

CS195-5 : Introduction to Machine Learning

Lecture 21

Greg Shakhnarovich

November 3, 2006

Announcements

Review: the EM algorithm for Gaussian mixtures

- 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:

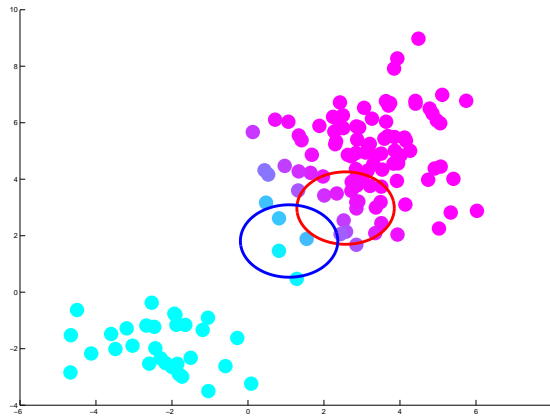
$$\hat{p}_c^{new} = \frac{1}{N} \sum_{i=1}^N \gamma_{ic},$$

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

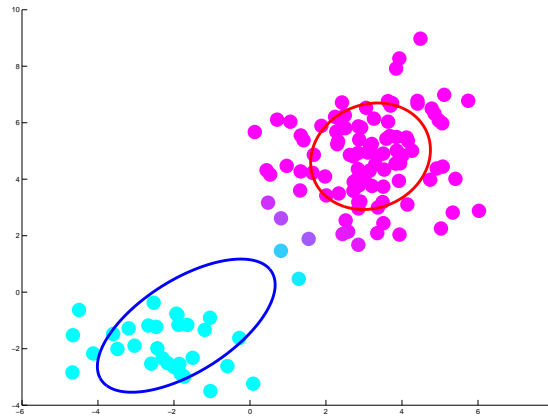
$$\hat{\Sigma}_c^{new} = \frac{1}{\sum_{i=1}^N \gamma_{ic}} \sum_{i=1}^N \gamma_{ic} (\mathbf{x}_i - \hat{\mu}_c^{new})(\mathbf{x}_i - \hat{\mu}_c^{new})^T.$$

EM for Gaussian mixture: an example

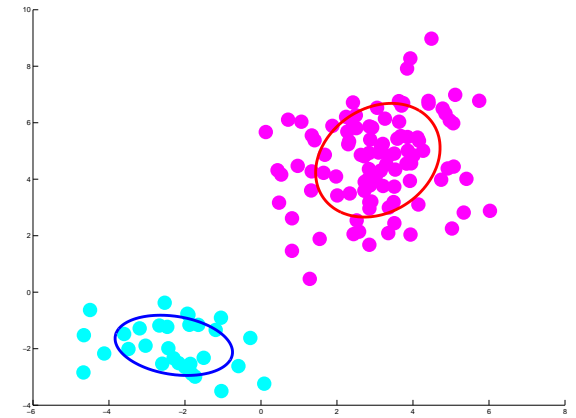
1st iteration



3rd iteration



7th (final) iteration



Plan for today

- EM in general
- Practical aspects
- Model selection

Generic EM for mixture models

- General mixture models: $p(\mathbf{x}) = \sum_{c=1}^k p_c p(\mathbf{x}; \theta_c)$
- Initialize \mathbf{p} , θ^{old} , and iterate until convergence:

E-step: compute responsibilities

$$\gamma_{ic} = \frac{p_c^{old} p(\mathbf{x}_i; \theta_c^{old})}{\sum_{l=1}^k p_l^{old} p(\mathbf{x}_i; \theta_l^{old})}.$$

M-step: re-estimate mixture parameters:

$$\mathbf{p}^{new}, \theta^{new} = \operatorname{argmax}_{\theta, \mathbf{p}} \sum_{i=1}^N \sum_{c=1}^k \gamma_{ic} (\log p_c + \log p(\mathbf{x}_i; \theta_c)).$$

The EM algorithm in general

- Observed data X , hidden variables Z .
 - E.g., *missing data*.
- Complete data log-likelihood: $\ell(X, Z; \theta)$
- Initialize θ^{old} , and iterate until convergence:

E-step: Compute the expected likelihood as a function of θ .

$$Q(\theta; \theta^{old}) = E_{p(Z | X, \theta^{old})} [\ell(X, Z; \theta) | X, \theta^{old}]$$

M-step: Compute

$$\theta^{new} = \underset{\theta}{\operatorname{argmax}} Q(\theta; \theta^{old}).$$

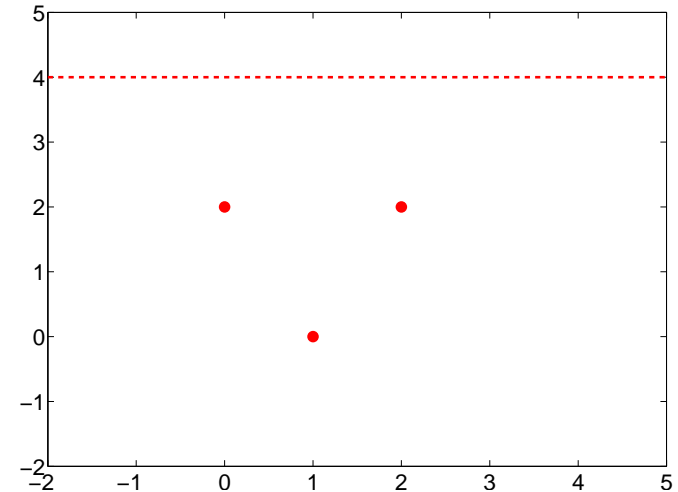
EM for missing data

- Suppose *some* of the data is missing.
- Examples:
 - A mixture of 4 Gaussians, two of which are given
 - A set of points with some coordinates in some points missing.
- E-step: compute posterior and expectation only over missing data.
- M-step: maximize likelihood over *complete* data, i.e. observed and missing.

Example: EM for missing data

- Four data points:

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} * \\ 4 \end{bmatrix}.$$



- Model: single Gaussian with diagonal covariance; $\theta = [\mu_1, \mu_2, \sigma_1^2, \sigma_2^2]^T$.
- The expected log-likelihood, under guessed θ^{old} :

$$Q(\theta; \theta^{old}) = \sum_{i=1}^3 \log \mathcal{N}(\mathbf{x}_i; \theta) + \int_{-\infty}^{\infty} \log p\left(\begin{bmatrix} x \\ 4 \end{bmatrix} \middle| \theta\right) \frac{p\left(\begin{bmatrix} x \\ 4 \end{bmatrix} \middle| \theta^{old}\right)}{\int_{-\infty}^{\infty} p\left(\begin{bmatrix} x' \\ 4 \end{bmatrix} \middle| \theta^{old}\right) dx'} dx$$

Example: EM for missing data (continued)

- Initial guess: $\theta^{old} = [0, 0, 1, 1]^T$ i.e., $\mu = \mathbf{0}$, $\Sigma = \mathbf{I}$.
- After some calculus:

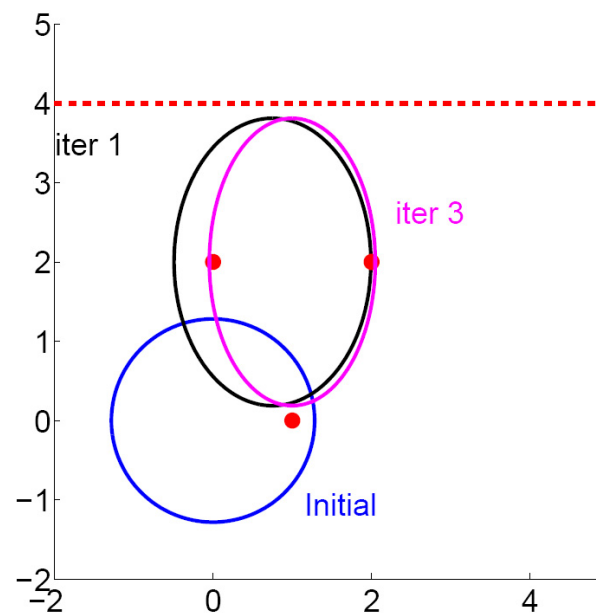
$$Q(\theta; \theta^{old}) = \sum_{i=1}^3 \mathcal{N}(\mathbf{x}_i; \mathbf{0}, \mathbf{I}) - \frac{1 + \mu_1^2}{2\sigma_1^2} - \frac{(4 - \mu_2)^2}{2\sigma_2^2} - \log 2\pi\sigma_1\sigma_2.$$

- Maximizing w.r.t. θ we get

$$\theta^{new} = [0.75, 2, 0.938, 2]^T$$

- After three iterations, converges to

$$\theta = [1, 2, .667, 2]^T$$



Why does EM work?

- Ultimately, we want to maximize likelihood of the *observed* data

$$\theta^* = \operatorname{argmax}_{\theta} \log p(X; \theta).$$

- Let $\ell^{(t)}$ be $\log p(X; \theta^{new})$ after t iterations.
- Can show (not today):

$$\ell^{(0)} \leq \ell^{(1)} \leq \dots \leq \ell^{(t)} \dots$$

EM and maximum likelihood

$$\ell^{(0)} \leq \ell^{(1)} \leq \dots \leq \ell^{(t)} \dots$$

- The idea of the proof:
 - in each iteration, E-step computes $Q(\theta; \theta^{old})$ which is a lower bound on $\log p(X; \theta)$;
 - M-step maximizes (“saturates”) that lower bound.
 - In the subsequent E-step, the new bound is at least as high as the previous one.
- I.e., EM monotonically increases the likelihood of the observed data.
 - as long as $\log p(X; \theta^{new}) < \infty$ EM necessarily converges – **but** possibly to a local maximum!

EM and maximum likelihood

$$\ell^{(0)} \leq \ell^{(1)} \leq \dots \leq \ell^{(t)} \dots$$

- The idea of the proof:
 - in each iteration, E-step computes $Q(\theta; \theta^{old})$ which is a lower bound on $\log p(X; \theta)$;
 - M-step maximizes (“saturates”) that lower bound.
 - In the subsequent E-step, the new bound is at least as high as the previous one.
- I.e., EM monotonically increases the likelihood of the observed data.
 - as long as $\log p(X; \theta^{new}) < \infty$ EM necessarily converges – **but** possibly to a local maximum!
 - One popular solution: restart a number of times (with different initializations), and choose the run with highest $\log p(X; \theta)$.

Caveat: log-sum computation

- Back to mixture models: need to compute $\log p(\mathbf{x} | \theta) = \log \sum_{c=1}^2 p_c p(\mathbf{x}; \theta_c)$
- Warning: underflow (especially in high-dimensional spaces)

```
x=[-2000 -2006]; log(sum(exp(x)))
```

```
Warning: Log of zero.
```

```
ans =  
-Inf
```

- Observation: $a = \log[\exp(a)] = \log[\exp(a + B)] - B$.
- Set a constant B so that the highest $p(\mathbf{x}; \theta_c)$ saturates positive precision on your machine.

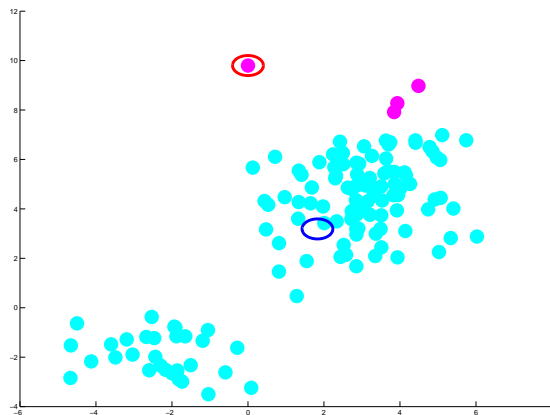
```
B=700-max(logp); log(sum(exp(x+B)))-B
```

```
ans =  
-2.0000e+003
```

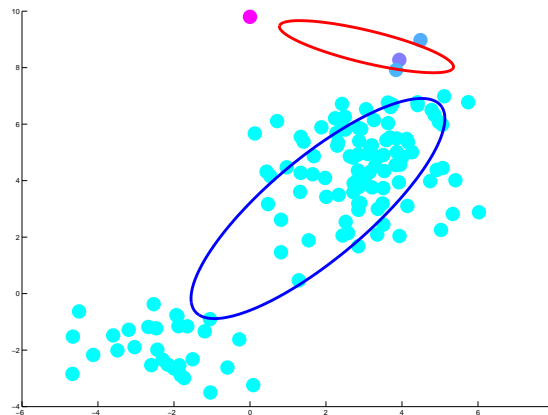
EM and overfitting

- We can be very unlucky with the initial guess.

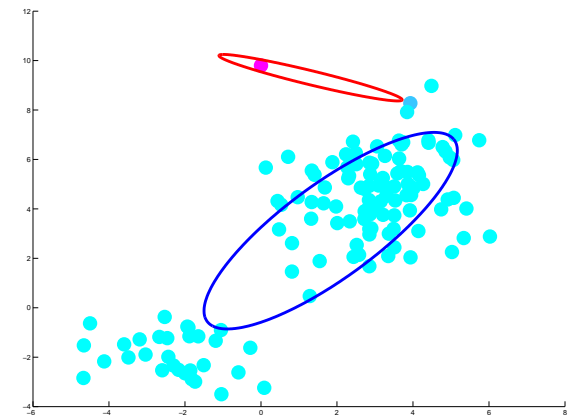
1st iteration



2nd



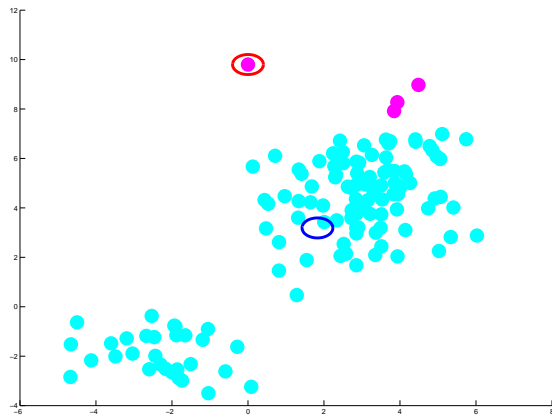
4th



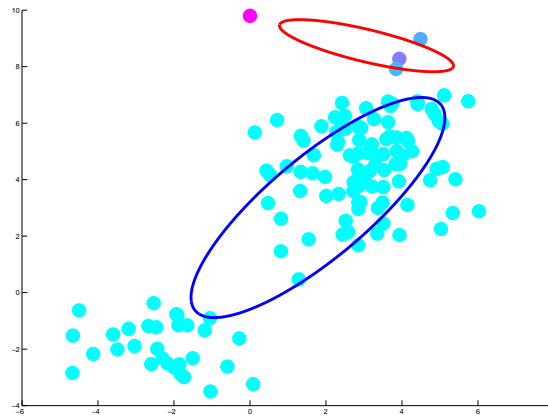
EM and overfitting

- We can be very unlucky with the initial guess.

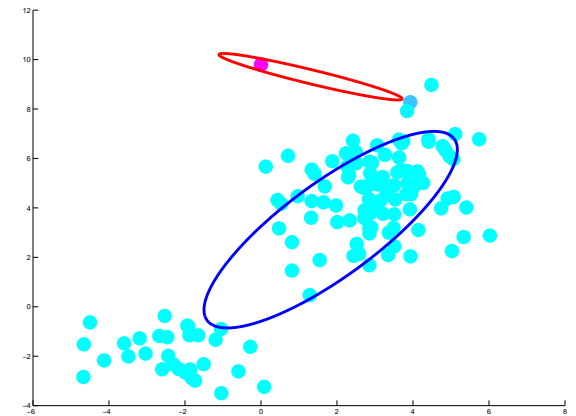
1st iteration



2nd



4th



- The problem:

$$\lim_{\sigma^2 \rightarrow 0} \mathcal{N}(\mathbf{x}; \mu = \mathbf{x}, \Sigma = \sigma^2 \mathbf{I}) = \infty.$$

Regularized EM

- Impose a prior on θ .
- Instead of maximizing the likelihood in the M-step, maximize the posterior:

$$\theta^{new} = \operatorname{argmax}_{\theta} \left\{ E_{p(Z|X;\theta)} [\log p(X, Z; \theta) | X; \theta^{old}] + \log p(\theta) \right\}.$$

- A common prior on a covariance matrix: the *Wishart* distribution

$$p(\mathbf{\Sigma}; \mathbf{S}, n) \propto \frac{1}{|\mathbf{\Sigma}|^{n/2}} \exp \left(-\frac{1}{2} \operatorname{Tr} (\mathbf{\Sigma}^{-1} \mathbf{S}) \right)$$

- Intuition: \mathbf{S} is the covariance of n “hallucinated” observations.

Model selection: setting k

- So far we have assumed known k .
- Idea 1: select k that maximizes the likelihood.

Model selection: setting k

- So far we have assumed known k .
- Idea 1: select k that maximizes the likelihood.
 - The solution: a separate, very narrow Gaussian component for every training example.
 - This solution yields infinite log-likelihood!

Model selection: setting k

- So far we have assumed known k .
- Idea 1: select k that maximizes the likelihood.
 - The solution: a separate, very narrow Gaussian component for every training example.
 - This solution yields infinite log-likelihood!
- We need a criterion to penalize such models.
- Occam's razor: try to find the *simplest* among all possible explanations.

Bayesian Information Criterion

- Let $\pi(\mathcal{M})$ be the number of free parameters in the model \mathcal{M} .
 - For a MoG model with k components in \mathbb{R}^d :

$$\pi(\mathcal{M}) = k \cdot (d + d(d + 1)/2) + k - 1.$$

- For each model, we find MAP estimate of the parameters on $X_N = [\mathbf{x}_1, \dots, \mathbf{x}_N]$:

$$L^*(\mathcal{M}) \triangleq \max_{\theta_{\mathcal{M}}} \{ \log p(X_N | \mathcal{M}; \theta_{\mathcal{M}}) + \log p(\theta_{\mathcal{M}}) \}.$$

- The BIC score for the model \mathcal{M} on data X_N :

$$BIC(\mathcal{M}) = L^*(\mathcal{M}) - \frac{\pi(\mathcal{M})}{2} \log N.$$

Next time

Connection between model selection (and learning in general) and information theory.