CHAPTER 7:

CLUSTERING

Semiparametric Density Estimation

- Parametric: Assume a *single model* of data for $p(\boldsymbol{x}|C_i)$
  - Estimation of small # of parameters.
- Semiparametric: $p(\boldsymbol{x}|C_i)$ 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

3

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

where G_i the mixture components/groups/clusters,

$P(G_i)$: the mixture proportions (priors),

$p(\boldsymbol{x} \mid G_i)$: component densities,

k : the # of components, a hyperparameter, predefined.

Given a sample X and k, estimate the component densities and proportions.

Gaussian mixture where $p(\boldsymbol{x} \mid G_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and parameters $\Phi = \{P(G_i), \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^{k}$ from the unlabeled sample $X=\{\boldsymbol{x}^t\}_t$ (unsupervised learning)

Classes vs. Clusters

4

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

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

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

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

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

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

- Unsupervised : $X = \{\boldsymbol{x}^t\}_t$
- Clusters G_i $i=1,\ldots,k$

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

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

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

Labels $\boldsymbol{r}^t_i$?

k-Means Clustering

5

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

$$\|\mathbf{x}^t - \mathbf{m}_i\| = \min_j \|\mathbf{x}^t - \mathbf{m}_j\|$$

- Reconstruction error

$$E\left(\{\mathbf{m}_i\}_{i=1}^k | X\right) = \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|^2$$

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

- The best reference vectors are those that minimize the total reconstruction error.

k-Means Clustering

6

- b_i^t : the estimated label, depending on $\mathbf{m}_i$
- k-means clustering: the iterative algorithm for the estimation of b_i^t.
- With random $\mathbf{m}_i$, iterate
  1. Estimate b_i^t for all $\mathbf{x}^t$ – the estimated label (E-step)
  2. Minimize reconstruction error: $E\left(\{\mathbf{m}_i\}_{i=1}^k | X\right) = \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|^2$

     → update $\mathbf{m}_i = \frac{\sum_t b_i^t \mathbf{x}^t}{\sum_t b_i^t}$ (M-step)

  Until $\mathbf{m}_i$ converges.

  - The reference vector is set to the mean of all the instances that it represents.
  - Local search that highly depend on the initial $\mathbf{m}_i$.

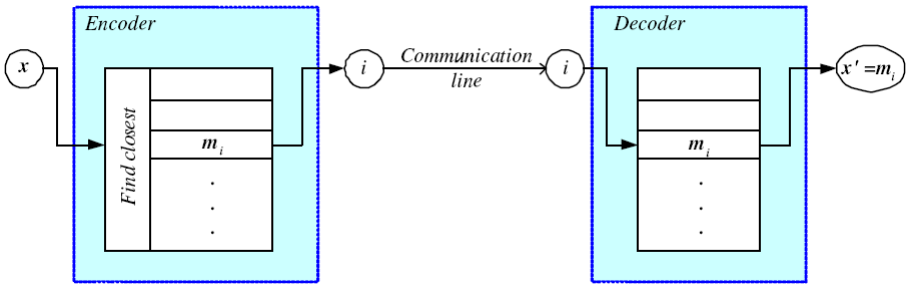
k-Means Clustering

7

- Initialization of $\mathbf{m}_i$:
  - Random selection of k instances,
  - Mean of all data + small random vector, or
  - K groups of the equal intervals from the Principal component → the means of groups
- Leader cluster algorithm:
  - An instance that is far away from existing centers create a new center there.
  - A center that covers a large # of instances ($\frac{\sum_t \boldsymbol{b}_i^t}{N} > \theta$) can be split into two.
  - A center of few instances can be removed and restarted from other part of X.
- The algorithm to find groups in the data by their centers.
- Application of clustering: vector quantization, preprocessing for classification/regression.

Encoding/Decoding

8



k-means Clustering

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

Repeat

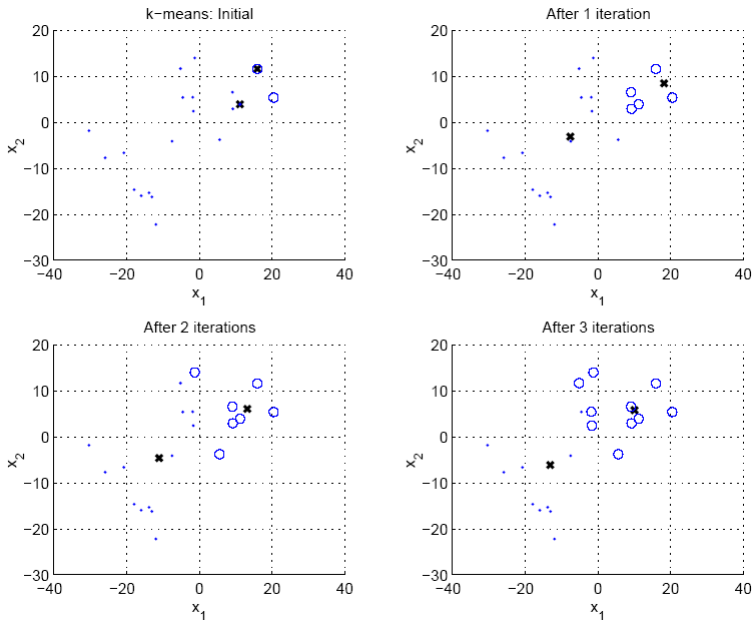
For all $\boldsymbol{x}^t \in \mathcal{X}$

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

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



Expectation-Maximization (EM)

11

- A probabilistic approach to find the component density parameters that maximize the likelihood of the sample.
- Log likelihood with a mixture model of $X=\{x^t\}_t$

$$L(\Phi|X) = \log \prod_t p(\mathbf{x}^t|\Phi)$$
$$= \sum_t log \sum_{i=1}^k p(\mathbf{x}^t|G_i)P(G_i)$$

where $\Phi = \{P(G_i), p(\mathbf{x}^t|G_i)\}_{i=1}^k$

- Assume hidden variables *z*, which when known, make optimization much simpler.
- Find the parameter vector Φ that maximizes the likelihood of the observed values of X, $L(\Phi|X)$, iteratively. If not feasible, $L_C(\Phi|X,Z)$.
- Complete likelihood, $L_C(\Phi|X,Z)$, in terms of ***x*** and ***z.***
- Incomplete likelihood, $L_C(\Phi|X)$, in terms of ***x.***

E- and M-steps

12

- Since the Z values are not observed, i.e. hidden, it's not directly complete data likelihood $L_C(\Phi|X,Z)$.
- Instead, work with its expectation, Q, given X and the current parameter values Φ^l. : Expectation step.
- Then, update the parameter values Φ^{l+1} that maximize $L_C(\Phi|X,Z)$. : Maximization step.

E- and M-steps

13

Iterate the two steps:

1. E(xpectation)-step:
   Estimate unknown *z* with the expectation *Q* of L_C, given X and current Φ^l.
2. M(aximization)-step:
   Find new Φ^{l+1} that maximize Q given *z*, X, and old Φ^l.

E-step: $Q(\Phi|\Phi^l) = E[L_C(\Phi|X,Z)|X,\Phi^l]$

M-step: $\Phi^{l+1} = \underset{\Phi}{\operatorname{argmax}}\, Q(\Phi|\Phi^l)$

An increase in Q increases incomplete likelihood

$$L(\Phi^{l+1}|X) \geq L(\Phi^l|X)$$

EM in Gaussian Mixtures

14

- E-step: Estimate the labels given current components.
- M-step: Update the component given the estimated label.
- Define a vector of indicator variables $\mathbf{z}^t = \{z_1^t, \ldots, z_k^t\}$ where $z^t_i = 1$ if $\boldsymbol{x}^t$ belongs to G_i ; 0 o.w. - (labels $\boldsymbol{r}^t_i$ of supv. learning); Assume $p(\boldsymbol{x}|G_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
- Since z is a multinomial distribution from *k* categories with their prior $P(G_i)$, $P(\mathbf{z}^t) = \prod_{i=1}^k P(G_i)^{z_i^t}$.
- Likelihood of x^t = its probability specified by the component:
  $p(\mathbf{x}^t|\mathbf{z}^t) = \prod_{i=1}^k p(\mathbf{x}^t|G_i)^{z_i^t}$
- Joint density: $p(\mathbf{x}^t, \mathbf{z}^t) = p(\mathbf{z}^t)p(\mathbf{x}^t|\mathbf{z}^t)$
- The complete data likelihood of iid X: $L_C(\Phi|X,Z) = \ldots.$

EM in Gaussian Mixtures

15

- The complete data likelihood of iid X:

$$\begin{aligned}\mathcal{L}_c(\Phi|X,Z) &= \log\prod_t p(x^t,z^t|\Phi)\\ &= \sum_t \log p(x^t,z^t|\Phi)\\ &= \sum_t \log P(z^t|\Phi) + \log p(x^t|z^t,\Phi)\\ &= \sum_t\sum_i z_i^t[\log\pi_i + \log p_i(x^t|\Phi)]\end{aligned}$$

where $\pi_i = P(G_i)$

- E-step: Estimate the labels given current components

Define $Q(\Phi|\Phi^l) = E[\log P(X,Z)\,|X,\Phi^l]$

$= E[L_C(\Phi|X,Z)|X,\Phi^l]$

$= \sum_t\sum_i E[z_i^t|X,\Phi^l][\log p(G_i) + \log p(x^t|G_i,\Phi^l)]$

EM in Gaussian Mixtures

16

- E-step: Estimate the labels given current components

Define $Q(\Phi|\Phi^l) = E[\log P(X,Z)\,|X,\Phi^l] = E[L_C(\Phi|X,Z)|X,\Phi^l]$

$= \sum_t\sum_i E[z_i^t|X,\Phi^l][\log p(G_i) + \log p(x^t|G_i,\Phi^l)]$

where

$$\begin{aligned}E[z_i^t|X,\Phi^l] &= E[z_i^t|x^t,\Phi^l] \quad x^t \text{ are iid}\\ &= P(z_i^t=1|x^t,\Phi^l) \quad z_i^t \text{ is a 0/1 random variable}\\ &= \frac{p(x^t|z_i^t=1,\Phi^l)P(z_i^t=1|\Phi^l)}{p(x^t|\Phi^l)} \quad \text{Bayes' rule}\\ &= \frac{p_i(x^t|\Phi^l)\pi_i}{\sum_j p_j(x^t|\Phi^l)\pi_j}\\ &= \frac{p(x^t|G_i,\Phi^l)P(G_i)}{\sum_j p(x^t|G_j,\Phi^l)P(G_j)}\\ &= P(G_i|x^t,\Phi^l) \equiv h_i^t\end{aligned}$$

i.e. the expected value of the hidden variable

= posterior probability that x^t is generated by component G_i.

EM in Gaussian Mixtures

17

- M-step: Maximize Q to get the next set of parameter Φ^{l+1}

$$\Phi^{l+1} = \underset{\Phi}{\operatorname{argmax}}\, Q(\Phi|\Phi^l) \qquad \text{where}$$

$$Q(\Phi|\Phi^l) = \sum_t \sum_i h_i^t [\log p(G_i) + \log p(x^t|G_i, \Phi^l)] \qquad P(G_i|x^t, \Phi^l) = h_i^t$$

$$= \sum_t \sum_i h_i^t \log p(G_i) + \sum_t \sum_i h_i^t \log p(x^t|G_i, \Phi^l)$$

$$\text{and } \sum_i p(G_i) = 1.$$

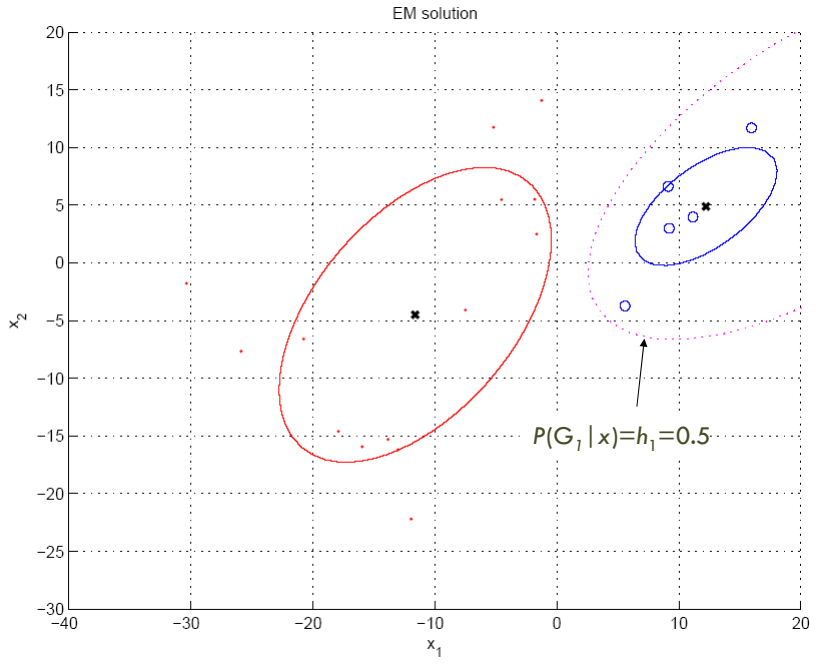
Solve it as the Lagrangian.

$$p(G_i) = \frac{\sum_t h_i^t}{N} \qquad \mathbf{m}_i^{l+1} = \frac{\sum_t h_i^t \mathbf{x}^t}{\sum_t h_i^t}$$

$$\mathbf{S}_i^{l+1} = \frac{\sum_t h_i^t (\mathbf{x}^t - \mathbf{m}_i^{l+1})(\mathbf{x}^t - \mathbf{m}_i^{l+1})^T}{\sum_t h_i^t} \quad \text{(if assume } p(\boldsymbol{x}^t|\Phi) \sim \mathrm{N}(\boldsymbol{m}_i, \mathbf{S}_i)\text{)}$$

where, for Gaussian components in e-step, we calculate

$$h_l^t = \frac{\pi_l |\mathbf{S}_l|^{-1/2} \exp[-(1/2)(\boldsymbol{x}^t - \boldsymbol{m}_l)^T \mathbf{S}_l^{-1} (\boldsymbol{x}^t - \boldsymbol{m}_l)]}{\sum_J \pi_J |\mathbf{S}_J|^{-1/2} \exp[-(1/2)(\boldsymbol{x}^t - \boldsymbol{m}_J)^T \mathbf{S}_J^{-1} (\boldsymbol{x}^t - \boldsymbol{m}_J)]}$$



18

Mixtures of Latent Variable Models

19

Regularize clusters

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

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

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

After Clustering

20

- 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

21

- 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

22

- 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:

$$p(\mathbf{x}|C_i) = \sum_{j=1}^{k_i} p(\mathbf{x}|G_{ij}) P(G_{ij})$$

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

Spectral Clustering

23

- Cluster using predefined pairwise similarities B_{rs} instead of using Euclidean or Mahalanobis distance
- Can be used even if instances not vectorially represented
- Steps:
  I. Use Laplacian Eigenmaps (chapter 6) to map to a new **z** space using B_{rs}
  II. Use *k*-means in this new **z** space for clustering

Hierarchical Clustering

24

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

  Minkowski (L_p) (Euclidean for p=2)

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

  City-block distance (Manhattan distance for p=2)

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

Agglomerative Clustering

25

- 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:

$$d(G_i, G_j) = \min_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

  - Complete-link:

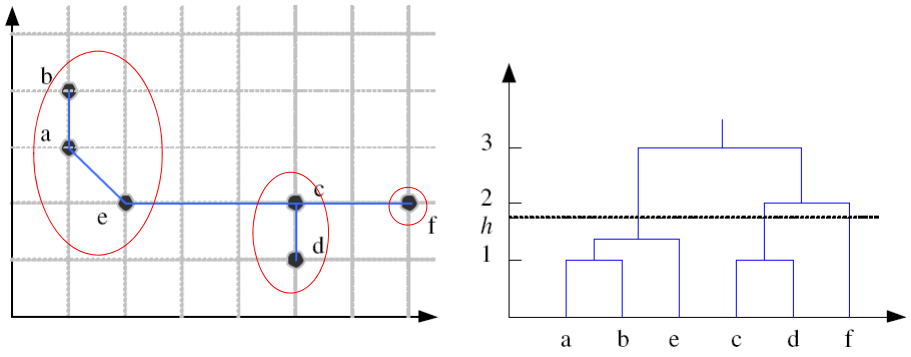
$$d(G_i, G_j) = \max_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

  - Average-link, centroid

$$d(G_i, G_j) = \text{ave}_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

Example: Single-Link Clustering

26



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