Machine Intelligence 2

5.3 Mixture Models and the EM-Algorithm

Prof. Dr. Klaus Obermayer

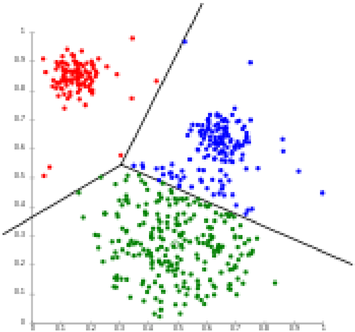
Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2018

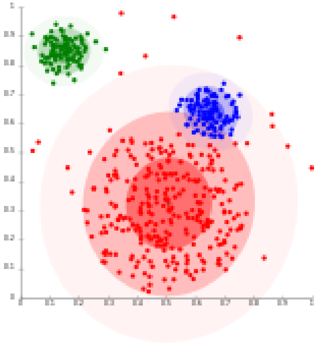
Mixture Models

Motivation

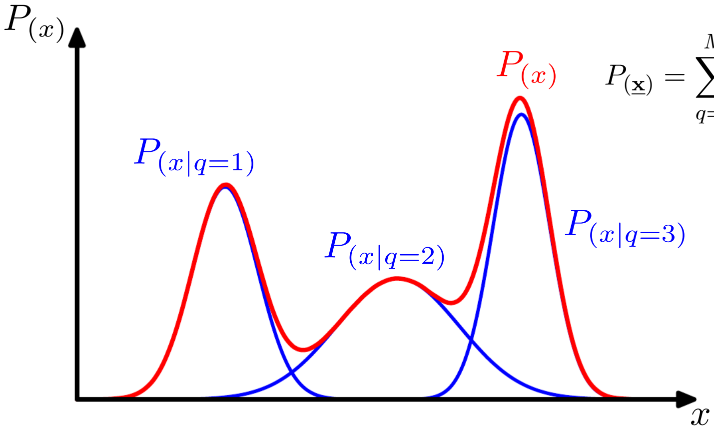
K-Means Clustering



Mixture of Gaussians



Parametric density estimation: Gaussian mixture model



$$P_{(\underline{\mathbf{x}})} = \sum_{q=1}^{M} P_{(\underline{\mathbf{x}}|q)} P_{(q)}$$

Component-based modeling of complex densities.
Source: Bishop, 2006 modified

Learning as model selection

data representation

$\downarrow$

model class

$\downarrow$

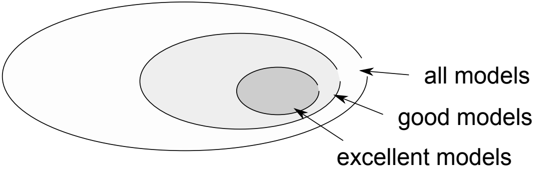
performance measure

$\downarrow$

optimization

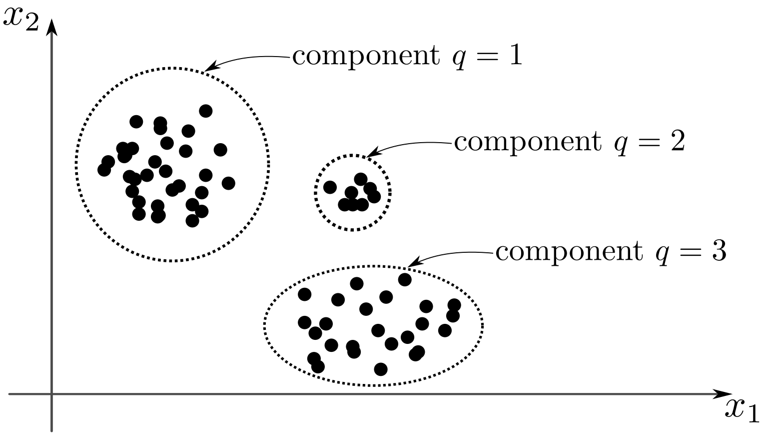
$\downarrow$

validation



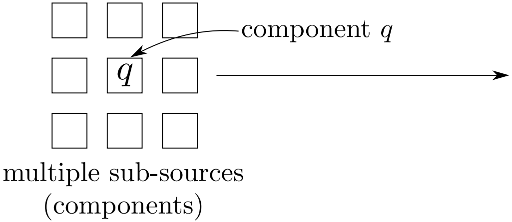
Data sources & representation

Data source $\rightarrow$ data $\underline{\mathbf{x}} \in \mathbb{R}^N \sim P_{(\underline{\mathbf{x}})}$



$\Rightarrow$ Assumption: Data is generated by multiple sources / classes.

Model class



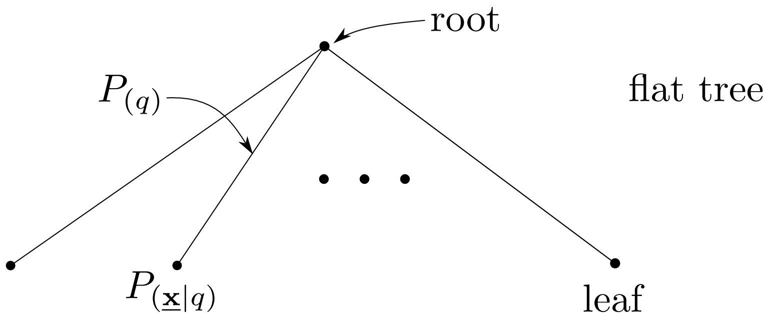
data $\underline{\mathbf{x}} \in \mathbb{R}^d \sim$

$$\underbrace{P_{(\underline{\mathbf{x}})} = \sum_{q=1}^{M} P_{(\underline{\mathbf{x}}|q)} P_{(q)}}_{\text{mixture model}}$$

$P_{(\underline{\mathbf{x}}|q)}$: components: probability density, that data point $\underline{\mathbf{x}}$ was created by component q

$P_{(q)}$: mixture parameters: probability, that component q creates a data point

Model class



$\rightarrow$ deeper trees possible: hierarchical mixture-models

$\rightarrow$ neural networks at the leaves: mixture of experts

Choice of basis functions

$$P_{(\underline{\mathbf{x}})} = \sum_{q=1}^{M} P_{(q)} P_{(\underline{\mathbf{x}}|q)}$$

$$P_{(\underline{\mathbf{x}}|q)} = \mathcal{N}\left(\underline{\mathbf{x}}; \underline{\mathbf{w}}_q, \sigma_q^2\right) = \frac{1}{(2\pi\sigma_q^2)^{N/2}} \exp\left\{-\frac{(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q)^2}{2\sigma_q^2}\right\}$$

$\rightsquigarrow$ (Gaussian mixture model)

parameters $P_{(q)}$, $\underline{\mathbf{w}}_q$ and σ_q must be determined for all components q

different basis functions are possible (problem specific)

Performance measure

Probability, that the dataset $\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}$ was generated by the model:

$$P_{\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}\right)} = \prod_{\alpha=1}^{p} P_{\left(\underline{\mathbf{x}}^{(\alpha)}\right)} = \prod_{\alpha=1}^{p} \left\{ \sum_{q=1}^{M} P_{\left(\underline{\mathbf{x}}^{(\alpha)}|q\right)} P_{(q)} \right\}$$

$$= \prod_{\alpha=1}^{p} \left\{ \sum_{q=1}^{M} \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}_q, \sigma_q^2\right) P_{(q)} \right\}$$

Principle of maximum likelihood:

$$P_{\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}\right)} \stackrel{!}{=} \max \quad \text{w.r.t. parameters}$$

Minimization of negative log-likelihood instead:

$$E^T = -\ln P_{\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}\right)} = -\sum_{\alpha=1}^{p} \ln \sum_{q=1}^{M} P_{\left(\underline{\mathbf{x}}^{(\alpha)}|q\right)} P_{(q)} \stackrel{!}{=} \min \quad \text{w.r.t parameters}$$

Relation to Soft-Clustering methods

Assumptions:

- Gaussian mixture model with M components
- same widths $\sigma_q^2 = \sigma^2 := \underbrace{\frac{1}{\beta}}_{\text{given}}$ for all basis functions
- same mixture parameters $P_{(q)} = \frac{1}{M}$

Cost function:

$$P_{(\underline{\mathbf{x}}^{(\alpha)})} = \sum_{q=1}^{M} \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}_q, \sigma^2\right) P_{(q)} = \frac{1}{M}\left(\frac{\beta}{2\pi}\right)^{N/2} \sum_{q=1}^{M} \exp\left\{-\frac{\beta}{2}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_q\right)^2\right\}$$

$$\begin{aligned} E^T = -\ln P_{(\{\underline{\mathbf{x}}^{(\alpha)}\})} &= -\ln \prod_{\alpha=1}^{p} \left\{\sum_{q=1}^{M} \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}_q, \sigma^2\right) P_{(q)}\right\} \\ &= -\sum_{\alpha=1}^{p} \ln \sum_{q=1}^{M} \exp\left\{-\frac{\beta}{2}\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)^2\right\} + \text{const}_{(\underline{\mathbf{w}}_q)} \end{aligned}$$

Relation to Soft-Clustering methods

Assignment probabilities:

$P_{(q|\underline{\mathbf{x}})}$: posterior probability of component q having generated a given data point $\underline{\mathbf{x}}$

$$P_{(q|\underline{\mathbf{x}})} = \frac{P_{(\underline{\mathbf{x}}|q)} P_{(q)}}{P_{(\underline{\mathbf{x}})}} \qquad \text{(Bayes' theorem)}$$

$\rightsquigarrow$ given the simplified Gaussian mixture model we obtain:

$$\begin{aligned} P_{(q|\underline{\mathbf{x}})} &= \frac{\left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2}\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)^2\right\} \cdot \frac{1}{M}}{\left(\frac{\beta}{2\pi}\right)^{N/2} \frac{1}{M} \sum_\gamma \exp\left\{-\frac{\beta}{2}\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_\gamma\right)^2\right\}} \\ &= \frac{\exp\left\{-\frac{\beta}{2}\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)^2\right\}}{\sum_{\gamma=1}^{M} \exp\left\{-\frac{\beta}{2}\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_\gamma\right)^2\right\}} \end{aligned}$$

$\Rightarrow$ assignment probability for Soft-Clustering

Relation to Soft-Clustering methods

$$E^T = - \sum_{\alpha=1}^{p} \ln \sum_{q=1}^{M} \exp\left\{ -\frac{\beta}{2} \left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q \right)^2 \right\} + \text{const}_{(\underline{\mathbf{w}}_q)}$$

Minimization of the cost function w.r.t. the weights:

$$\frac{\partial E^T}{\partial \underline{\mathbf{w}}_r} = - \sum_{\alpha=1}^{p} \frac{\exp\left\{ -\frac{\beta}{2} \left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_r \right)^2 \right\}}{\sum_{q=1}^{M} \exp\left\{ -\frac{\beta}{2} \left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q \right)^2 \right\}} \beta \left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_r \right) \stackrel{!}{=} 0$$

$$\underline{\mathbf{w}}_r = \frac{\sum_{\alpha=1}^{p} P_{(r|\underline{\mathbf{x}}^{(\alpha)})} \underline{\mathbf{x}}^{(\alpha)}}{\sum_{\alpha=1}^{p} P_{(r|\underline{\mathbf{x}}^{(\alpha)})}} \quad \rightsquigarrow \quad \text{center of mass condition for Soft-Clustering!}$$

$\Rightarrow$ Gaussian mixture model with components of equal size and strength is equivalent to Soft-Clustering

Relation to Soft-Clustering methods

New interpretation of Soft-Clustering:

- parameter estimation for a Gaussian mixture model with components of equal widths and strengths
- β is given
- implicit assumption: every cluster contains the same number of data points

Mixture models can be viewed as an extension of Soft-Clustering methods:

- clusters with different widths
- clusters with different number of data points

Optimization for Gaussian mixtures

Supporting equations:

$$\frac{\partial}{\partial \underline{\mathbf{w}}_q} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} = \frac{\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)}{\sigma_q^2} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} \qquad \frac{\partial}{\partial \sigma_q} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} = \left\{ -\frac{N}{\sigma_q} + \frac{\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)^2}{\sigma_q^3} \right\} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)}$$

Cost function: $\quad E^T = -\ln P_{\{\underline{\mathbf{x}}^{(\alpha)}\}} = -\sum_{\alpha=1}^{p} \ln \left(\sum_{q=1}^{M} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} P_{(q)} \right) \stackrel{!}{=} \min$

Minimization w.r.t. weights:

$$\frac{\partial E^T}{\partial \underline{\mathbf{w}}_r} = -\sum_{\alpha=1}^{p} \frac{\frac{\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_r\right)}{\sigma_r^2} P_{(\underline{\mathbf{x}}^{(\alpha)}|r)} P_{(r)}}{\sum_{q=1}^{M} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} P_{(q)}}$$

$$= \sum_{\alpha=1}^{p} \frac{\left(\underline{\mathbf{w}}_r - \underline{\mathbf{x}}^{(\alpha)}\right)}{\sigma_r^2} P_{(r|\underline{\mathbf{x}}^{(\alpha)})} \stackrel{!}{=} 0$$

$$\boxed{\underline{\mathbf{w}}_r = \frac{\sum_{\alpha=1}^{p} P_{(r|\underline{\mathbf{x}}^{(\alpha)})} \underline{\mathbf{x}}^{(\alpha)}}{\sum_{\alpha=1}^{p} P_{(r|\underline{\mathbf{x}}^{(\alpha)})}}}$$

$\Rightarrow$ mean of the data $\underline{\mathbf{x}}$ assigned to cluster r

Optimization for Gaussian mixtures

Supporting equations:

$$\frac{\partial}{\partial \underline{\mathbf{w}}_q} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} = \frac{\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)}{\sigma_q^2} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} \qquad \frac{\partial}{\partial \sigma_q} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} = \left\{ -\frac{N}{\sigma_q} + \frac{\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)^2}{\sigma_q^3} \right\} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)}$$

Cost function: $E^T = -\ln P_{\{\underline{\mathbf{x}}^{(\alpha)}\}} = -\sum_{\alpha=1}^{p} \ln \sum_{q=1}^{M} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} P_{(q)} \stackrel{!}{=} \min$

Minimization w.r.t. width of components:

$$\frac{\partial E^T}{\partial \sigma_r} = -\sum_{\alpha=1}^{p} \frac{\left\{ -\frac{N}{\sigma_r} + \frac{\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_r\right)^2}{\sigma_r^3} \right\} P_{(\underline{\mathbf{x}}^{(\alpha)}|r)} P_{(r)}}{\sum_{q=1}^{M} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} P_{(q)}}$$

$$= \frac{1}{\sigma_r} \sum_{\alpha=1}^{p} \left\{ N - \frac{\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_r\right)^2}{\sigma_r^2} \right\} P_{(r|\underline{\mathbf{x}}^{(\alpha)})} \stackrel{!}{=} 0$$

$$\boxed{\sigma_r^2 = \frac{1}{N} \frac{\sum_{\alpha=1}^{p} \left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_r\right)^2 P_{(r|\underline{\mathbf{x}}^{(\alpha)})}}{\sum_{\alpha=1}^{p} P_{(r|\underline{\mathbf{x}}^{(\alpha)})}}}$$

$\Rightarrow$ width of cluster: variance of data

Optimization for Gaussian mixtures

Cost function: $E^T = -\ln P_{\{\underline{\mathbf{x}}^{(\alpha)}\}} = -\sum_{\alpha=1}^{p} \ln \sum_{q=1}^{M} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} P_{(q)} \overset{!}{=} \min$

Minimization w.r.t. mixture parameters using Lagrange multipliers:

$$\frac{\partial}{\partial P_{(r)}} \left\{ E^T + \lambda \left(\sum_{q=1}^{M} P_{(q)} - 1 \right) \right\} \overset{!}{=} 0$$

$$= -\sum_{\alpha=1}^{p} \frac{P_{(\underline{\mathbf{x}}^{(\alpha)}|r)}}{\underbrace{\sum_{q=1}^{M} P_{(\underline{\mathbf{x}}^{(\alpha)}|q)} P_{(q)}}_{P_{(\underline{\mathbf{x}}^{(\alpha)})}}} + \lambda \overset{\frac{P_{(\underline{\mathbf{x}}|r)}}{P_{(\underline{\mathbf{x}})}} = \frac{P_{(r|\underline{\mathbf{x}})}}{P_{(r)}}}{=} -\sum_{\alpha=1}^{p} \frac{P_{(r|\underline{\mathbf{x}}^{(\alpha)})}}{P_{(r)}} + \lambda \overset{!}{=} 0$$

$P_{(r)} = \frac{1}{\lambda} \sum_{\alpha=1}^{p} P_{(r|\underline{\mathbf{x}}^{(\alpha)})}$, from $\sum_{r=1}^{M} P_{(r)} = p \overset{!}{=} 1$ follows $\lambda = p$ and

$$\boxed{P_{(r)} = \frac{1}{p} \sum_{\alpha=1}^{p} P_{(r|\underline{\mathbf{x}}^{(\alpha)})}}$$

$\Rightarrow$ "number" of data points per cluster (weighted by probability)

Optimization for Gaussian mixtures

Algorithm 1: Fixed-point iteration (Expectation-Maximization-algorithm)

initialization: $P_{(q)}^{\text{old}} = \frac{1}{M}, \quad \underline{\boldsymbol{\mu}} = \frac{1}{p}\sum_{\alpha=1}^{p} \underline{\mathbf{x}}^{(\alpha)}, \quad \underline{\mathbf{w}}_q^{\text{old}} = \underline{\boldsymbol{\mu}} + \underline{\eta}_q,$

$$(\sigma_q^2)^{\text{old}} = \frac{1}{p}\sum_{\alpha=1}^{p}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\boldsymbol{\mu}}\right)^2 + \underline{\varepsilon}_q, \quad \underline{\eta}_q, \underline{\varepsilon}_q\text{: small random vectors}$$

repeat

1. E-Step: Calculation of the assignment probabilities for $q = 1, \ldots, M$

$$P_{(q|\underline{\mathbf{x}}^{(\alpha)})} \overset{\text{Bayes}}{=} \frac{P_{(\underline{\mathbf{x}}^{(\alpha)}|q)}^{\text{old}} P_{(q)}^{\text{old}}}{\sum_{r=1}^{M} P_{(\underline{\mathbf{x}}^{(\alpha)}|r)}^{\text{old}} P_{(r)}^{\text{old}}} \qquad P_{(\underline{\mathbf{x}}^{(\alpha)}|q)}^{\text{old}} = \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}_q^{\text{old}}, (\sigma_q^2)^{\text{old}}\right)$$

2. M-Step: Calculation of the new parameter values for $q = 1, \ldots, M$

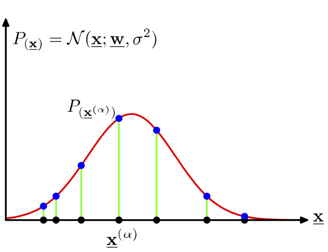
$$\underline{\mathbf{w}}_q^{\text{new}} = \frac{\sum_{\alpha=1}^{p} P_{(q|\underline{\mathbf{x}}^{(\alpha)})} \underline{\mathbf{x}}^{(\alpha)}}{\sum_{\alpha=1}^{p} P_{(q|\underline{\mathbf{x}}^{(\alpha)})}}$$

$$\left(\sigma_q^2\right)^{\text{new}} = \frac{1}{N} \frac{\sum_{\alpha=1}^{p}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_q^{\text{old}}\right)^2 P_{(q|\underline{\mathbf{x}}^{(\alpha)})}}{\sum_{\alpha=1}^{p} P_{(q|\underline{\mathbf{x}}^{(\alpha)})}}$$

$$P_{(q)}^{\text{new}} = \frac{1}{p}\sum_{\alpha=1}^{p} P_{(q|\underline{\mathbf{x}}^{(\alpha)})}$$

until *parameter values converge*

Degenerated solutions: "collapse" of components



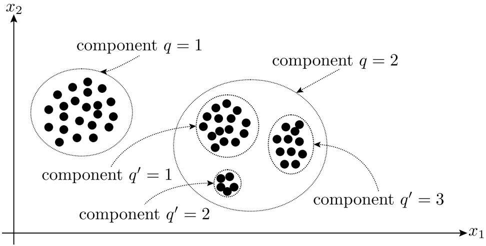
$$\mathcal{N}(\underline{\mathbf{x}}^{(\beta)}; \underline{\mathbf{w}}_q, \sigma_q^2) = \mathcal{N}(\underline{\mathbf{x}}^{(\beta)}; \underline{\mathbf{x}}^{(\beta)}, \sigma_q^2)$$

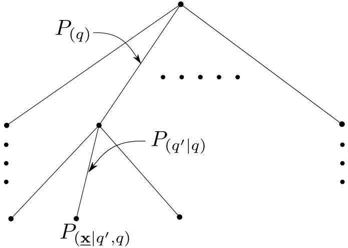
$$= \frac{1}{\left(2\pi\sigma_q^2\right)^{1/2}} \exp\left\{-\frac{\left(\underline{\mathbf{x}}^{(\beta)} - \underline{\mathbf{x}}^{(\beta)}\right)^2}{2\sigma_q^2}\right\} = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_q} \overset{\sigma_q^2 \to 0}{\longrightarrow} \infty$$

$\leadsto$ model validation (testset method) to detect overfitting

$\leadsto$ Maximum-a-posteriori instead of maximum likelihood approaches using a prior for each component which penalizes components with small variance σ_q^2

Hierarchical Gaussian mixtures





$$P_{(\underline{\mathbf{x}})} = \sum_q P_{(q)} \sum_{q'} P_{(q'|q)} P_{(\underline{\mathbf{x}}|q',q)}$$

Summary

Gaussian mixture model:

$$P_{(\underline{\mathbf{x}})} = \sum_{q=1}^{M} P_{(q)} P_{(\underline{\mathbf{x}}|q)}$$

$$P_{(\underline{\mathbf{x}}|q)} = \frac{1}{\left(2\pi\sigma_q^2\right)^{N/2}} \exp\left\{ -\frac{\left(\underline{\mathbf{x}} - \underline{\mathbf{w}}_q\right)^2}{2\sigma_q^2} \right\}$$

Maximum likelihood:

$$P_{\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}|\text{ parameter}\right)} \stackrel{!}{=} \max$$

Relation to Soft-Clustering:

- cost functions are identical, if: $\sigma_q^2 = \text{const.}_{(q)} = \frac{1}{\beta}$ and $P_{(q)} = \text{const.}_{(q)} = \frac{1}{M}$
- new interpretation of Soft-Clustering:
  - estimation of parameter $\underline{\mathbf{w}}_q$ of a Gaussian mixture model
  - β defines the size of the cluster $\widehat{=}$ resolution

Remarks

- equivalent solutions (permutation of components)
- improved initialization by application of the (much faster) K-means method:
  - prototypes $\rightsquigarrow$ component means $\underline{\mathbf{w}}_q$
  - intra-cluster spreads $\rightsquigarrow$ component variances σ_q^2
- extension to general Gaussian components ($\sigma_q^2 \to \underline{\boldsymbol{\Sigma}}_q$) straightforward (cf. Bishop 2006)
- mixture model: example of latent variable model

The Expectation-Maximization Algorithm

Latent variables

Example: mixture model

- observed data set $\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)} \in \mathbb{R}^N$
- every data point $\underline{\mathbf{x}}^{(\alpha)}$ is generated by one component $q = 1, \dots, M$
  - $\rightsquigarrow$ assignment variables: $\underline{\mathbf{m}}^{(\alpha)} = \left(m_1^{(\alpha)}, \dots, m_M^{(\alpha)}\right)^T \in \{0,1\}^M$

$$m_q^{(\alpha)} = \begin{cases} 1, & \text{if component } q \text{ has generated data point} \\ 0, & \text{otherwise} \end{cases} \qquad \sum_{q=1}^{M} m_q^{(\alpha)} = 1$$

- complete data set: $\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{m}}^{(1)} \dots, \underline{\mathbf{x}}^{(p)}, \underline{\mathbf{m}}^{(p)}$
- hidden / latent variables: $\underline{\mathbf{m}}^{(1)}, \dots, \underline{\mathbf{m}}^{(p)}$

Latent variable models and maximum likelihood

Calculation of the likelihood of the observed data requires marginalization of $P(\underline{\mathbf{x}}, \underline{\mathbf{m}}|\underline{\mathbf{w}})$:

$$P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}|\underline{\mathbf{w}}\right) \overset{iid}{=} \prod_{\alpha=1}^{p} P\left(\underline{\mathbf{x}}^{(\alpha)}|\underline{\mathbf{w}}\right) = \prod_{\alpha=1}^{p} \sum_{\underline{\mathbf{m}}} P\left(\underline{\mathbf{x}}^{(\alpha)}, \underline{\mathbf{m}}|\underline{\mathbf{w}}\right)$$

Log-likelihood is computationally costly to maximize / no closed-form solution due to sum in logarithm:

$$\ln P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}|\underline{\mathbf{w}}\right) = \sum_{\alpha=1}^{p} \ln\left(\sum_{\underline{\mathbf{m}}} P\left(\underline{\mathbf{x}}^{(\alpha)}, \underline{\mathbf{m}}|\underline{\mathbf{w}}\right)\right)$$

The Expectation-Maximization (EM) algorithm

Maximize the joint distribution over observed and latent variables (specifically useful if $P(\underline{\mathbf{x}}, \underline{\mathbf{m}} | \underline{\mathbf{w}})$ is from the exponential family: Gaussian, Bernoulli etc.)

$$\ln P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \underline{\mathbf{w}}\right) \stackrel{!}{=} \max_{(\underline{\mathbf{w}})}$$

Problem: values of the hidden variables are unknown.

The Expectation-Maximization (EM) algorithm

choose initial values for the parameters $\underline{\mathbf{w}}_{\text{old}}$ (e.g., by random) and tolerance θ

repeat

1. Evaluation of posterior distribution: $P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \,\middle|\, \left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \underline{\mathbf{w}}_{\text{old}}\right)$
2. E-Step: Compute expectation of complete data log-likelihood
   w.r.t posterior of $\left\{\underline{\mathbf{m}}^{(\alpha)}\right\}$

$$\mathcal{Q}(\underline{\mathbf{w}}, \underline{\mathbf{w}}_{\text{old}}) = \sum_{\left\{\underline{\mathbf{m}}^{(\alpha)}\right\}} P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \underline{\mathbf{w}}_{\text{old}}\right) \ln P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \underline{\mathbf{w}}\right)$$

3. M-Step: Determine new parameters that maximize the expectation

$$\underline{\mathbf{w}}_{\text{new}} = \arg\max_{(\underline{\mathbf{w}})} \mathcal{Q}(\underline{\mathbf{w}}, \underline{\mathbf{w}}_{\text{old}})$$

$$\underline{\mathbf{w}}_{\text{old}} \leftarrow \underline{\mathbf{w}}_{\text{new}}$$

until $|\underline{\mathbf{w}}_{old} - \underline{\mathbf{w}}_{new}| < \theta$

Remarks

- The EM algorithm converges to a local maximum of the log-likelihood function (cf. Bishop 2006)
- local optima (e.g., multimodal likelihood function) $\rightsquigarrow$ different initial conditions or simulated annealing methods
- EM is applicable to many latent variable problems: e.g., hidden Markov models, missing data situations
- EM is particularly efficient if the complete data distribution is from exponential family (log of exp)
- further extensions:
  - continuous latent variables (replace sums by integrals in marginalization / expectation)
  - maximum a posteriori estimation using a prior distribution $P_0(\underline{\mathbf{w}})$
  - non-tractable E- or M-steps: approximate inference or generalized EM-algorithms

Gaussian mixtures revisited

$$P(\underline{\mathbf{x}}) = \sum_{q=1}^{M} \rho(q)\mathcal{N}(\underline{\mathbf{x}}|\underline{\mathbf{w}}_q, \sigma_q^2) \stackrel{!}{=} \sum_{\underline{\mathbf{m}}} P(\underline{\mathbf{x}}, \underline{\mathbf{m}}) = \sum_{\underline{\mathbf{m}}} P(\underline{\mathbf{m}})P(\underline{\mathbf{x}}|\underline{\mathbf{m}})$$

mixture parameters: $\rho(q), \quad 0 \leq \rho(q) \leq 1, \sum_{q=1}^{M} \rho(q) = 1$

prior distribution of latent variables

$$P(\underline{\mathbf{m}}) = \prod_{q=1}^{M} \rho(q)^{m_q}$$

conditional distribution of the observed variables given the latent variables

$$P(\underline{\mathbf{x}}|\underline{\mathbf{m}}) = \prod_{q=1}^{M} \mathcal{N}^{m_q}(\underline{\mathbf{x}}|\underline{\mathbf{w}}_q, \sigma_q^2)$$

joint distribution

$$P(\underline{\mathbf{x}}, \underline{\mathbf{m}}) = P(\underline{\mathbf{x}}|\underline{\mathbf{m}}) \cdot P(\underline{\mathbf{m}}) = \prod_{q=1}^{M} \rho(q)^{m_q} \mathcal{N}^{m_q}(\underline{\mathbf{x}}|\underline{\mathbf{w}}_q, \sigma_q^2)$$

Gaussian mixtures & the EM algorithm

joint distribution: $P(\underline{\mathbf{x}}, \underline{\mathbf{m}}) = \prod_{q=1}^{M} \rho(q)^{m_q} \mathcal{N}^{m_q}(\underline{\mathbf{x}} | \underline{\mathbf{w}}_q, \sigma_q^2)$

likelihood:

$$P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \Big| \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}\right) = \prod_{\alpha=1}^{p} \prod_{q=1}^{M} \rho(q)^{m_q^{(\alpha)}} \mathcal{N}^{m_q^{(\alpha)}}(\underline{\mathbf{x}}^{(\alpha)} | \underline{\mathbf{w}}_q, \sigma_q^2)$$

log-likelihood:

$$\ln P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \Big| \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}\right) = \sum_{\alpha=1}^{p} \sum_{q=1}^{M} m_q^{(\alpha)} \left(\ln \rho(q) + \ln \mathcal{N}(\underline{\mathbf{x}}^{(\alpha)} | \underline{\mathbf{w}}_q, \sigma_q^2)\right)$$

log within sum & log of normal: much easier to handle

Gaussian mixtures & the EM algorithm

posterior distribution:

$$P(\underline{\mathbf{m}}|\underline{\mathbf{x}}) = \frac{P(\underline{\mathbf{x}}, \underline{\mathbf{m}})}{P(\underline{\mathbf{x}})} = \frac{\prod_{q=1}^{M} \left[\rho(q)\mathcal{N}(\underline{\mathbf{x}}|\underline{\mathbf{w}}_q, \sigma_q^2)\right]^{m_q}}{\sum_{q=1}^{M} \rho(q)\mathcal{N}(\underline{\mathbf{x}}|\underline{\mathbf{w}}_q, \sigma_q^2)}$$

posterior distribution of hidden data given observed:

$$\begin{aligned}
&P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\}\right) \\
&\overset{\text{iid. data}}{=} \prod_{\alpha=1}^{p} \frac{P\left(\underline{\mathbf{x}}^{(\alpha)}, \underline{\mathbf{m}}^{(\alpha)} \middle| \{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\}\right)}{P\left(\underline{\mathbf{x}}^{(\alpha)} \middle| \{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\}\right)} \\
&= \prod_{\alpha=1}^{p} \frac{\prod_{q=1}^{M} \left[\rho(q)\mathcal{N}(\underline{\mathbf{x}}^{(\alpha)}|\underline{\mathbf{w}}_q, \sigma_q^2)\right]^{m_q^{(\alpha)}}}{\sum_{q=1}^{M} \rho(q)\mathcal{N}(\underline{\mathbf{x}}^{(\alpha)}|\underline{\mathbf{w}}_q, \sigma_q^2)}
\end{aligned}$$

Gaussian mixtures & the EM algorithm

$$P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\}\middle|\left\{\underline{\mathbf{x}}^{(\alpha)}\right\},\left\{\underline{\mathbf{w}}_q,\sigma_q^2,\rho(q)\right\}\right)$$

$$= \quad \prod_{\alpha=1}^{p} \frac{\prod_{q=1}^{M}\left[\rho(q)\mathcal{N}(\underline{\mathbf{x}}^{(\alpha)}|\underline{\mathbf{w}}_q,\sigma_q^2)\right]^{m_q^{(\alpha)}}}{\sum_{q=1}^{M}\rho(q)\mathcal{N}(\underline{\mathbf{x}}^{(\alpha)}|\underline{\mathbf{w}}_q,\sigma_q^2)}$$

expected value under posterior:

$$\langle m_q^{(\alpha)}\rangle_{P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\}\middle|\left\{\underline{\mathbf{x}}^{(\alpha)}\right\},\left\{\underline{\mathbf{w}}_q,\sigma_q^2,\rho(q)\right\}\right)} = \textbf{ see blackboard}$$

$$= \frac{\rho(q)\mathcal{N}(\underline{\mathbf{x}}^{(\alpha)}|\underline{\mathbf{w}}_q,\sigma_q^2)}{\sum\limits_{r=1}^{M}\rho(r)\mathcal{N}(\underline{\mathbf{x}}^{(\alpha)}|\underline{\mathbf{w}}_r,\sigma_r^2)}$$

$$= \rho(q|\underline{\mathbf{x}}^{(\alpha)}) \text{ (from mixture EM-Algorithm)}$$

Gaussian mixtures & the EM algorithm

using this we can evaluate

$$
\begin{aligned}
&\mathcal{Q}\left(\left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}, \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{old}}\right) \\
&= \left\langle \ln P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}\right) \right\rangle_{P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{old}}\right)} \\
&= \left\langle \sum_{\alpha=1}^{p} \sum_{q=1}^{M} m_q^{(\alpha)} \left(\ln \rho(q) + \ln \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)} \middle| \underline{\mathbf{w}}_q, \sigma_q^2\right)\right) \right\rangle_{P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{old}}\right)} \\
&= \sum_{\alpha=1}^{p} \sum_{q=1}^{M} \left[\langle m_q^{(\alpha)} \rangle_{P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{old}}\right)} \cdot \left(\ln \rho(q) + \ln \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)} \middle| \underline{\mathbf{w}}_q, \sigma_q^2\right)\right) \right] \\
&= \sum_{\alpha=1}^{p} \sum_{q=1}^{M} \left[\rho\left(q \middle| \underline{\mathbf{x}}^{(\alpha)}\right) \cdot \left(\ln \rho(q) + \ln \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)} \middle| \underline{\mathbf{w}}_q, \sigma_q^2\right)\right) \right]
\end{aligned}
$$

$\rightsquigarrow$ E-step: calculation of

$$
\rho\left(q \middle| \underline{\mathbf{x}}^{(\alpha)}\right) = \frac{\rho(q)_{\text{old}} \cdot \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)} \middle| \underline{\mathbf{w}}_{q,\text{old}}, \sigma_{q,\text{old}}^2\right)}{\sum_{r=1}^{M} \rho(r)_{\text{old}} \cdot \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)} \middle| \underline{\mathbf{w}}_{r,\text{old}}, \sigma_{r,\text{old}}^2\right)}
$$

Gaussian mixtures & the EM algorithm

calculation of new parameters:

$$\left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{new}} = \underset{\left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}}{\operatorname{argmax}} \mathcal{Q}\left(\left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}, \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{old}}\right)$$

$$\mathcal{Q}\left(\left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}, \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{old}}\right) = \sum_{\alpha=1}^{p} \sum_{q=1}^{M} \left[\rho\left(q \middle| \underline{\mathbf{x}}^{(\alpha)}\right) \cdot \left(\ln \rho(q) + \ln \mathcal{N}\left(\underline{\mathbf{x}}^{(\alpha)} \middle| \underline{\mathbf{w}}_q, \sigma_q^2\right)\right)\right]$$

$$\left.\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \underline{\mathbf{w}}_q} = 0 \quad &\Rightarrow \quad \underline{\mathbf{w}}_{q,\text{new}} = \frac{\sum_{\alpha=1}^{p} \rho(q|\underline{\mathbf{x}}^{(\alpha)})\underline{\mathbf{x}}^{(\alpha)}}{\sum_{\alpha=1}^{p} \rho(q|\underline{\mathbf{x}}^{(\alpha)})} \\
\frac{\partial \mathcal{Q}}{\partial \sigma_q^2} = 0 \quad &\Rightarrow \quad \sigma_{q,\text{new}}^2 = \frac{1}{N} \frac{\sum_{\alpha=1}^{p} \left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_{q,\text{old}}\right)^2 \rho(q|\underline{\mathbf{x}}^{(\alpha)})}{\sum_{\alpha=1}^{p} \rho(q|\underline{\mathbf{x}}^{(\alpha)})} \\
\frac{\partial \mathcal{Q}}{\partial \rho(q)} = 0 \quad &\Rightarrow \quad \rho(q)_{\text{new}} = \frac{1}{p} \sum_{\alpha=1}^{p} \rho(q|\underline{\mathbf{x}}^{(\alpha)})
\end{aligned}\right\} \begin{array}{c}\text{expressions from}\\ \text{mixture EM-algorithm}\\ \text{recovered}\end{array}$$

$\rightsquigarrow$ M-step: optimal parameters for given

$$\langle m_q^{(\alpha)} \rangle_{P\left(\left\{\underline{\mathbf{m}}^{(\alpha)}\right\} \middle| \left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \left\{\underline{\mathbf{w}}_q, \sigma_q^2, \rho(q)\right\}_{\text{old}}\right)} = \rho\left(q \middle| \underline{\mathbf{x}}^{(\alpha)}\right)$$