A.3 Mixture of Gaussians

The mixture of Gaussians (MoG) is a probability density model suitable for data $\mathbf{x}$ in D dimensions. The data is described as a weighted sum of K normal distributions

$$Pr(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \lambda_k \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k],$$

where $\boldsymbol{\mu}_{1\ldots K}$ and $\boldsymbol{\Sigma}_{1\ldots K}$ are the means and covariances of the normal distributions and $\lambda_{1\ldots K}$ are positive valued weights that sum to one.

The MoG can be expressed as the marginalization of the joint distribution $Pr(\mathbf{x}, h|\boldsymbol{\theta})$ between the data and a discrete hidden variable $h \in \{1\ldots K\}$,

$$\begin{aligned} Pr(\mathbf{x}|h, \boldsymbol{\theta}) &= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h] \\ Pr(h|\boldsymbol{\theta}) &= \text{Cat}_h[\boldsymbol{\lambda}]. \end{aligned}$$

This view of the MoG means that it can be fit using the EM algorithm

Algorithm 7: Maximum likelihood learning for mixtures of Gaussians

Input : Training data $\{\mathbf{x}_i\}_{i=1}^{I}$, number of clusters K
Output: ML estimates of parameters $\boldsymbol{\theta} = \{\lambda_{1\ldots K}, \boldsymbol{\mu}_{1\ldots K}, \boldsymbol{\Sigma}_{1\ldots K}\}$
begin
  Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ [a]
  repeat
    `// Expectation Step`
    for i=*1* **to** I **do**
      for k=*1* **to** K **do**
        $l_{ik} = \lambda_k \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$ `// numerator of Bayes' rule`
      end
      `// Compute posterior (responsibilities) by normalizing`
      $r_{ik} = l_{ik}/(\sum_{k=1}^{K} l_{ik})$
    end
    `// Maximization Step` [b]
    $\lambda_k^{[t+1]} = \sum_{i=1}^{I} r_{ik}/(\sum_{k=1}^{K}\sum_{i=1}^{I} r_{ik})$
    $\boldsymbol{\mu}_k^{[t+1]} = \sum_{i=1}^{I} r_{ik}\mathbf{x}_i/(\sum_{i=1}^{I} r_{ik})$
    $\boldsymbol{\Sigma}_k^{[t+1]} = \sum_{i=1}^{I} r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k^{[t+1]})(\mathbf{x}_i - \boldsymbol{\mu}_k^{[t+1]})^T/(\sum_{i=1}^{I} r_{ik})$.
    `// Compute Data Log Likelihood and EM Bound`
    $L = \sum_{i=1}^{I} \log\left[\sum_{k=1}^{K} \lambda_k \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]\right]$
    $B = \sum_{i=1}^{I}\sum_{k=1}^{K} r_{ik} \log\left[\lambda_k \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]/r_{ik}\right]$
  until *No further improvement in* L
end

[a] One possibility is to set the weights $\lambda_\bullet = 1/K$, the means $\boldsymbol{\mu}_\bullet$ to the values of K randomly chosen datapoints and the variances $\boldsymbol{\Sigma}_\bullet$ to the variance of the whole dataset. [b] For a diagonal covariance retain only the diagonal of the $\boldsymbol{\Sigma}_k$ update.