

CS195-5 : Introduction to Machine Learning

Lecture 20

Greg Shakhnarovich

November 1, 2006

Announcements

Review

- Locally-weighted regression:
Fit a parametric model to neighbors, weighted by a kernel.

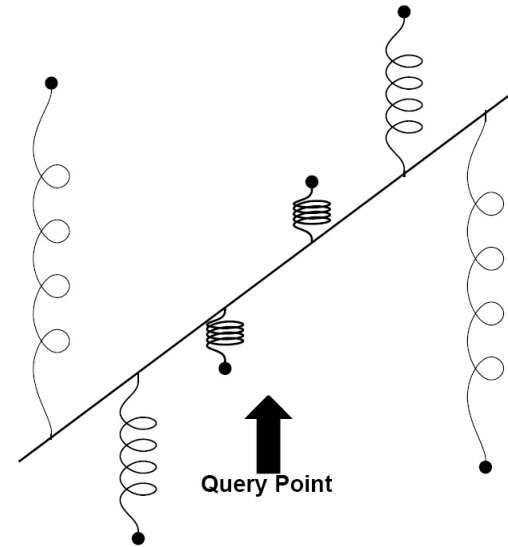
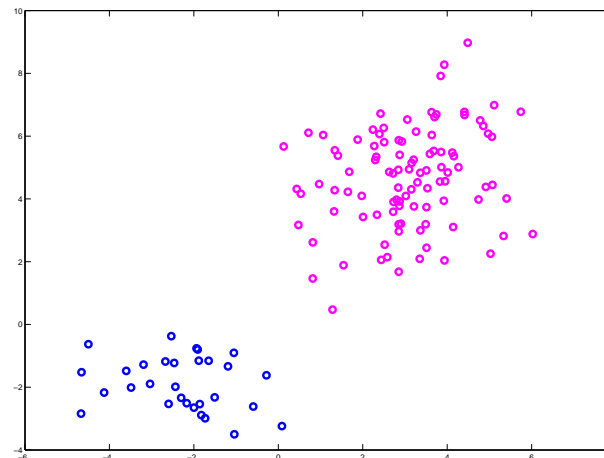


Figure 5: Weighted springs.

- Parametric mixture models:

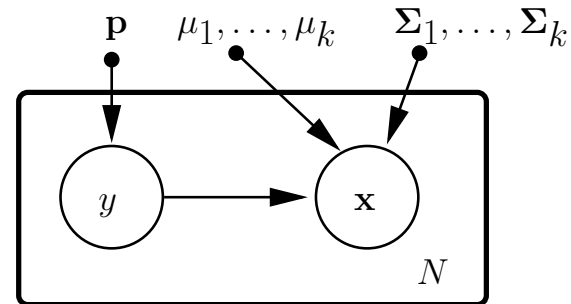
$$p(\mathbf{x}; \mathbf{p}) = \sum_{c=1}^k p(y = c)p(\mathbf{x} | y = c).$$



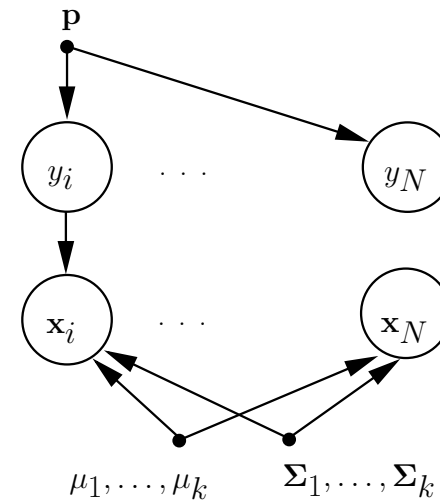
Generative model for a Gaussian mixture

$$p(\mathbf{x}; \theta, \mathbf{p}) = \sum_{c=1}^k p_c \cdot \mathcal{N}(\mathbf{x}; \mu_c, \Sigma_c),$$

- The graphical model



- The *plate* notation is a shorthand for



Plan for today

- The Expectation Maximization (EM) algorithm.

Likelihood of a mixture model

- The log-likelihood of \mathbf{p}, θ :

$$\log p(X_N; \mathbf{p}, \theta) = \sum_{i=1}^N \log \sum_{c=1}^k p_c \mathcal{N}(\mathbf{x}_i; \mu_c, \Sigma_c).$$

- No closed-form solution because of the sum inside log.
 - Since we need to take into account all possible components that could have generated \mathbf{x}_i .

Mixture density estimation

- Suppose that we do observe $y_i \in \{1, 2\}$ for each $i = 1, \dots, N$.
- Let us introduce a set of binary *indicator variables* $\mathbf{z}_i = [z_{i1}, \dots, z_{ik}]$ where

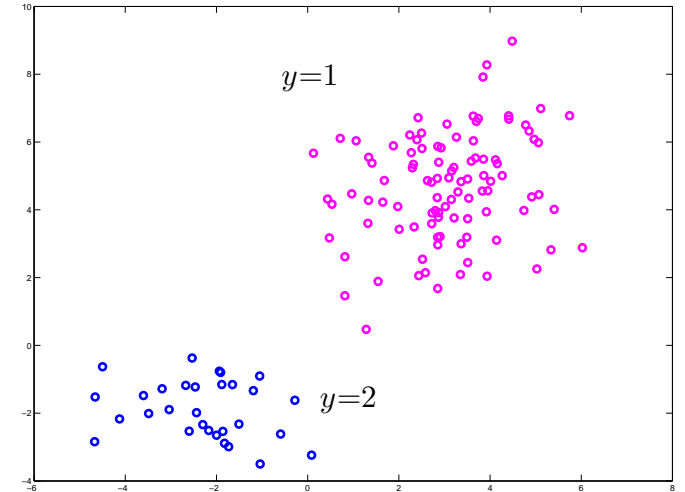
$$z_{ic} = 1 = \begin{cases} 1 & \text{if } y_i = c, \\ 0 & \text{otherwise.} \end{cases}$$

- The count of examples from c -th component:

$$N_c = \sum_{i=1}^N z_{ic}.$$

Mixture density estimation: known labels

- If we know \mathbf{z}_i , the ML estimates of the Gaussian components, just like in class-conditional model, are



$$\hat{p}_c = \frac{N_c}{N},$$

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i=1}^N z_{ic} \mathbf{x}_i,$$

$$\hat{\Sigma}_c = \frac{1}{N_c} \sum_{i=1}^N z_{ic} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T.$$

Credit assignment

- When we don't know \mathbf{z} , we face a *credit assignment* problem: which component is responsible for \mathbf{x}_i ?
- Suppose we do know component parameters $\theta = [\mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k]$ and mixing probabilities $\mathbf{p} = [p_1, \dots, p_k]$.
- The *posterior* of indicators using Bayes' theorem:

$$\gamma_{ic} \triangleq p(z_{ic} | \mathbf{x}; \theta, \mathbf{p}) = \frac{p_c \cdot p(\mathbf{x}; \mu_c, \Sigma_c)}{\sum_{l=1}^k p_l \cdot p(\mathbf{x}; \mu_l, \Sigma_l)}$$

- We will call γ_{ic} the *responsibility* of the c -th component for \mathbf{x} .
 - Note: $\sum_{c=1}^k \gamma_{ic} = 1$.

Expected likelihood

- The “complete data” likelihood (when \mathbf{z} are known):

$$p(X_N, Z_N; \mathbf{p}, \theta) \propto \prod_{i=1}^N \prod_{c=1}^k (p_c \mathcal{N}(\mathbf{x}_i; \mu_c, \Sigma_c))^{z_{ic}}.$$

and the log:

$$\log p(X_N, Z_N; \mathbf{p}, \theta) = \text{const} + \sum_{i=1}^N \sum_{c=1}^k z_{ic} (\log p_c + \log \mathcal{N}(\mathbf{x}_i; \mu_c, \Sigma_c)).$$

- We can't compute it, but can take the *expectation* w.r.t. the posterior of \mathbf{z} :

$$E_{z_{ic} \sim p(\cdot | X_N, \theta, \mathbf{p})} [\log p(X_N, Z_N; \mathbf{p}, \theta)].$$

Expected likelihood

- Expectation of z_{ic} :

$$E_{z_{ic} \sim p(\cdot | X_N, \theta, \mathbf{p})} [z_{ic}] = \sum_{z \in \{0,1\}} z p(z_{ic} = z | \mathbf{x}_i; \theta, \mathbf{p}) = \gamma_{ic}.$$

- The expected likelihood of the data:

$$E_{z_{ic} | \mathbf{x}_i, \theta, \mathbf{p}} [\log p(X_N, Z_N; \mathbf{p}, \theta)] = \text{const} \\ + \sum_{i=1}^N \sum_{c=1}^k \gamma_{ic} (\log p_c + \log \mathcal{N}(\mathbf{x}_i; \mu_c, \Sigma_c)).$$

Expectation maximization

$$E_{z_{ic}|\mathbf{x}_i,\theta,\mathbf{p}} [\log p(X_N, Z_N; \mathbf{p}, \theta)] = \sum_{i=1}^N \sum_{c=1}^k \gamma_{ic} (\log p_c + \log \mathcal{N}(\mathbf{x}_i; \mu_c, \Sigma_c)).$$

- We can find \mathbf{p}, θ that maximize this *expected* likelihood – by setting derivatives to zero and, for \mathbf{p} , using Lagrange multipliers to enforce $\sum_c p_c = 1$.

$$\hat{p}_c = \frac{\sum_{i=1}^N \gamma_{ic}}{N},$$

$$\hat{\mu}_c = \frac{1}{\sum_{i=1}^N \gamma_{ic}} \sum_{i=1}^N \gamma_{ic} \mathbf{x}_i,$$

$$\hat{\Sigma}_c = \frac{1}{\sum_{i=1}^N \gamma_{ic}} \sum_{i=1}^N \gamma_{ic} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T.$$

Summary so far

- If we know the **parameters** and **indicators** (assignments) we are done.
- If we know the **indicators** but not the parameters, we can do ML estimation of the parameters – and we are done.
- If we know the **parameters** but not the indicators, we can compute the posteriors of indicators;
 - With known posteriors, we can estimate parameters that maximize the *expected* likelihood – and then we are done.
- But in reality we know neither the parameters nor the indicators.

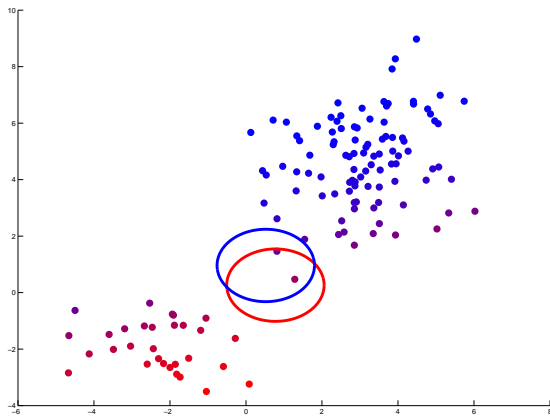
The EM algorithm

- Start with a guess of θ, \mathbf{p} .
 - Typically, random θ and $p_c = 1/k$.
- Iterate between:
 - E-step** Compute values of expected assignments, i.e. calculate γ_{ic} , using current estimates of θ, \mathbf{p} .
 - M-step** Maximize the *expected* likelihood, under current γ_{ic} .

EM for Gaussian mixture: an example

- Colors represent γ_{ic} after the E-step.

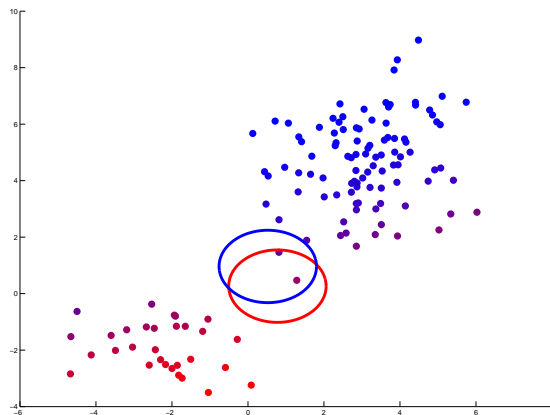
1st iteration



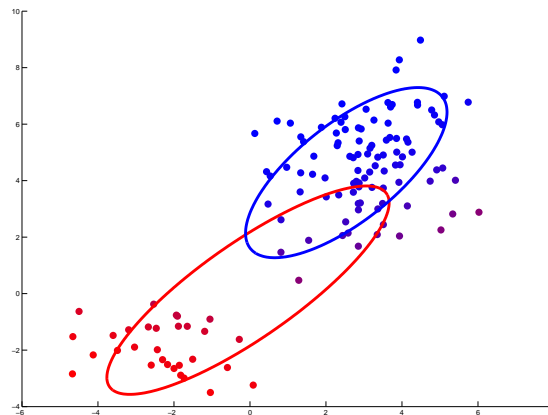
EM for Gaussian mixture: an example

- Colors represent γ_{ic} after the E-step.

1st iteration



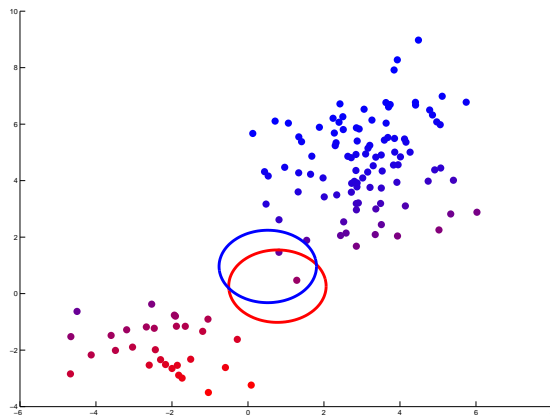
2nd iteration



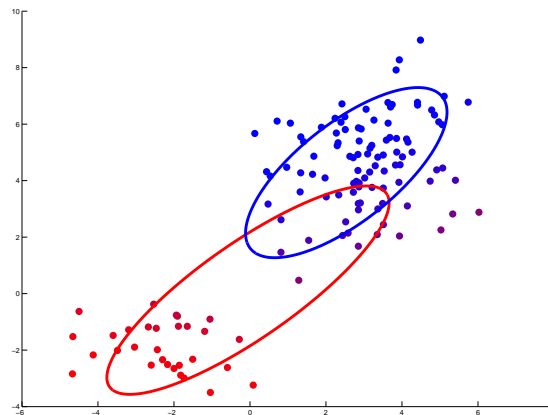
EM for Gaussian mixture: an example

- Colors represent γ_{ic} after the E-step.

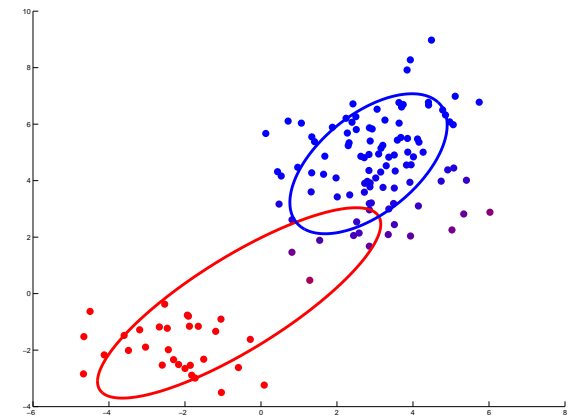
1st iteration



2nd iteration



3rd iteration

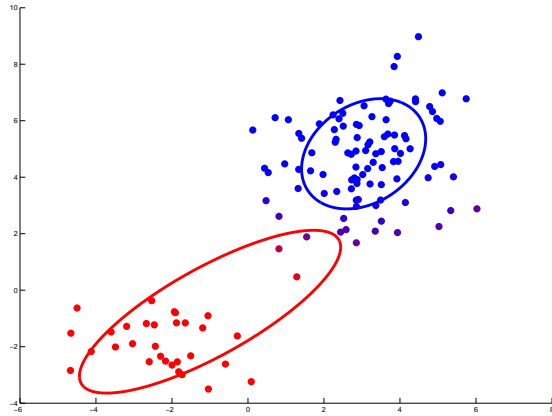


EM for Gaussian mixture: an example

4th iteration

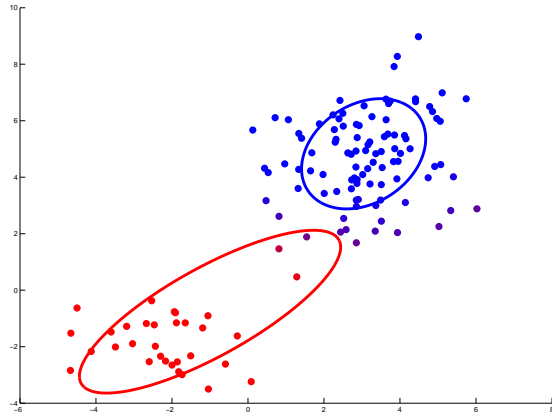
7th iteration

9th iteration

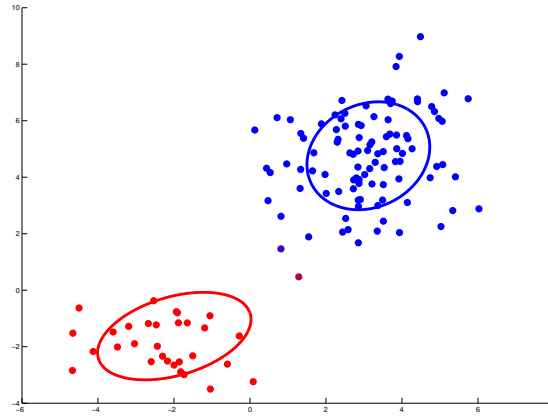


EM for Gaussian mixture: an example

4th iteration



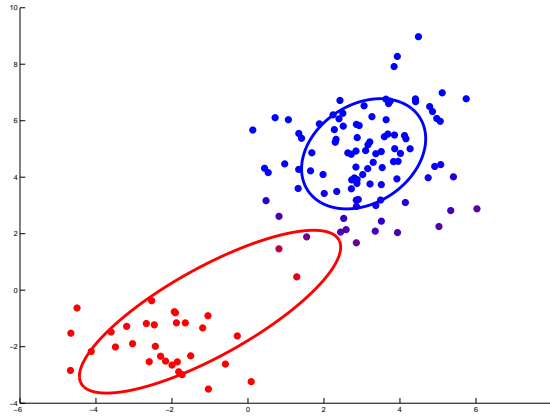
7th iteration



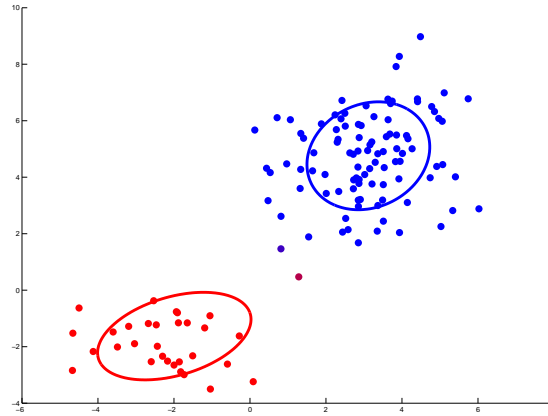
9th iteration

EM for Gaussian mixture: an example

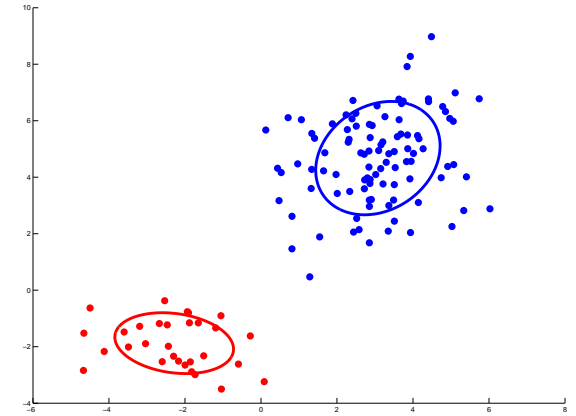
4th iteration



7th iteration



9th iteration



The EM for Gaussian mixtures- summary

- Initialize: random μ_c^{old} , Σ_c^{old} , $p_c^{old} = 1/k$ for $c = 1, \dots, k$.
- Iterate until convergence:

E-step estimate responsibilities:

$$\gamma_{ic} = \frac{p_c^{old} \mathcal{N}(\mathbf{x}_i; \mu_c^{old}, \Sigma_c^{old})}{\sum_{l=1}^k p_l^{old} \mathcal{N}(\mathbf{x}_i; \mu_l^{old}, \Sigma_l^{old})}$$

M-step re-estimate mixture parameters:

$$\hat{p}_c^{new} = \frac{\sum_{i=1}^N \gamma_{ic}}{N},$$

$$\hat{\mu}_c^{new} = \frac{1}{\sum_{i=1}^N \gamma_{ic}} \sum_{i=1}^N \gamma_{ic} \mathbf{x}_i,$$

$$\hat{\Sigma}_c^{new} = \frac{1}{\sum_{i=1}^N \gamma_{ic}} \sum_{i=1}^N \gamma_{ic} (\mathbf{x}_i - \hat{\mu}_c^{new})(\mathbf{x}_i - \hat{\mu}_c^{new})^T.$$

Next time

More on the EM algorithm.
Model selection.