mixture of Mixtures

- In classification, the input comes from a mixture of classes (supervised).
- If each class is also a mixture, e.g., of Gaussians, (unsupervised), we have a mixture of mixtures:

$$p(\mathbf{x} \mid C_i) = \sum_{j=1}^{k_i} p(\mathbf{x} \mid \mathcal{G}_{ij}) P(\mathcal{G}_{ij})$$

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x} \mid C_i) P(C_i)$$

Hierarchical Clustering

- Cluster based on similarities/distances
- Distance measure between instances $\boldsymbol{x}^r$ and $\boldsymbol{x}^s$

Minkowski (L_p) (Euclidean for $p = 2$)

$$d_m(\boldsymbol{x}^r, \boldsymbol{x}^s) = \left[\sum_{j=1}^{d} \left(x_j^r - x_j^s\right)^p\right]^{1/p}$$

City-block distance

$$d_{cb}(\boldsymbol{x}^r, \boldsymbol{x}^s) = \sum_{j=1}^{d} \left|x_j^r - x_j^s\right|$$

Agglomerative Clustering

- Start with *N* groups each with one instance and merge two closest groups at each iteration
- Distance between two groups $\mathcal{G}_i$ and $\mathcal{G}_j$:
  - Single-link:

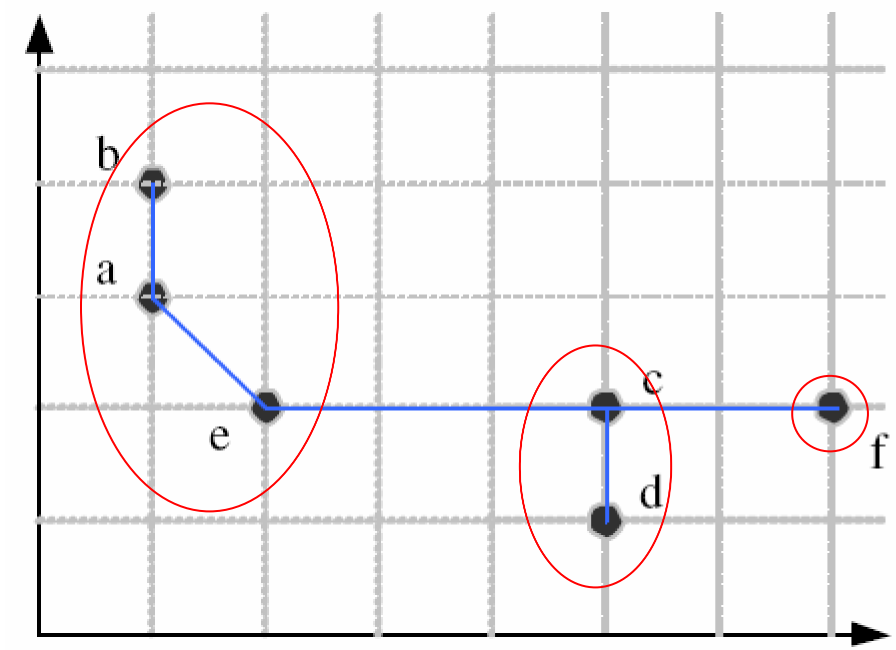
$$d(\mathcal{G}_i, \mathcal{G}_j) = \min_{\boldsymbol{x}^r \in \mathcal{G}_i, \boldsymbol{x}^s \in \mathcal{G}_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

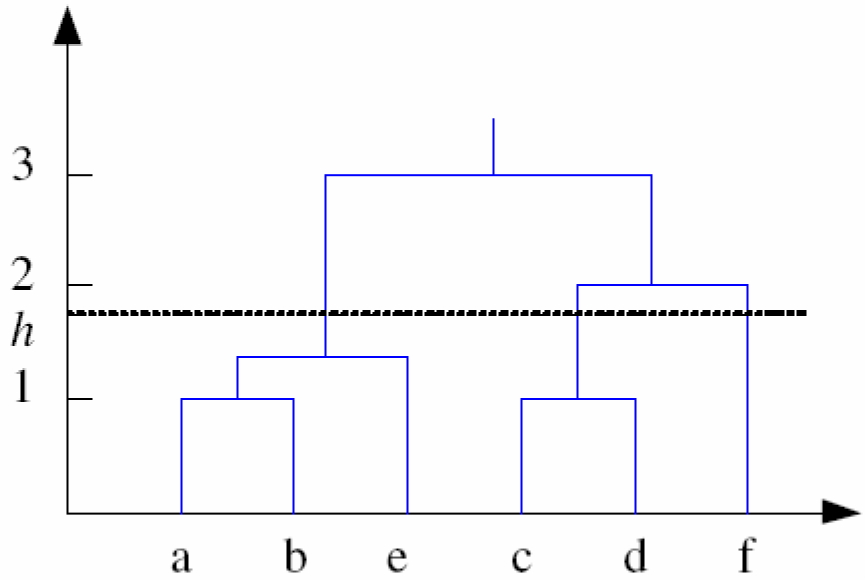
  - Complete-link:

$$d(\mathcal{G}_i, \mathcal{G}_j) = \max_{\boldsymbol{x}^r \in \mathcal{G}_i, \boldsymbol{x}^s \in \mathcal{G}_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

  - Average-link, centroid

Example: Single-Link Clustering





Dendrogram

Choosing k

- Defined by the application, e.g., image quantization
- Plot data (after PCA) and check for clusters
- Incremental (leader-cluster) algorithm: Add one at a time until “elbow” (reconstruction error/log likelihood/intergroup distances)
- Manual check for meaning