

CS599—EM Background

- The EM Algorithm
 - + Maximum Likelihood as a general probabilistic learning procedure
 - + EM as general procedure for maximum likelihood estimation
 - + EM for Mixture Models
 - + Foundations of the EM Algorithm

Maximum Likelihood as a General Learning Procedure

- How to use Maximum Likelihood:

- + define parameterized probabilistic model of the data

$$p(\mathbf{x}; \theta) \quad \text{or} \quad p(\mathbf{y} | \mathbf{x}; \theta)$$

- + define likelihood, usually assuming independently drawn data

$$\ell(\theta) = \prod_{n=1}^N p(\mathbf{x}^n; \theta) \quad \text{or} \quad \ell(\theta) = \prod_{n=1}^N p(\mathbf{y} | \mathbf{x}; \theta)$$

- + find parameters by minimizing the negative log likelihood

$$-\ln \ell(\theta) = -\sum_{n=1}^N \ln p(\mathbf{x}^n; \theta) \quad \text{or} \quad -\ln \ell(\theta) = \sum_{n=1}^N \ln p(\mathbf{y} | \mathbf{x}; \theta)$$

$$\frac{\partial(-\ln \ell(\theta))}{\partial \theta} = 0$$

- Note:

- u sometimes parameters can be found analytically
 - u if not, gradient descent can be performed in the $-\ln$ lik.
 - u there is often a dangerous undesirable global minimum of the likelihood
 - u likelihood landscape has also local minima
 - u we can often find neural networks which perform the equivalent of likelihood estimation

Example I: Multinomial Classification

- A Classical Example from Dempster, Laird , & Rubin , 1977

+ We observe four dimensional multinomial data vectors:

$$\mathbf{y} = (y_1, y_2, y_3, y_4)^T$$

+ From a data sample, we know the frequency of the data in sample of $n=197$:

$$\# y_1 = 125, \# y_2 = 18, \# y_3 = 20, \# y_4 = 34$$

+ From an expert, we know the following parameterization of the probabilities of classes:

$$P(y_1) = \frac{1}{2} + \frac{1}{4}\Psi \quad P(y_2) = \frac{1}{4}(1 - \Psi)$$

$$P(y_3) = \frac{1}{4}(1 - \Psi) \quad P(y_4) = \frac{1}{4}\Psi$$

+ Goal: Find a maximum likelihood estimate of Ψ

Example I: Max. Likelihood Estimation

- The log likelihood is:

$$\begin{aligned} L = -\ln \ell &= -\ln \left(\frac{n!}{y_1! y_2! y_3! y_4!} P(y_1)^{y_1} P(y_2)^{y_2} P(y_3)^{y_3} P(y_4)^{y_4} \right) \\ &= -\ln \left(\frac{n!}{y_1! y_2! y_3! y_4!} \left(\frac{1}{2} + \frac{1}{4} \Psi \right)^{y_1} \left(\frac{1}{4} (1 - \Psi) \right)^{y_2} \left(\frac{1}{4} (1 - \Psi) \right)^{y_3} \left(\frac{1}{4} \Psi \right)^{y_4} \right) \\ &= \text{const} - (y_1 \ln(2 + \Psi) + (y_2 + y_3) \ln(1 - \Psi) + y_4 \ln \Psi) \end{aligned}$$

- Derivative

$$\frac{\partial(L)}{\partial \Psi} = - \left(\frac{y_1}{2 + \Psi} + \frac{y_2 + y_3}{1 - \Psi} + \frac{y_4}{\Psi} \right) = 0$$

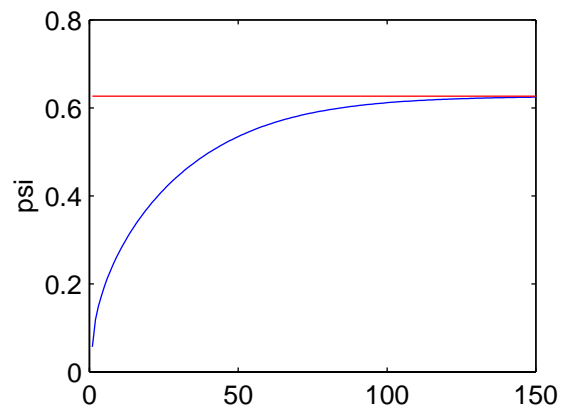
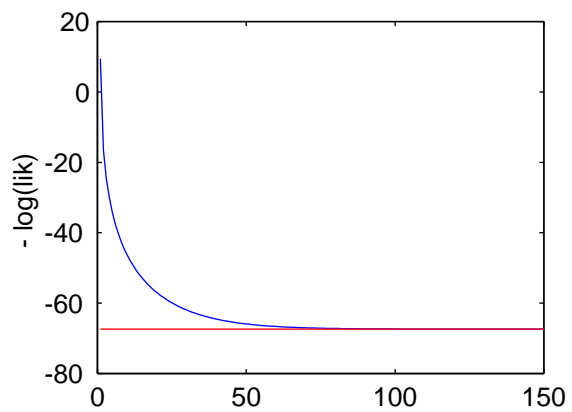
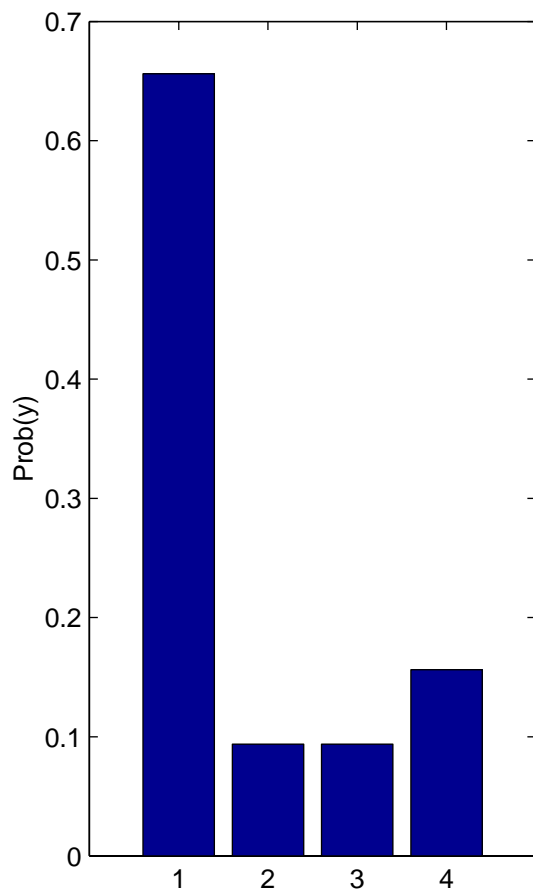
- Analytical Solution:

$$\Psi = 0.6268 \quad \text{and} \quad -\ln \ell = -67.3841$$

Example I: Gradient Descent in likelihood

- Gradient Descent:

$$\Psi^{n-1} = \Psi^n - \alpha \left. \frac{\partial L}{\partial \Psi} \right|_{\Psi = \Psi^n}$$



Example II: Mixture Density Estimation

- One of the most frequently used max. likelihood problems that DOES NOT have an analytical solution

$$p(\mathbf{x}; \theta) = \sum_{k=1}^c p_k(\mathbf{x}; \theta_k) P(k), \quad \sum_{k=1}^c P(k) = 1$$
$$p_k(\mathbf{x}; \theta_k) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

- log. likelihood

$$L = -\ln p(\mathbf{X}; \theta) = -\ln \prod_{n=1}^N \sum_{k=1}^c p_k(\mathbf{x}^n; \theta_k) P(k)$$
$$= -\sum_{n=1}^N \ln \sum_{k=1}^c p_k(\mathbf{x}^n; \theta_k) P(k)$$

- derivatives

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^N h_k^n \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^n - \boldsymbol{\mu}_k); \quad h_k^n = P(k | \mathbf{x}^n) = \frac{p(\mathbf{x}^n | k) P(k)}{\sum_j p(\mathbf{x}^n | j) P(j)}$$

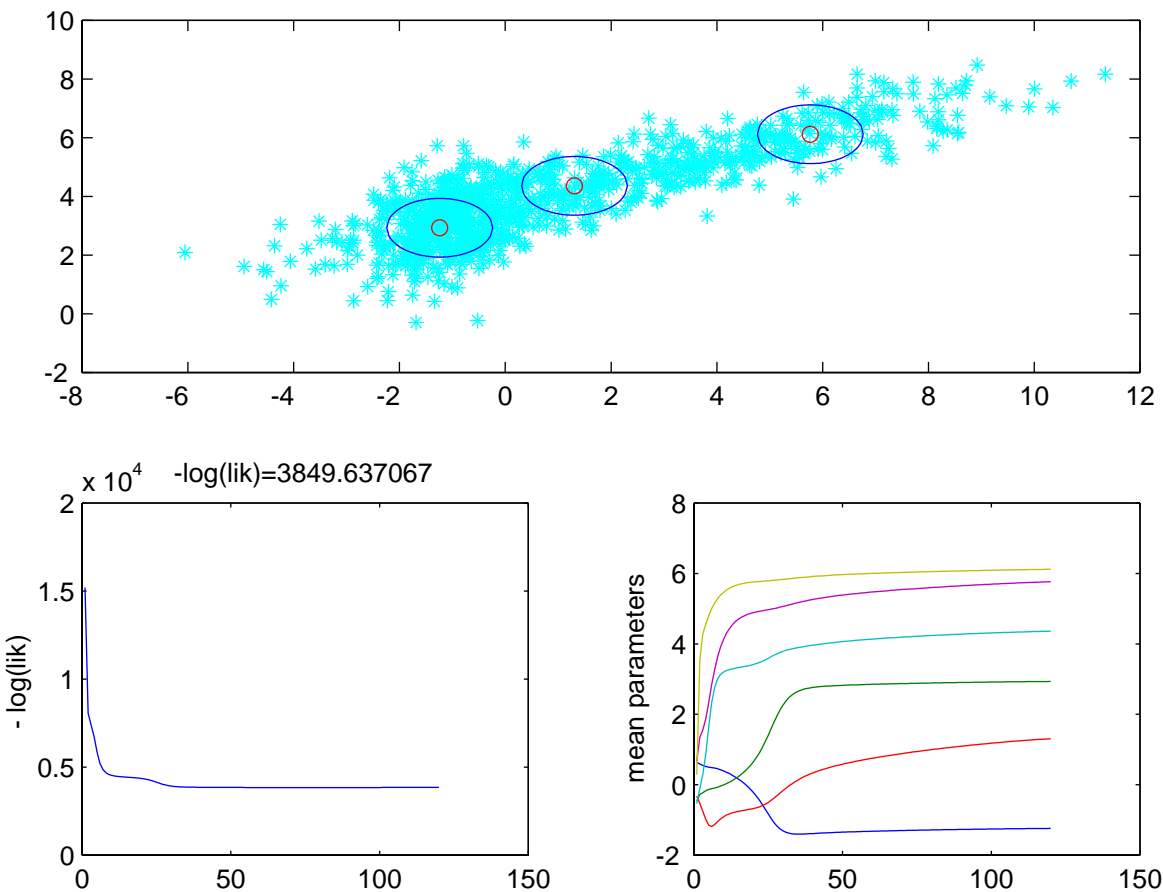
$$\frac{\partial L}{\partial P(k)} = \sum_{n=1}^N \frac{h_k^n}{P(k)}, \text{ where } \sum_{k=1}^c P(k) = 1; \quad \frac{\partial L}{\partial \boldsymbol{\Sigma}_k} = \dots \text{don' t mention, don' t ask } \dots$$

Example II: Gradient Descent

- only “mean”-adaptation and $P(k)$ adaptation

$$\mu_k^{n+1} = \mu_k^n - \alpha \left. \frac{\partial L}{\partial \mu_k} \right|_{\mu_k = \mu_k^n}$$

$$P(k)^{n+1} = P(k)^n - \alpha \left. \frac{\partial L}{\partial P(k)} \right|_{P(k) = P(k)^n}$$



The Expectation- Maximization Algorithm

- The EM algorithm is a method to do max. likelihood estimation
- Idea of EM:
 - + we have some observed data, but max. likelihood estimation of our model is complicated
 - + but maybe if there were some extra variables, the max. likelihood estimation in the “augmented space” would be MUCH simplified
 - + IMPORTANT: the extra variables may be hypothetical!
- Official Notation:
 - + real data space is called “incomplete data space”
 - + augmented data space is called “complete data space”
- Applicability of EM
 - + missing data problems (the most natural application)
 - + but any other max. likelihood problem, too => need creativity to recognize where EM can be used!

The EM Algorithm: Formalization

- Typical Notation of EM:

$p(\mathbf{x};\theta)$ the "incomplete data density"

$\ell = \prod_n p(\mathbf{x};\theta)$ the "incomplete data likelihood"

$L = -\ln \ell$

$p_c(\mathbf{x}, \mathbf{z}; \theta)$ the "complete data density"

\mathbf{z} the "unobserved variables"

$\ell_c = \prod_n p_c(\mathbf{x}, \mathbf{z}; \theta)$ the "complete data likelihood"

$L_c = -\ln \ell_c$

- The pre-requisite of EM:

$$p(\mathbf{x};\theta) = \int_{\mathbf{z}} p_c(\mathbf{x}, \mathbf{z}; \theta) d\mathbf{z}$$

The EM Algorithm: General Procedure

- Start with an initial guess for $\theta = \theta^n$
- Since we cannot observe L_c , find the conditional expectation of L_c given the observed data

$$Q(\theta; \theta^n) = E_{\theta^n} \{-L_c(\theta) | \mathbf{x}\} \quad \text{Expectation-Step}$$

– Note: Very often this reduces to finding the expectation $E\{\mathbf{z}\}$

- Maximize Q with respect to the parameters

$$\theta^{n+1} = \max_{\theta} Q(\theta; \theta^n) \quad \text{Maximization-Step}$$

- Iterate until convergence
- **NOTE: THIS ALGORITHM CONVERGES WITH PROBABILITY ONE TO A (LOCAL) MAXIMUM OF Q WITHOUT ANY LEARNING RATES!**

Example I: EM for Multinomial Classification

- Incomplete Data

$$\mathbf{y} = (y_1, y_2, y_3, y_4)^T$$

- Complete Data

$$(\mathbf{z}, \mathbf{y}) = (z_{11}, z_{12}, y_2, y_3, y_4)^T$$

$$P(y_1) = \frac{1}{2} + \frac{1}{4}\Psi \quad P(y_2) = \frac{1}{4}(1 - \Psi)$$

$$P(y_3) = \frac{1}{4}(1 - \Psi) \quad P(y_4) = \frac{1}{4}\Psi$$

thus becomes \rightarrow

$$P(z_{11}) = \frac{1}{2} \quad P(z_{12}) = \frac{1}{4}\Psi \quad P(y_2) = \frac{1}{4}(1 - \Psi)$$

$$P(y_3) = \frac{1}{4}(1 - \Psi) \quad P(y_4) = \frac{1}{4}\Psi$$

$$z_{11} + z_{12} = y_1$$

- Why this choice of the complete data?

$$L = -(y_1 \ln(2 + \Psi) + (y_2 + y_3) \ln(1 - \Psi) + y_4 \ln \Psi)$$

$$L_c = -(z_{12} \ln(\Psi) + (y_2 + y_3) \ln(1 - \Psi) + y_4 \ln \Psi)$$

Example I: EM (cont'd)

- Derivative (this is the maximization step)

$$\frac{\partial L_c}{\partial \Psi} = 0 = \frac{\partial}{\partial \Psi} ((z_{12} + y_4) \ln(\Psi) + (y_2 + y_3) \ln(1 - \Psi))$$

$$\frac{\partial L_c}{\partial \Psi} = \frac{(z_{12} + y_4)}{\Psi} - \frac{(y_2 + y_3)}{1 - \Psi} = 0$$

$$\Psi = \frac{z_{12} + y_4}{z_{12} + y_2 + y_3 + y_4}$$

This has become very simple!

- The Expectation Step

$$\begin{aligned} Q(\Psi; \Psi^n) &= E_{\Psi^n} \{-L_c(\Psi) | \mathbf{y}\} \\ &= E_{\Psi^n} \{(z_{12} + y_4) \ln(\Psi^n) + (y_2 + y_3) \ln(1 - \Psi^n) | \mathbf{y}\} \\ &= (E\{z_{12} | \mathbf{y}\} + y_4) \ln(\Psi^n) + (y_2 + y_3) \ln(1 - \Psi^n) \end{aligned}$$

z_{11} is a binomial variable (conditional on y_1) with sample size y_1 and

probability $\frac{1/2}{1/2 + 1/4\Psi}$

$$E\{z_{11} | \mathbf{y}\} = E\{z_{11} | y_1\} = y_1 \frac{1/2}{1/2 + 1/4\Psi}$$

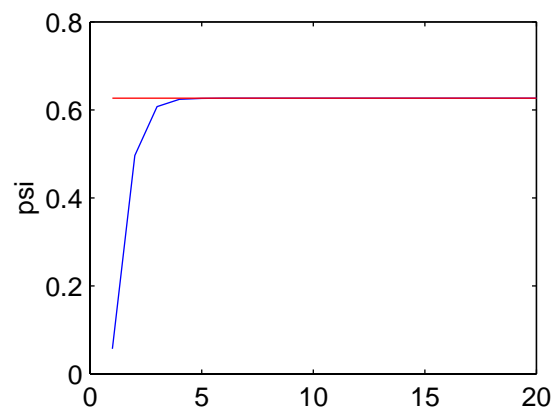
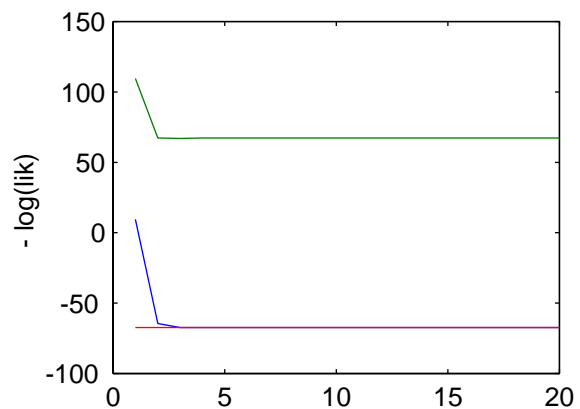
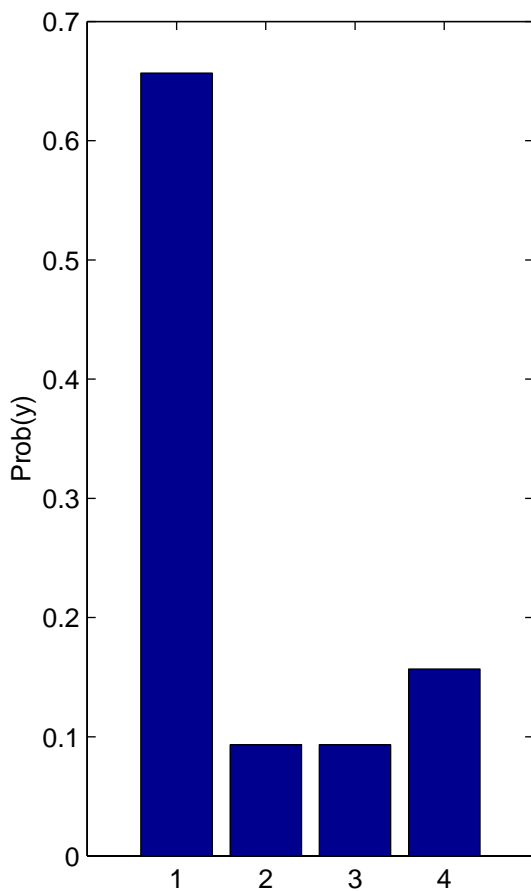
$$^{12} z_{12} = y_1 - z_{11} = y_1 \frac{1/4\Psi}{1/2 + 1/4\Psi}$$

Example I: EM (cont'd)

- Summary:

– E-Step: $E\{z_{11} | \mathbf{y}\} = y_1 \frac{1/2}{1/2 + 1/4\Psi}$; $E\{z_{12} | \mathbf{y}\} = y_1 - z_{11}$

– M-Step $\Psi = \frac{z_{12} + y_4}{z_{12} + y_2 + y_3 + y_4}$



Example II: Mixture Densities

- Incomplete Data

\mathbf{x}

- Complete Data

$$\mathbf{z} = [0, \dots, 0, 1, 0, \dots, 0]^T$$

+ where \mathbf{z} is an indicator variable for every data point, indicating which mixture component created it

$$\begin{aligned} L_c &= -\ln p(\mathbf{X}, \mathbf{Z}; \theta) = -\ln \prod_{n=1}^N \sum_{k=1}^c z_k^n p_k(\mathbf{x}^n; \theta_k) P(k) \\ &= -\sum_{n=1}^N \ln \sum_{k=1}^c z_k^n p_k(\mathbf{x}^n; \theta_k) P(k) \\ &= -\sum_{n=1}^N \sum_{k=1}^c z_k^n \ln p_k(\mathbf{x}^n; \theta_k) + z_k^n \ln P(k) \\ &= -\sum_{k=1}^c \sum_{n=1}^N z_k^n \ln p_k(\mathbf{x}^n; \theta_k) + z_k^n \ln P(k) \end{aligned}$$

Example II: EM Mixture Densities (cont'd)

- Maximization:

$$\frac{\partial L_c}{\partial \theta_k} = \sum_{n=1}^N z_k^n (\mathbf{x}^n - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}^{-1} = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^N z_k^n \mathbf{x}^n}{\sum_{n=1}^N z_k^n}$$

$$P(k) = \frac{1}{N} \sum_{n=1}^N z_k^n$$

- Expectation

$$z_k^n = \frac{p_k(\mathbf{x}^n; \boldsymbol{\theta}_k) P(k)}{\sum_j p_j(\mathbf{x}^n; \boldsymbol{\theta}_j) P(j)}$$

Example II: EM Mixture Densities (cont'd)

- Result

