Mixture Densities

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

where G_i the components/groups/clusters,

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

$p(\boldsymbol{x} \mid \mathsf{G}_i)$ component densities

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

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

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

Classes vs. Clusters

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

$$p(\boldsymbol{x}) = \sum_{i=1}^{K} p(\boldsymbol{x} \mid \mathsf{C}_i) P(\mathsf{C}_i)$$

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

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

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

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

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

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

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

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

Labels, $\boldsymbol{r}^t_i$?

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 X and current Φ
2. M–step: Find new Φ' given z, X, and old Φ.

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

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

An increase in Q increases incomplete likelihood

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

EM in Gaussian Mixtures

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

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

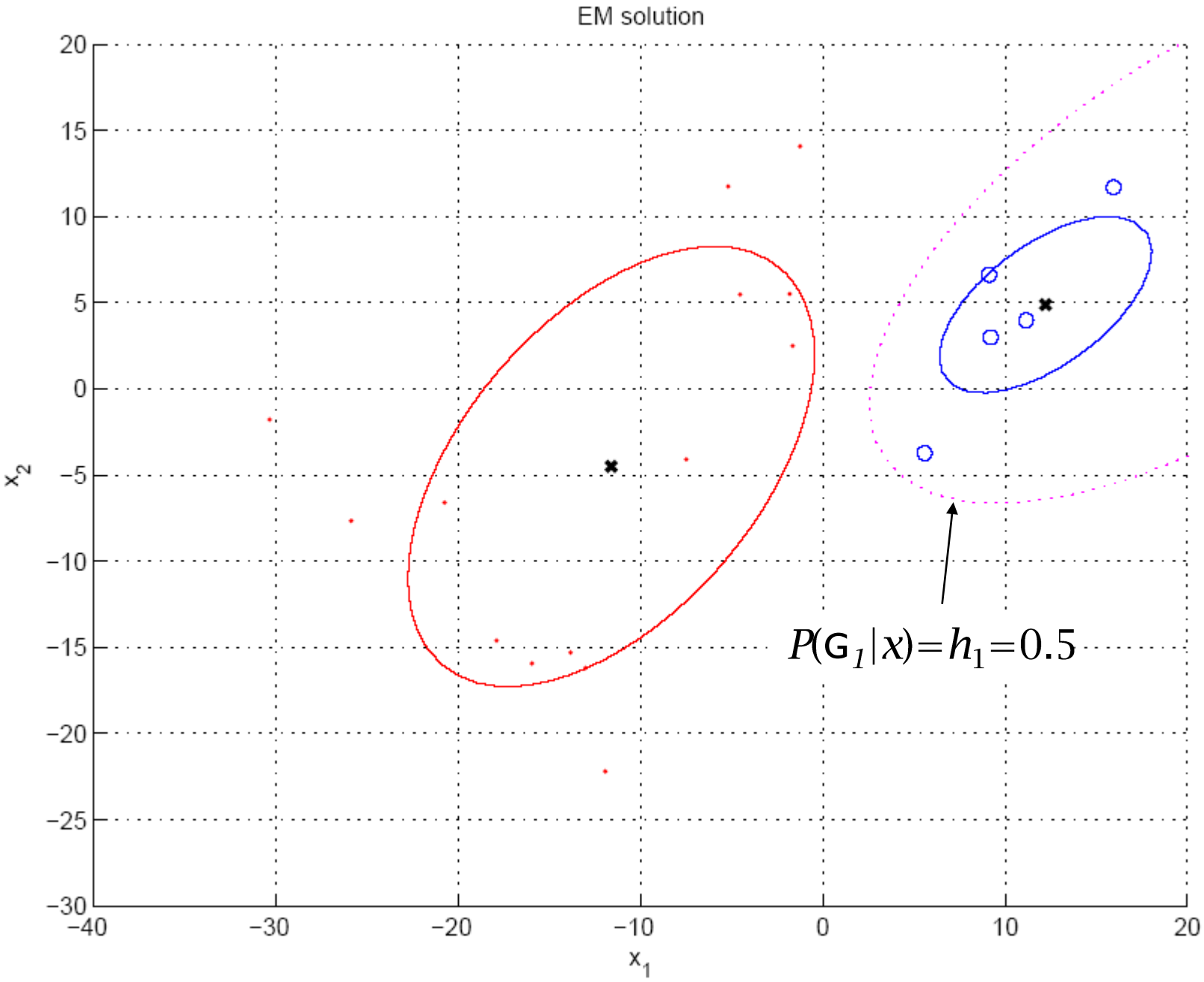
$$= P\left(\mathsf{G}_i | \boldsymbol{x}^t, \boldsymbol{\Phi}^l\right) \equiv h^t_i$$

- M-step:

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

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

Use estimated labels in place of unknown labels



Mixtures of Latent Variable Models

- Regularize clusters

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

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

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:

$$p(\boldsymbol{x} \mid \mathsf{C}_i) = \sum_{j=1}^{k_i} p(\boldsymbol{x} \mid \mathsf{G}_{ij}) P(\mathsf{G}_{ij})$$

$$p(\boldsymbol{x}) = \sum_{i=1}^{K} p(\boldsymbol{x} \mid \mathsf{C}_i) P(\mathsf{C}_i)$$