CHAPTER 7:
Clustering
Semiparametric Density Estimation

- **Parametric:** Assume a single model for $p (x | C_i)$ (Chapters 4 and 5)
- **Semiparametric:** $p (x | C_i)$ 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

\[ p(x) = \sum_{i=1}^{k} p(x | G_i) P(G_i) \]

where \( G_i \) the components/groups/clusters,
\( P(G_i) \) mixture proportions (priors),
\( p(x | G_i) \) component densities

Gaussian mixture where \( p(x | G_i) \sim N(\mu_i, \Sigma_i) \) parameters \( \Phi = \{P(G_i), \mu_i, \Sigma_i\}_{i=1}^{k} \)
Classes vs. Clusters

- **Supervised:** $X = \{x^t, r^t\}_t$
- **Classes $C_i$** $i=1,...,K$
  \[ p(x) = \sum_{i=1}^{K} p(x | C_i) P(C_i) \]
  where $p(x | C_i) \sim N(\mu_i, \Sigma_i)$
  \[ \Phi = \{P(C_i), \mu_i, \Sigma_i\}_{i=1}^{K} \]
  \[ \hat{P}(C_i) = \frac{\sum_{t} r_i^t}{N} \]
  \[ m_i = \frac{\sum_{t} r_i^t x^t}{\sum_{t} r_i^t} \]
  \[ S_i = \frac{\sum_{t} r_i^t (x^t - m_i)(x^t - m_i)^T}{\sum_{t} r_i^t} \]

- **Unsupervised:** $X = \{x^t\}_t$
- **Clusters $G_i$** $i=1,...,k$
  \[ p(x) = \sum_{i=1}^{k} p(x | G_i) P(G_i) \]
  where $p(x | G_i) \sim N(\mu_i, \Sigma_i)$
  \[ \Phi = \{P(G_i), \mu_i, \Sigma_i\}_{i=1}^{k} \]
  \[ \text{Labels, } r_i^t \, ? \]

Lecture Notes for E ALPAYDIN 2004
Introduction to Machine Learning © The MIT Press (V1.1)
**k-Means Clustering**

- Find $k$ reference vectors (prototypes/codebook vectors/codewords) which best represent data
- Reference vectors, $m_j$, $j = 1, \ldots, k$
- Use nearest (most similar) reference:
  \[
  \|x^t - m_i\| = \min_j \|x^t - m_j\|
  \]

- Reconstruction error
  \[
  E\left(\{m_i\}_{i=1}^k \mid X\right) = \sum_t \sum_i b_{i}^{t} \|x^t - m_i\|
  \]

  \[
  b_{i}^{t} = \begin{cases} 1 & \text{if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\ 0 & \text{otherwise} \end{cases}
  \]
Encoding/Decoding

![Diagram of Encoding and Decoding Process]

Lecture Notes for E ALPAYDIN 2004
Introduction to Machine Learning © The MIT Press (V1.1)

Wednesday, May 9, 2012
**k-means Clustering**

 Initialize $m_i, i = 1, \ldots, k$, for example, to $k$ random $x^t$

 Repeat

 For all $x^t \in X$

 $$b_i^t \left\{ \begin{array}{ll} 1 & \text{if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\ 0 & \text{otherwise} \end{array} \right.$$ 

 For all $m_i, i = 1, \ldots, k$

 $$m_i \leftarrow \frac{\sum_t b_i^t x^t}{\sum_t b_i^t}$$

 Until $m_i$ converge
Expectation-Maximization (EM)

• Log likelihood with a mixture model

\[ \mathcal{L}(\Phi | X) = \log \prod_t p(x^t | \Phi) \]

\[ = \sum_t \log \sum_{i=1}^k p(x^t | G_i) \mathcal{P}(G_i) \]

• Assume hidden variables \(z\), which when known, make optimization much simpler

• Complete likelihood, \( L_c(\Phi | X,Z) \), in terms of \(x\) and \(z\)
E- and M-steps

- Iterate the two steps
  1. E-step: Estimate $z$ given $X$ and current $\Phi$
  2. M-step: Find new $\Phi'$ given $z$, $X$, and old $\Phi$.

  $E$-step: $Q(\Phi | \Phi') = E \left[ \mathcal{L}_c(\Phi | X, Z) | X, \Phi' \right]$

  $M$-step: $\Phi'^{+1} = \operatorname{arg\,max}_\Phi Q(\Phi | \Phi')$

An increase in $Q$ increases incomplete likelihood

$$\mathcal{L}(\Phi'^{+1} | X) \geq \mathcal{L}(\Phi' | X)$$
EM in Gaussian Mixtures

- \( z_t^i = 1 \) if \( x^t \) belongs to \( G_i \), 0 otherwise (labels \( r_t^i \) of supervised learning); assume \( p(x^t | G_i) \sim N(\mu_i, \Sigma_i) \)

**E-step:**

\[
E[z_t^i | \mathcal{X}, \Phi'] = \frac{p(x^t | G_i, \Phi') p(G_i)}{\sum_j p(x^t | G_j, \Phi') p(G_j)} = p(G_i | x^t, \Phi') \equiv h_t^i
\]

**M-step:**

\[
P(G_i) = \frac{\sum_t h_t^i}{N} \quad m_t^{i+1} = \frac{\sum_t h_t^i x^t}{\sum_t h_t^i} \quad S_t^{i+1} = \frac{\sum_t h_t^i (x^t - m_t^{i+1}) (x^t - m_t^{i+1})}{\sum_t h_t^i}
\]

*Use estimated labels in place of unknown labels*
$P(G_1 | x) = h_1 = 0.5$
Mixtures of Latent Variable Models

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

\[
p(x_t | G_i) = \mathcal{N}(m_i, V_i V_i^T + \phi_i)
\]

Can use EM to learn \(V_i\) (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:

\[
p(x \mid C_i) = \sum_{j=1}^{k_i} p(x \mid G_{ij}) p(G_{ij})
\]

\[
p(x) = \sum_{i=1}^{K} p(x \mid C_i) p(C_i)
\]
Hierarchical Clustering

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

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

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

City-block distance

$$d_{cb}(x^r, x^s) = \sum_{j=1}^{d} |x^r_j - x^s_j|$$
Agglomerative Clustering

- Start with $N$ groups each with one instance and merge two closest groups at each iteration.
- Distance between two groups $G_i$ and $G_j$:
  - Single-link:
    $$ d(G_i, G_j) = \min_{x^r \in G_i, x^s \in G_j} d(x^r, x^s) $$
  - Complete-link:
    $$ d(G_i, G_j) = \max_{x^r \in G_i, x^s \in G_j} d(x^r, x^s) $$
  - 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)
- Manually check for meaning