CIS606 – Lecture 6

Woon Wei Lee, Jacob Crandall
Spring 2014, 10:00am-11:15am,
Mondays and Thursdays

For today:

- Admin stuff: project
- EM algorithm wrap-up
- Tutorial

Gaussian Mixture Models (GMM)

- **Most well known application of EM**
  - Simply another class of mixture model!
  - Recall: MNB model was:

$$P(x;\theta)=\sum_{z=1}^{k} p(x|z;\theta)\,p(z;\theta)$$
$$=\sum_{z=1}^{k}\left[\prod_{j=1}^{m} p(x_j|z)\,p(z;\theta)\right]$$

  - GMM definition very similar:

$$P(x;\theta)=\sum_{i=1}^{k} p(x|z;\theta)\,p(z;\theta)$$
$$=\sum_{z=1}^{k} N(\mu_z,\sigma_z^2)\,p(z;\theta)$$
$$=\sum_{z=1}^{k}\left[p(z;\theta)\frac{1}{\sqrt{2\pi}^{d}|\Sigma_z|^{1/2}}e^{-\frac{1}{2}(x-\mu_z)^T\Sigma_z^{-1}(x-\mu_z)}\right]$$

Masdar INSTITUTE

Cont'd

- **E-Step**

$$Q(z(i)|x(i);\theta)=\frac{p(x(i)|z;\theta)\,p(z(i);\theta)}{p(x(i);\theta)}$$

$$=\frac{N(\mu_z,\sigma_z^2)\,p(z(i);\theta)}{\sum_{z(i)=1}^{k}N(\mu_z,\sigma_z^2)\,p(z(i);\theta)}$$

$$=\frac{p(z(i);\theta)\frac{1}{\sqrt{2\pi}^d|\Sigma_z|^{1/2}}e^{-\frac{1}{2}(x(i)-\mu_z)^T\Sigma_z^{-1}(x(i)-\mu_z)}}{\sum_{z=1}^{k}p(z(i);\theta)\frac{1}{\sqrt{2\pi}^d|\Sigma_z|^{1/2}}e^{-\frac{1}{2}(x(i)-\mu_z)^T\Sigma_z^{-1}(x(i)-\mu_z)}}$$

Cont'd

- **M-Step**
  - Expected complete data log-likelihood is:

$$R(\theta;\hat{\theta})=\sum_{i=1}^{N}\sum_{z}Q(z(i)|x(i);\hat{\theta})\log\left[p(z(i);\theta)\frac{1}{\sqrt{2\pi}^{d}|\Sigma_z|^{1/2}}e^{-\frac{1}{2}(x(i)-\mu_z)^T\Sigma_z^{-1}(x(i)-\mu_z)}\right]$$

$$=\sum_{i=1}^{N}\sum_{z}Q(z(i)|x(i);\hat{\theta})\left[\underbrace{\log p(z(i);\theta)}_{\text{(a)}}+\underbrace{\log\frac{1}{\sqrt{2\pi}^{d}|\Sigma_z|^{1/2}}}_{\text{(b)}}\underbrace{-\frac{1}{2}(x(i)-\mu_z)^T\Sigma_z^{-1}(x(i)-\mu_z)}_{\text{(c)}}\right]$$

  - Maximizing w.r.t. μ_z (only element (c) involved):

$$\frac{\partial R(\theta;\hat{\theta})}{\partial\mu_z}=\sum_{i=1}^{N}Q(z(i)|x(i);\hat{\theta})\frac{1}{2}\Sigma_z^{-1}(x(i)-\mu_z)$$

$$\sum_{i=1}^{N}Q(z(i)|x(i);\hat{\theta})(x(i)-\mu_z)=[0,0,\ldots,0]^T$$

$$\mu_z^{*}=\frac{\sum_{i=1}^{N}Q(z(i)|x(i);\hat{\theta})\,x(i)}{\sum_{i=1}^{N}Q(z(i)|x(i);\hat{\theta})}$$

Cont'd

- Maximizing w.r.t. p(z(i);θ) (only element (a) involved):

$$\frac{\partial R(\theta;\hat{\theta})}{\partial p(z(i);\theta)}=\frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}{p(z(i)|\theta)}-\lambda=0$$

$$p(z(i);\theta)=\frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}{\lambda}$$

(Constrained optimization)

$$\sum_{z}\left[\frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}{\lambda}\right]=1$$

$$\rightarrow\lambda=\sum_{z}\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})$$

$$p(z(i);\theta)=\frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}{\sum_{z}\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}$$

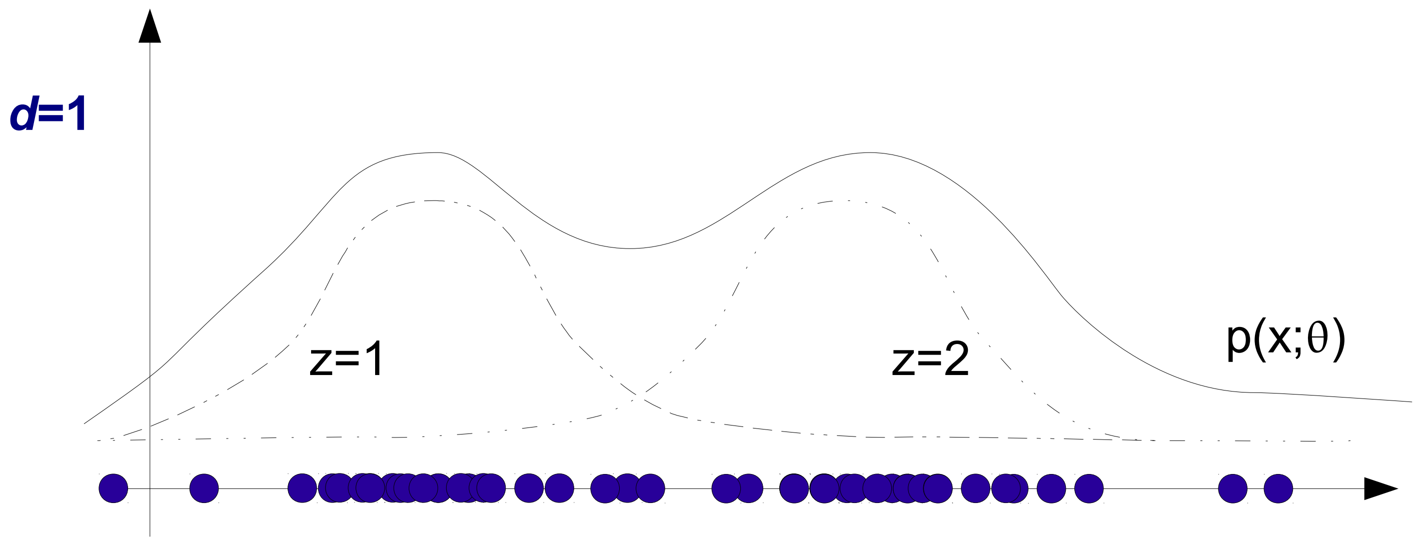
M-Step equations

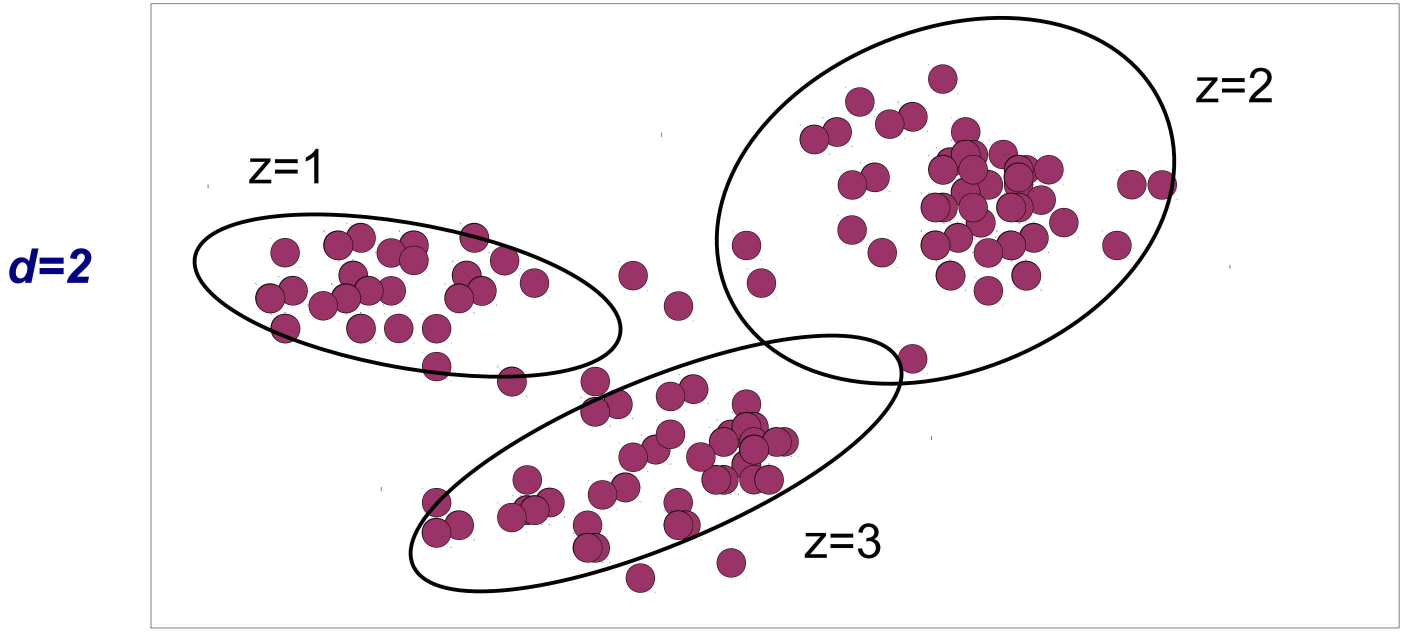
$$\mu_z^* = \frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})\, x(i)}{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}$$

$$p(z(i);\theta)^* = \frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}{\sum_{z}\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}$$

$$\Sigma_z^* = \frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})\left[(x(i)-\mu_z)(x(i)-\mu_z)^T\right]}{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}$$

GMM as a form of clustering





Masdar INSTITUTE

Potential Problems

- **Unknown number of components!**
  - Similar problem with k-means
  - Measures of clustering quality can be used as before to evaluate candidate "*k*"s
- **Singularities**
  - A more problem occurs when the diversity of one of the Gaussians collapses.

    → e.g.: only a single data point – degenerate!
  - Results in a rank-deficient covariance matrix
  - Determinant goes to zero → not possible to invert

(Appendix – additional stuff)

- **Optimize with respect to Σ_z**
  - Differentiate $R(\theta;\hat{\theta})$ w.r.t. $\Sigma_z^{-1} \rightarrow$ only elements (b) and (c) involved:

$$\frac{\partial R(\theta;\hat{\theta})}{\partial \Sigma_z^{-1}} = \underbrace{\frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \left[\log \frac{1}{|\Sigma_z|^{1/2}} \right]}{\partial \Sigma_z^{-1}}}_{\text{(b)}} - \underbrace{\frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \left[\frac{1}{2} (x(i)-\mu_z)^T \Sigma_z^{-1} (x(i)-\mu_z) \right]}{\partial \Sigma_z^{-1}}}_{\text{(c)}}$$

$$= \frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \log |\Sigma_z^{-1}|}{\partial \Sigma_z^{-1}} - \frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \left[Trace\left((x(i)-\mu_z)^T \Sigma_z^{-1} (x(i)-\mu_z) \right) \right]}{\partial \Sigma_z^{-1}}$$

Identity
Trace(AB)=Trace(BA)

$$= \frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \log |\Sigma_z^{-1}|}{\partial \Sigma_z^{-1}} - \frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \left[Trace\left(\Sigma_z^{-1} (x(i)-\mu_z)(x(i)-\mu_z)^T \right) \right]}{\partial \Sigma_z^{-1}}$$

Masdar INSTITUTE

Cont'd

From the “Matrix Cookbook”

2.8.2 Symmetric

If $\mathbf{A}$ is symmetric, then $\mathbf{S}^{ij} = \mathbf{J}^{ij} + \mathbf{J}^{ji} - \mathbf{J}^{ij}\mathbf{J}^{ij}$ and therefore

$$\frac{df}{d\mathbf{A}} = \left[\frac{\partial f}{\partial \mathbf{A}}\right] + \left[\frac{\partial f}{\partial \mathbf{A}}\right]^T - \operatorname{diag}\left[\frac{\partial f}{\partial \mathbf{A}}\right] \qquad (127)$$

That is, e.g., ([5]):

$$\boxed{\frac{\partial \operatorname{Tr}(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A} + \mathbf{A}^T - (\mathbf{A} \circ \mathbf{I}), \text{ see (131)}} \qquad (128)$$

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \det(\mathbf{X})(2\mathbf{X}^{-1} - (\mathbf{X}^{-1} \circ \mathbf{I})) \qquad (129)$$

$$\boxed{\frac{\partial \ln \det(\mathbf{X})}{\partial \mathbf{X}} = 2\mathbf{X}^{-1} - (\mathbf{X}^{-1} \circ \mathbf{I})} \qquad (130)$$

- Apply to (c)
- Apply to (b)

$$\frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \log|\Sigma_z^{-1}|}{\partial \Sigma_z^{-1}} - \frac{\partial \sum_{i=1}^{N} \sum_{z} Q(z(i)|x(i);\hat{\theta}) \left[Trace\left(\Sigma_z^{-1}(x(i)-\mu_z)(x(i)-\mu_z)^T\right)\right]}{\partial \Sigma_z^{-1}}$$

$$= \sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})(2\Sigma_z - I \circ \Sigma_z) - \sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})(2\mathrm{S} - I \circ S) = 0 \quad \left[where\ S = (x(i)-\mu_z)(x(i)-\mu_z)^T\right]$$

$$2\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})(\Sigma_z - S) - I \circ \sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})(\Sigma_z - S) = 0$$

$$\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})(\Sigma_z - (x(i)-\mu_z)(x(i)-\mu_z)^T) = 0 \quad [2\mathrm{A} - I \circ A = 0 \rightarrow A = 0]$$

$$\Sigma_z = \frac{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})\left[(x(i)-\mu_z)(x(i)-\mu_z)^T\right]}{\sum_{i=1}^{N} Q(z(i)|x(i);\hat{\theta})}$$

Recommended Reads

- “A gentle tutorial of the EM algorithm”, Jeff Bilmes
- Andrew Ng's EM algorithm lecture notes
- “Latent Semantic Models for Collaborative Filtering” Thomas Hoffman
- “The Expectation Maximization Algorithm” - Frank Dellaert
- “The Matrix Cookbook”, Petersen and Pedersen
- “Elements of Statistical Learning”, pg. 272 onwards
- Various Wikis, Google Scholar, etc.