The Expectation Maximization or EM algorithm

Carl Edward Rasmussen

November 15th, 2017

Contents

Notation

Probabilistic models may have *visible* (or *observed*) variables y, *latent* variables, (or *hidden* or *unobserved* variables or *missing data*) z and parameters θ.

Example: in a Gaussian mixture model, the visible variables are the observations, the latent variables are the assignments of data points to mixture components and the parameters are the means, variances, and weights of the mixture components.

The likelihood, $p(y|\theta)$, is the probability of the visible variables given the parameters. The goal of the EM algorithm is to find parameters θ which maximize the likelihood. The EM algorithm is iterative and converges to a local maximum.

Throughout, $q(z)$ will be used to denote an arbitrary distribution of the latent variables, z. The exposition will assume that the latent variables are continuous, but an analogue derivation for discrete z can be obtained by substituting integrals with sums.

The lower bound

Bayes' rule:

$$p(z|y,\theta) \;=\; \frac{p(y|z,\theta)p(z|\theta)}{p(y|\theta)} \quad\Leftrightarrow\quad p(y|\theta) \;=\; \frac{p(y|z,\theta)p(z|\theta)}{p(z|y,\theta)}.$$

Multiply and divide by an arbitrary (non-zero) distribution $q(z)$:

$$p(y|\theta) \;=\; \frac{p(y|z,\theta)p(z|\theta)}{q(z)}\,\frac{q(z)}{p(z|y,\theta)},$$

take logarithms:

$$\log p(y|\theta) \;=\; \log\frac{p(y|z,\theta)p(z|\theta)}{q(z)} \;+\; \log\frac{q(z)}{p(z|y,\theta)},$$

and average both sides wrt $q(z)$:

$$\log p(y|\theta) \;=\; \underbrace{\int q(z)\log\frac{p(y|z,\theta)p(z|\theta)}{q(z)}\,dz}_{\text{lower bound functional } \mathcal{F}(q(z),\theta)} \;+\; \underbrace{\int q(z)\log\frac{q(z)}{p(z|y,\theta)}\,dz}_{\text{non-negative } \mathcal{KL}(q(z)\|p(z|y,\theta))}.$$

The EM algorithm

From initial (random) parameters $\theta^{t=0}$ iterate $t = 1, \ldots, T$ the two steps:

E step: for fixed θ^{t-1}, maximize the lower bound $\mathcal{F}(q(z), \theta^{t-1})$ wrt $q(z)$. Since the log likelihood $\log p(y|\theta)$ is independent of $q(z)$ maximizing the lower bound is equivalent to minimizing $\mathcal{KL}(q(z)||p(z|y, \theta^{t-1}))$, so $q^t(z) = p(z|y, \theta^{t-1})$.

M step: for fixed $q^t(z)$ maximize the lower bound $\mathcal{F}(q^k(z), \theta)$ wrt θ. We have:

$$\mathcal{F}(q(z), \theta) = \int q(z) \log \big(p(y|z, \theta) p(z|\theta)\big) dz - \int q(z) \log q(z) dz,$$

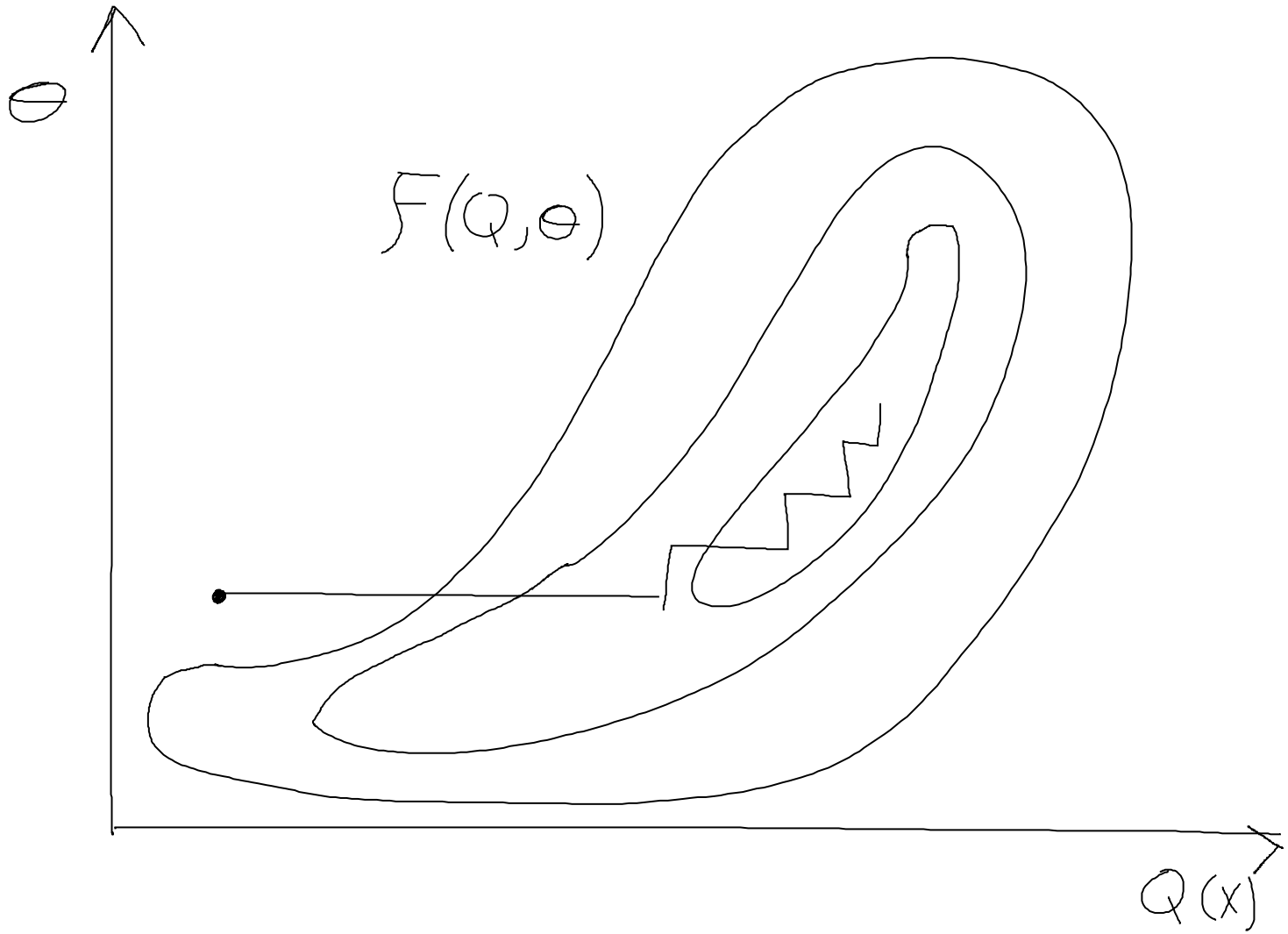
whose second term is the entropy of $q(z)$, independent of θ, so the M step is

$$\theta^t = \operatorname*{argmax}_\theta \int q^t(z) \log \big(p(y|z, \theta) p(z|\theta)\big) dz.$$

Although the steps work with the lower bound, each iteration cannot decrease the log likelihood as

$$\log p(y|\theta^{t-1}) \overset{\text{E step}}{=} \mathcal{F}(q^t(z), \theta^{t-1}) \overset{\text{M step}}{\leqslant} \mathcal{F}(q^t(z), \theta^t) \overset{\text{lower bound}}{\leqslant} \log p(y|\theta^t).$$

EM as Coordinate Ascent in $\mathcal{F}$



Example: Mixture of Gaussians

In a Gaussian mixture model, the parameters are $\theta = \{\mu_j, \sigma_j^2, \pi_j\}_{j=1\ldots k}$ the mixture means, variances and mixing proportions for each of the k components. There is one latent variable per data-point $z_i, i = 1 \ldots n$ taking on values $1 \ldots k$.

The probability of the observations given the latent variables and the parameters, and the prior on latent variables are

$$p(y_i|z_i = j, \underline{\theta}) \;=\; \exp\left(-\tfrac{(y_i-\mu_j)^2}{2\sigma_j^2}\right)/\sqrt{2\pi\sigma_j^2}, \qquad p(z_i = j|\underline{\theta}) \;=\; \pi_j,$$

so the E step becomes:

posterior probability of data point i has latent value j. Each data point i has different posterior prob because of their different location in the corresponding Gaussian.

$$q(z_i = j) \;\propto\; u_{ij} = \pi_j \exp(-(y_i - \mu_j)^2/2\sigma_j^2)/\sqrt{2\pi\sigma_j^2} \;\Rightarrow\; q(z_i = j) = r_{ij} = \frac{u_{ij}}{u_i},$$

where $u_i = \sum_{j=1}^{k} u_{ij}$. This shows that the posterior for each latent variable, z_i follows a discrete distribution with probability given by the product of the prior and likelihood, renormalized. Here, r_{ij} is called the *responsibility* that component j takes for data point i.

Example: Mixture of Gaussians continued

The lower bound is

$$\mathcal{F}(q(z),\theta) = \sum_{i=1}^{n}\sum_{j=1}^{k} q(z_i=j)\left[\log(\pi_j) - \tfrac{1}{2}(y_i-\mu_j)^2/\sigma_j^2 - \tfrac{1}{2}\log(\sigma_j^2)\right] + \text{const.}$$

The M step, optimizing $\mathcal{F}(q(z),\theta)$ wrt the parameters, θ

$$\frac{\partial\mathcal{F}}{\partial\mu_j} = \sum_{i=1}^{n} q(z_i=j)\frac{y_i-\mu_j}{\sigma_j^2} = 0 \;\Rightarrow\; \mu_j = \frac{\sum_{i=1}^{n} q(z_i=j)\,y_i}{\sum_{i=1}^{n} q(z_i=j)},$$

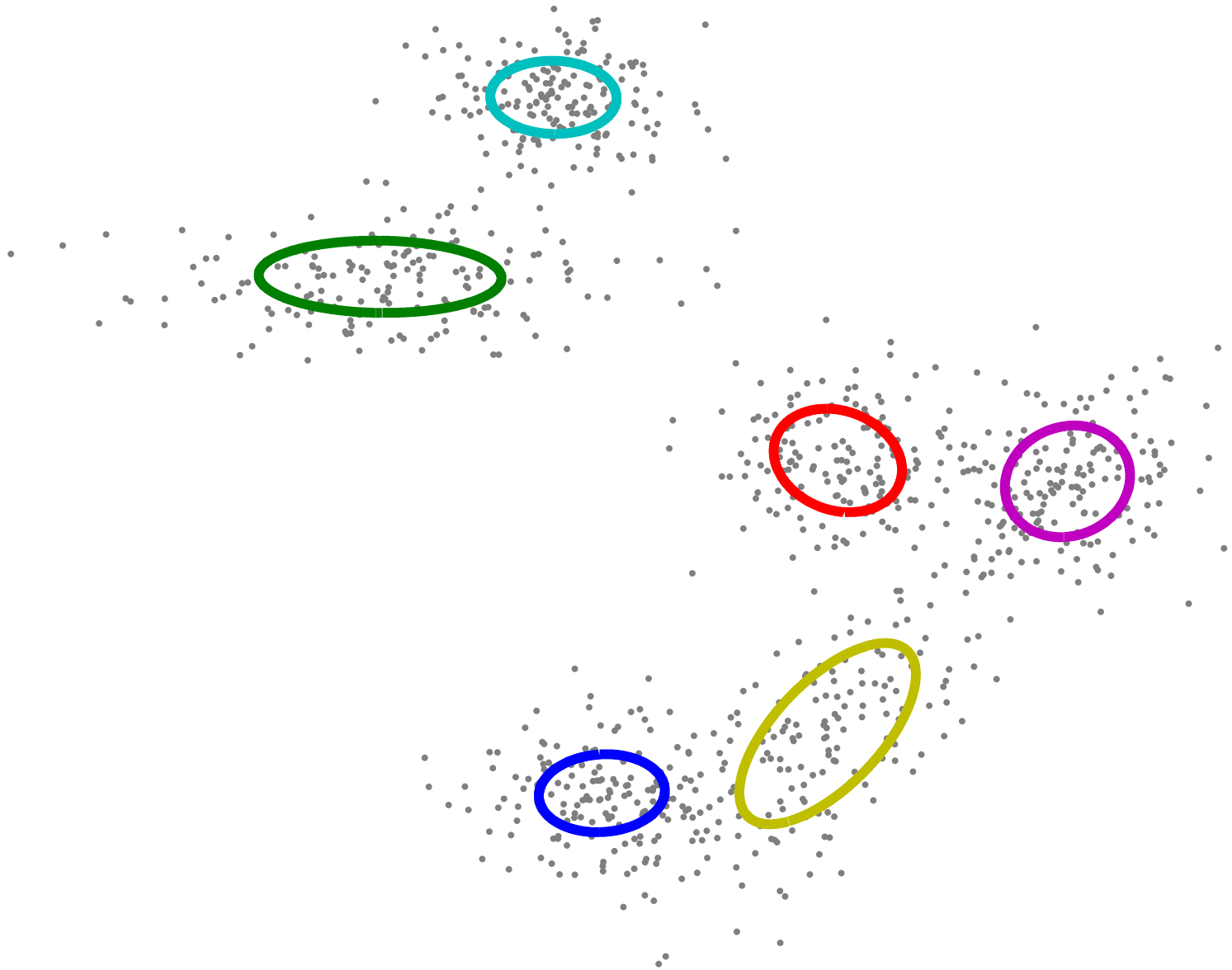
$$\frac{\partial\mathcal{F}}{\partial\sigma_j^2} = \sum_{i=1}^{n} q(z_i=j)\left[\frac{(y_i-\mu_j)^2}{2\sigma_j^4} - \frac{1}{2\sigma_j^2}\right] = 0 \;\Rightarrow\; \sigma_j^2 = \frac{\sum_{i=1}^{n} q(z_i=j)\,(y_i-\mu_j)^2}{\sum_{i=1}^{n} q(z_i=j)},$$

$$\frac{\partial[\mathcal{F} + \lambda(1-\sum_{j=1}^{k}\pi_j)]}{\partial\pi_j} = 0 \;\Rightarrow\; \pi_j = \frac{1}{n}\sum_{i=1}^{n} q(z_i=j),$$

which have nice interpretations in terms of weighted averages.

Clustering with MoG

Clustering with MoG



Appendix: some properties of KL divergence

The (asymmetric) Kullbach Leibler divergence (or relative entropy) $\mathcal{KL}(q(x)||p(x))$ is non-negative. To minimize, add a Lagrange multiplier enforcing proper normalization and take variational derivatives:

$$\frac{\delta}{\delta q(x)}\Big[\int q(x)\log\frac{q(x)}{p(x)}dx+\lambda\big(1-\int q(x)dx\big)\Big] = \log\frac{q(x)}{p(x)}+1-\lambda.$$

Find stationary point by setting the derivative to zero:

$$q(x) = \exp(\lambda-1)p(x), \quad \text{normalization conditon } \lambda = 1, \quad \text{so } q(x) = p(x),$$

which corresponds to a minimum, since the second derivative is positive:

$$\frac{\delta^2}{\delta q(x)\delta q(x)}\mathcal{KL}(q(x)||p(x)) = \frac{1}{q(x)} > 0.$$

The minimum value attained at $q(x) = p(x)$ is $\mathcal{KL}(p(x)||p(x)) = 0$, showing that $\mathcal{KL}(q(x)||p(x))$

- is non-negative
- attains its minimum 0 when $p(x)$ and $q(x)$ are equal (almost everywhere).