L12 – Mixture Density Models, EM Algorithm

Mixture Density Models

- Flexible models – able to fit lots of densities
- Fit parameters by maximum likelihood. Nonlinear equations require iterative fitting procedure. Standard is Expectation – Maximization (EM).
- "Soft" version of clustering.

General form is $p(x|\Theta) = \sum_{j=1}^{k} \alpha_j \; p(x|j)$

$p(x|j) \equiv p(x|\theta_j)$ are component densities with parameter (vectors) θ_j

$\Theta \equiv (\alpha_1, \ldots, \alpha_k, \theta_1, \theta_k)$

$\alpha_j \geq 0, \quad \sum_{j=1}^{k} \alpha_j = 1 \qquad \alpha_j$ is prior probability for mixture component j

Generative Model

Mixture model form $$p(x \mid \Theta) = \sum_{j=1}^{k} \alpha_j \; p(x \mid j)$$

$p(x \mid j) \equiv p(x \mid \theta_j)$ are component densities with parameter (vectors) θ_j

$\Theta \equiv \left(\alpha_1, \ldots, \alpha_k, \, \theta_1, \, \ldots, \theta_k \right)$

$\alpha_j \geq 0, \quad \sum_{j=1}^{k} \alpha_j = 1 \quad \alpha_j$ is prior probability for mixture component j

Generating *x* is a two-fold sampling procedure:

1. Pick a component density with probability α_j
2. Generate a sample *x* from *p(x | j)*

Mixture Models

Most common example is mixture of Gaussians

$$p(x \mid \Theta) = \sum_{j=1}^{k} \alpha_j \quad p(x \mid j)$$

with

$$p(x \mid j) = \frac{1}{\sqrt{(2\pi)^n \left|\Sigma_j\right|}} \exp - (x -) \ \Sigma^{1} (x - \mu))$$

There's a universal approximation th uch mixtures that states that with enough components, f Gaussians fit by maximum likelihood can arbitrarily closely match any density on a compact subset of R^n.

1. Jonathan Li and Andrew Barron. Mixture Density Estimation, in Solla, Leen, and Mueller (eds.) *Advances in Neural Information Processing Systems 12*, The MIT Press, 2000.

GEORGETOWN UNIVERSITY

Gaussian Mixture Model

Flexible --- can make lots of shapes!

Fitting Mixture Models

Suppose we have a data set

$$D = \{ x_a \ , \ a = 1, \ldots, N \} \quad \text{with each } x_a \text{ a vector in } R^n$$

we'd like to adjust the parameters

$$\Theta \equiv (\alpha_1, \ldots, \alpha_k, \theta_1, \ldots, \theta_k)$$

so as to maximize the data [illegible]ood

$$L(\Theta) = \ln P(D \mid \Theta) = \sum_{a=1}^{N} \ln\left(\sum_{j=1}^{k} \alpha_j \ p(x_a \mid \theta_j) \right)$$

GEORGETOWN UNIVERSITY

Fitting Mixture Models

The data log likelihood

$$L(\Theta) = \ln P(D \mid \Theta) = \sum_{a=1}^{N} \ln\left(\sum_{j=1}^{k} \alpha_j \;\; p(x_a \mid \theta_j) \right)$$

cannot be maximized in one step --- the maximization equations don't have a closed form solution.

Instead, use an itera algorithm

For the moment, re suppressing the mixture form of $p(x|\Theta)$

$$L(\Theta) = \ln P(D \mid \Theta) = \sum_{a=1}^{N} \ln p(x_a \mid \Theta) = \sum_{a=1}^{N} \ln\left(\sum_{i_a=1}^{k} p(i_a, x_a \mid \Theta) \right)$$

where i_a is the <u>unknown</u> index of the component responsible for generating x_a.

Fitting Mixture Models

Next, we write a lower bound for *L*. Introduce an average over <u>any</u> probability distribution on the unknown indices i_a , $Q(i_a)$

$$L = \sum_{a=1}^{N} \ln p\left(x_a \mid \Theta\right) = \sum_{a=1}^{N} \ln \left\{ \sum_{i_a=1}^{k} p(i_a, x_a | \Theta) \right\} = \sum_{a=1}^{N} \ln \left\{ \sum_{i_a=1}^{k} Q(i_a) \frac{p(i_a, x_a | \Theta)}{Q(i_a)} \right\}$$

Jensen's inequality gives

$$L = \sum_{a=1}^{N} \ln \left\{ \sum_{i_a=1}^{k} Q(i_a) \frac{p(i_a, x_a)}{Q(i_a)} \right\} \geq \sum_{a=1}^{N} \sum_{i_a=1}^{k} Q(i_a) \ln\left(\frac{p(i_a, x_a)}{Q(i_a)} \right)$$

$$= \sum_{a=1}^{N} \sum_{i_a=1}^{k} Q(i_a) \ln p(i_a, x_a) - \sum_{a=1}^{N} \sum_{i_a=1}^{k} Q(i_a) \ln Q(i_a) \equiv \Gamma(\Theta)$$

The equality holds when $Q(i_a)$ is the posterior distribution on the unknown indices

$$Q(i_a) = p(i_a \mid x_a , \Theta)$$

EM Algorithm

Iterative optimization algorithm: Expectation Maximization (EM) maximizes Γ (which maximizes *L).* There are multiple optima, EM only finds a local optimum.

Initialize the algorithm to some choice of the parameters. At the n+1th iteration :

E Step: With Θ fixe ex distribution as

$$\boxed{Q_{n+1}(i_a) = h_{i,a}(n+1) \equiv p(i|x_a, \Theta(n)) = \frac{\alpha_i(n)\ p\left(x_a|\theta_i(n)\right)}{\sum_{j=1}^{k} \alpha_j(n)\ p\left(x_a|\theta_j(n)\right)}}$$

GEORGETOWN UNIVERSITY

EM

M Step: With $Q = h_{i,a}(n+1)$ fixed, maximize Γ with respect to Θ

$$\Theta(n+1) = \arg\max_{\Theta} \Gamma\big(\Theta, h_{i,a}(n+1)\big) = \arg\max_{\Theta} \sum_{a=1}^{N}\sum_{i=1}^{k} h_{i,a}(n+1) \ln\left(\alpha_i \; p(x_a \mid \theta_i)\right)$$

subject to the condition $\sum_{i=1}^{k} \alpha_i = 1$

This gives

$$\boxed{\alpha_i(n+1) = \frac{1}{N}\sum_{a=1}^{N} h_{ia} = \frac{1}{N}\sum_{a=1}^{N} p(i \mid x_a, \Theta(n))}$$

for the α_i $i=1,\ldots k$

EM

<u>M Step </u>(continued) With $Q = h_{ia}(n+1)$ fixed,
maximize $\Gamma(\Theta,h)$ with respect to the θ_j

$$\Theta(n+1) = \arg\max_{\Theta} \Gamma\big(\Theta, h_{i,a}(n+1)\big) = \arg\max_{\Theta} \sum_{a=1}^{N}\sum_{i=1}^{k} h_{i,a}(n+1)\ \ln\big(\alpha_i\ p(x_a\,|\,\theta_i)\big)$$

Maximize Γ with res [illegible] $\theta_j\,\Gamma = 0$)
separately, so the above reduces to

$$\boxed{\theta_j(n+1) = \arg\max_{\theta_j} \sum_{a=1}^{N} h_{j,a}(n+1)\ \ln\big(\alpha_j\, p(x_a\,|\,\theta_j)\big)}$$

Example – Mixture of Gaussians

Component densities $$p(x \mid \theta_j) = \frac{1}{\sqrt{(2\pi)^n \, |\Sigma_i|}} \exp -\tfrac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)$$

E-Step $$h_{a,i}(n+1) \equiv p(i_a \mid x_a, \Theta(n)) = \frac{\alpha_i(n)\ p(x_a \mid \theta_i(n))}{\sum_{j=1}^{k} \alpha_j(n)\ p(x_a \mid \theta_j(n))}$$

M-Step

$$\alpha_i(n+1) = \frac{1}{N} \sum_{a=1}^{N} h_{i,a}(n+1) \qquad \mu_i(n+1) = \frac{\sum_{a=1}^{N} h_{i,a}(n+1)\ x_a}{\sum_{a} h_{i,a}(n+1)}$$

$$\Sigma_i(n+1) = \frac{\sum_{a=1}^{N} h_{i,a}(n+1)\left(x_a - \mu_i(n+1)\right)\left(x_a - \mu_i(n+1)\right)^T}{\sum_{a} h_{i,a}(n+1)}$$

Gaussian Mixtures

Let’s interpret equations for the M-Step

$$\alpha_i(n+1) = \frac{1}{N}\sum_{a=1}^{N} h_{i,a}(n+1)$$

New estimate of prior for i^{th} component is the average over the data points of the posteriors for i^{th} component.

$$\mu_i(n+1) = \frac{\sum_{a=1}^{N} h_{i,a}(n+1)\; x_a}{\sum_{a} h_{i,a}(n+1)}$$

nent is weighted average
g is fraction of the data
point x_a attrib
onent i

$$\Sigma_i(n+1) = \frac{\sum_{a=1}^{N} h_{i,a}(n+1)\,\left(x_a - \mu_i(n+1)\right)\left(x_a - \mu_i(n+1)\right)^T}{\sum_{a} h_{i,a}(n+1)}$$

New covariance is constructed from weighted outer product.

Gaussian Mixture Model

Flexible --- can make lots of shapes!

GEORGETOWN UNIVERSITY

EM Summary --- Gaussian Mixtures

Initialize parameters

$\alpha_i(0) = 1/k$ all components equally likely

$\mu_i(0) = x_i$ *k* ra raining data

$\Sigma_i(0)$ a positive symmetric, po matrix e.g. $\sigma^2 I$

EM Summary --- Gaussian Mixtures

Iterate

E-Step (estimate posteriors) $$h_{a,i}(n+1) \equiv p(i_a \mid x_a, \Theta(n)) = \frac{\alpha_i(n)\ p(x_a \mid \theta_i(n))}{\sum_{j=1}^{k} \alpha_j(n)\ p(x_a \mid \theta_j(n))}$$

M-Step

Re-estimate priors $$\alpha_i(n+1) = \frac{1}{N}\sum_{a=1}^{N} h_{i,a}(n+1)$$

Re-estimate means $$\mu_i(n+1) = \frac{\sum_{a=1}^{N} h_{i,a}(n+1)\ x_a}{\sum_a h_{i,a}(n+1)}$$

Re-estimate covariances

$$\Sigma_i(n+1) = \frac{\sum_{a=1}^{N} h_{i,a}(n+1)\left(x_a - \mu_i(n+1)\right)\left(x_a - \mu_i(n+1)\right)^T}{\sum_a h_{i,a}(n+1)}$$

Caveats

In high dimensions *n*, there are loads of covariance matrix elements. Likely to overfit.

Fixes – <u>constrain</u> covariance matrices to have fewer components

Diagonal

$$\Sigma_i = \begin{pmatrix} \lambda_{i1} & & 0 \\ & \lambda_{i2} & \\ 0 & & \ddots \end{pmatrix}$$

Spherically symmetric $\Sigma_i = \sigma_i^2 I$, with I the identity matrix

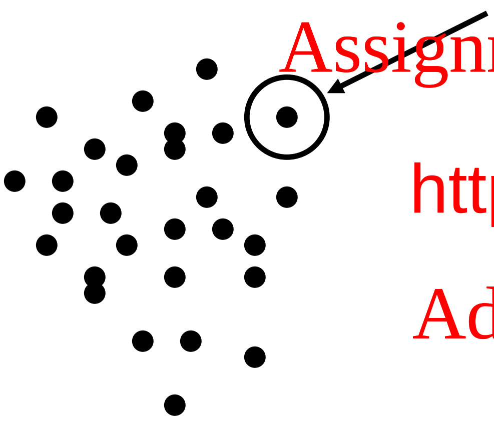
Some other clever form (???)

Note that any constraints modify the M-step equations for the covariance --- <u>can you derive the forms</u>?

Caveats

There are regions of the parameter space where the likelihood goes through the roof but the resulting model is bad

One Gaussian wrapped around this one point x'. Let $\mu_i = x'$, and take σ_i^2 *to 0.*



$$p(x' \mid \mu_i, \sigma_i^2) \rightarrow \infty$$

hout bound. This is particularly likel sions where the average distanc points becomes large.

Regularization (has a grounding in Bayesian priors and MAP estimation). After re-estimation, add a <u>small</u> diagonal matrix to the covariance

$$\Sigma_i(n+1) \rightarrow \Sigma_i(n+1) + \varepsilon I$$

References

- Dempster, Laird, and Rubin. Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statistical Soc. B,* 39, 1-39, 1977.
- Rener and Walker. Mixture densities, maximum likelihood and the EM algorithm, SI 4.
- Ormoneit and Tre *formation Processing Systems 8,* The MIT Press, 1996.
- Jonathan Li and Andrew Barron. y Estimation, in Solla, Leen, and Mueller (eds.) *Advances in Neural Information Processing Systems 12*, The MIT Press, 2000.