MLPM Tutorial 7

November 26, 2019

1. A mixture model over D-dimensional bit vectors can by defined by

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \left[p(z_i = k|\boldsymbol{\theta}) \prod_{j=1}^{D} \mathrm{Bern}(x_{ij}|\mu_{kj}) \right]$$

   We are to learn the parameters $\boldsymbol{\theta}$ using EM.

   (a) Derive the E-step.

   (b) Show that the M-step for ML estimation is given by

$$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}}$$

   (c) Show that the M-step for MAP estimation with $\mathrm{Beta}(\alpha, \beta)$ prior is given by

$$\mu_{kj} = \frac{\alpha - 1 + \sum_i r_{ik} x_{ij}}{\alpha + \beta - 2 + \sum_i r_{ik}}$$

2. Consider the Gaussian mixture model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_k \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

   Define the log likelihood as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta})$$

   The posterior responsibility of mixture k for datapoint i is

$$r_{ik} = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

(a) Show that the gradient of the log-likelihood wrt $\boldsymbol{\mu}_k$ is

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} r_{ik}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)$$

(b) Derive the gradient of the log-likelihood wrt π_k. (For now ignore any constraints on π_k.)

(c) One way to handle the constraints that $\sum_{k=1}^{K}\pi_k = 1$ is to reparameterise using the softmax function

$$\pi_k = \frac{e^{w_k}}{\sum_{k'=1}^{K} e^{w_{k'}}}$$

where $w_k \in \mathbb{R}$ are unconstrained parameters. Show that

$$\frac{\partial}{\partial w_k}\ell(\theta) = -N\pi_k + \sum_{i=1}^{N} r_{ik}$$

Hint: Find the derivative $\frac{\partial \pi_j}{\partial w_k}$ and use the chain rule.

(d) Derive the gradient of the log-likelihood wrt $\boldsymbol{\Sigma}_k$. (For now ignore any constraints on π_k.)

(e) One way to handle the constraint that $\boldsymbol{\Sigma}_k$ be a symmetric positive definite matrix is to reparameterise using a Cholesky decomposition $\boldsymbol{\Sigma}_k = \mathbf{R}_k^T\mathbf{R}_k$ where $\mathbf{R}_k$ is an upper-triangular, but otherwise unconstrained matrix. Derive the gradient of the log-likelihood wrt $\mathbf{R}_k$.

3. Using the same mixture of K Gaussians as described above show that

$$\mathbb{E}\left[\mathbf{x}\right] = \sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k$$

$$\mathrm{cov}\left[\mathbf{x}\right] = \sum_{k=1}^{K}\pi_k(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T) - \mathbb{E}\left[\mathbf{x}\right]\mathbb{E}\left[\mathbf{x}\right]^T$$

Hint: use the fact that $\mathrm{cov}\left[\mathbf{x}\right] = \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] - \mathbb{E}\left[\mathbf{x}\right]\mathbb{E}\left[\mathbf{x}\right]^T$.

4. Consider a simple two variable belief network $p(y, x) = p(y|x)p(x)$ where both $x \in \{0, 1\}$ and $y \in \{0, 1\}$ are binary variables. You have a set of training data $\{(x_i, y_i)\}_{i=1}^{N}$ in which some x_i's are missing. We are specifically interested in finding $p(x)$ from this data. A colleague suggests that one can set $p(x)$ by simply looking at datapoints where x is observed, and then setting $p(x = 1)$ to be the fraction of observed x that is in state 1. Explain how this procedure relates to maximum likelihood and EM.