

Homework 9 Solutions

START HERE: Instructions

- Thanks a lot to **John A.W.B. Constanzo** for providing and allowing to use the latex source files for quick preparation of the HW solution.
- The homework is due at 9:00am on April 20, 2015. Anything that is received after that time will not be considered.
- Answers to every theory questions will be also submitted electronically on Autolab (PDF: Latex or handwritten and scanned). Make sure you prepare the answers to each question separately.
- Collaboration on solving the homework is allowed (after you have thought about the problems on your own). However, when you do collaborate, you should list your collaborators! You might also have gotten some inspiration from resources (books or online etc...). This might be OK only after you have tried to solve the problem, and couldn't. In such a case, you should cite your resources.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.
- Latex source of this homework: http://alex.smola.org/teaching/10-701-15/homework/hw9_latex.zip.

Mixtures of Exponential Family

In this problem we will study approximate inference on a general Bayesian Mixture Model. In particular, we will derive both Expectation-Maximization (EM) algorithm and Gibbs Sampling for the mixture model.

A typical finite-dimensional mixture model is a hierarchical model consisting of the following components:

- N random variables $x_i, i = 1, \dots, N$ corresponding to observations, each assumed to be distributed according to a mixture of K components, with each component belonging to the same exponential family of distributions (e.g., all normal, all multinomial, etc.) but with different parameters

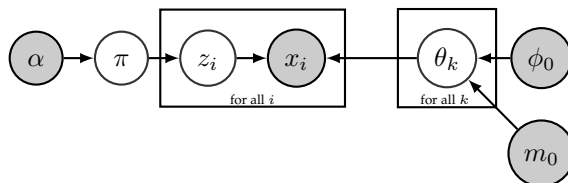
$$p(x_i|\theta) = \exp(\langle \phi(x_i), \theta \rangle - g(\theta)). \quad (1)$$

- N corresponding random latent variables z_i specifying the identity of the mixture component of each observation, each distributed according to a K -dimensional categorical distribution.
- A set of K mixture weights $\pi_k, k = 1, \dots, K$, each of which is a probability (a real number between 0 and 1 inclusive), all of which sum to 1.
- A Dirichlet prior on the mixture weights having hyper-parameters α .
- A set of K parameters $\theta_k, k = 1, \dots, K$, each specifying the parameter of the corresponding mixture component. For example, observations distributed according to a mixture of one-dimensional Gaussian distributions will have a mean and variance for each component. Observations distributed according to a mixture of V -dimensional categorical distributions (e.g., when each observation is a word from a vocabulary of size V) will have a vector of V probabilities, collectively summing to 1. Moreover, we put a shared conjugate prior on these parameters:

$$p(\theta; m_0, \phi_0) = \exp(\langle \phi_0, \theta \rangle - m_0 g(\theta) - h(m_0, \phi_0)). \quad (2)$$

Homework 9 Solutions

This model can be represented graphically as:



The generative model is given as follows:

1. For each topic $k \in \{1, \dots, K\}$
 - Draw $\theta_k \sim p(\cdot | m_0, \phi_0)$
2. Draw mixture weights $\pi \sim \text{Dirichlet}(\alpha)$
3. For each observation $i \in \{1, \dots, N\}$
 - Draw a component index $z_i \sim \text{Categorical}(\pi)$
 - Draw a datum $x_i \sim p(\cdot | \theta_{z_i})$

1 Expectation-Maximization on Mixture Models

Expectation-Maximization (EM) is an iterative algorithm for finding MLE or MAP estimates of parameters when the model depends on unobserved latent variables \mathbf{Z} . When the variables are fully observed except the model parameter θ , then we can simply run maximize the likelihood defined as $L(\theta; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z} | \theta)$ where $\mathbf{X} = \{x_i, i = 1, \dots, N\}$ and $\mathbf{Z} = \{z_i \in \{1, \dots, K\}, i = 1, \dots, N\}$. However, when \mathbf{Z} is not observed, we have to integrate out (or sum out) \mathbf{Z} and maximize over the collapsed likelihood i.e. $L(\theta; \mathbf{X}) = p(\mathbf{X} | \theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \theta)$.

The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood (ℓ) evaluated using the current estimate for the parameters (initially, random values), and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Question 1

E-step

Write down E-step in terms of π_k , the mixture component weight, and θ_k , the parameter for k -th component, for every component $k = 1, \dots, K$.

1. Write down the maximization function of posterior given π_k, θ_k .
2. Solve $n_{ik} = \mathbb{E}[\delta(z_i = k)]$ that maximizes the posterior from previous question.

Homework 9 Solutions

Answer

The posterior is

$$\begin{aligned} P(\theta, \pi | X, \alpha, \phi_0, m_0) &\propto P(X | \theta, \pi, \alpha, \phi_0, m_0) \overset{\text{(irrelevant)}}{P(\theta, \pi | \alpha, \phi_0, m_0)} \\ &= P(\theta, \pi | \alpha, \phi_0, m_0) \prod_i P(x_i | \theta, \pi). \end{aligned} \quad (3)$$

It is easier to deal with sums than products, so we can instead maximize the logarithm of (3) to obtain

$$\begin{aligned} \log [P(\theta, \pi | \alpha, \phi_0, m_0) \prod_i P(x_i | \theta, \pi)] \\ &= \log P(\theta, \pi | \alpha, \phi_0, m_0) + \sum_i \log \sum_k P(x_i, z_i = k | \theta, \pi) \\ &= \log P(\theta, \pi | \alpha, \phi_0, m_0) + \sum_i \log \sum_k Q(z_i = k | x_i) \frac{P(x_i, z_i = k | \theta, \pi)}{Q(z_i = k | x_i)} \\ &\geq \log P(\theta, \pi | \alpha, \phi_0, m_0) + \sum_i \sum_k Q(z_i = k | x_i) \log \frac{P(x_i, z_i = k | \theta, \pi)}{Q(z_i = k | x_i)} \\ &\triangleq F(Q, \theta, \pi) \end{aligned} \quad (4)$$

where the inequality above is given by Jensen's inequality. Note that, since

$$P(x_i, z_i = k | \theta, \pi) = P(z_i = k | x_i, \theta, \pi) P(x_i | \theta, \pi)$$

we have

$$\begin{aligned} F(Q, \theta, \pi) &= \log P(\theta, \pi | \alpha, \phi_0, m_0) + \sum_i \sum_k Q(z_i = k | x_i) \left[\log \frac{P(z_i = k | x_i, \theta, \pi)}{Q(z_i = k | x_i)} + \log P(x_i | \theta, \pi) \right] \\ &= -D_{KL}(Q(z_i = k | x_i) \| P(z_i = k | x_i, \theta, \pi)) + \sum_i \log P(x_i | \theta, \pi) \sum_k Q(z_i = k | x_i) \end{aligned} \quad (5)$$

which is maximized over Q when $D_{KL} = 0$, that is, when $Q(z_i = k | x_i) = P(z_i = k | x_i, \theta, \pi)$.

We have

$$\begin{aligned} Q_{ik}^{(t)} &= P(z_i = k | x_i, \pi^{(t)}, \theta^{(t)}) \\ &= \frac{P(x_i | z_i = k, \pi^{(t)}, \theta^{(t)}) P(z_i = k | \pi^{(t)}, \theta^{(t)})}{\sum_{j=1}^K P(x_i | z_i = j, \pi^{(t)}, \theta^{(t)}) P(z_i = j | \pi^{(t)}, \theta^{(t)})} \\ &= \frac{\exp[\langle \phi(x_i), \theta_k^{(t)} \rangle - g(\theta_k^{(t)})] \pi_k^{(t)}}{\sum_{j=1}^K \exp[\langle \phi(x_i), \theta_j^{(t)} \rangle - g(\theta_j^{(t)})] \pi_j^{(t)}}. \end{aligned} \quad (6)$$

Homework 9 Solutions

Question 2

M-step

1. Write down M-step as an optimization problem in terms of \mathbf{X} and $g(\cdot)$, ϕ_0 , m_0 , α and n_{ik} for every component $k = 1, \dots, K$.
2. Prove the convexity of the optimization problem.
3. Obtain closed form solution for M-step. Express it in terms of $\eta^{-1}(\cdot)$ and use the fact that $\eta(\theta) = \nabla_{\theta} g(\theta)$ is an invertible function.

Answer

The optimization problem is

$$\begin{aligned} \theta^{(t+1)}, \pi^{(t+1)} &= \underset{\theta, \pi}{\text{maximize}} F(Q^{(t)}, \theta, \pi) \\ \text{subject to} \quad &\sum_k \pi_k = 1. \end{aligned} \quad (7)$$

The objective functional can be rewritten (ignoring terms of the form $-Q \log Q$),

$$\begin{aligned} \log P(\theta, \pi \mid \alpha, \phi_0, m_0) + \sum_{i,k} Q_{ik}^{(t)} \log P(x_i, z_i = k \mid \theta, \pi) &= \\ = \log P(\theta, \pi \mid \alpha, \phi_0, m_0) + \sum_{i,k} Q_{ik}^{(t)} [\log P(x_i \mid z_i = k, \theta, \pi) + \log P(z_i = k \mid \theta, \pi)] &= \\ = \sum_k \log P(\theta_k \mid \phi_0, m_0) + \sum_k \log P(\pi_k \mid \alpha) + \sum_{i,k} Q_{ik}^{(t)} [\langle \phi(x_i), \theta_k \rangle - \mathbf{1}^\top g(\theta_k) + \log \pi_k] &= \\ \propto \sum_k \left(\alpha_k - 1 + \sum_i Q_{ik}^{(t)} \right) \log \pi_k + \sum_k \left[\left\langle \phi_0 + \sum_i Q_{ik}^{(t)} \phi(x_i), \theta_k \right\rangle - \left(m_0 + \mathbf{1} \sum_i Q_{ik}^{(t)} \right)^\top g(\theta_k) \right] &= \\ \triangleq \tilde{F}. & \end{aligned} \quad (8)$$

We can see that in θ_k , the objective is the difference between a linear function and a convex function (recall $g(\cdot)$ is convex because its second derivative is the variance), so the objective is *concave* in θ_k (which is what is desired for maximization). By the same token, $\log \pi_k$ is also concave, so the objective is concave in π_k provided $\alpha_k - 1 + \sum_i Q_{ik}^{(t)} > 0$, though there seems to be no guarantee of this unless we include the assumption that $\alpha \geq 1$.

The Lagrangian is

$$L(F; \theta, \pi, \beta) = F(Q^{(t)}, \theta, \pi) - \beta \left(\sum_k \pi_k - 1 \right). \quad (9)$$

(continued)

Homework 9 Solutions

Answer

(continued from previous page)

Denote $q_k^{(t)} = \sum_i Q_{ik}^{(t)}$. Notice that this quantity is the current estimate of the number of x_i drawn from distribution k .

Setting all partial derivatives equal to zero, we obtain

$$\begin{aligned}\frac{\partial F}{\partial \pi_k} &= \frac{\alpha_k - 1 + q_k}{\pi_k} &= \beta \\ \frac{\partial F}{\partial \theta_{kj}} &= (\phi_0 + \sum_i Q_{ik}^{(t)} \phi(x_i))_j - (m_0 + \mathbf{1}q_k^{(t)})^\top \frac{\partial g(\theta_k)}{\partial \theta_{kj}} &= 0 \\ \frac{\partial F}{\partial \beta} &= \sum_k \pi_k - 1 &= 0\end{aligned}\tag{10}$$

and hence

$$\begin{aligned}\theta_k^{(t+1)} &= \eta^{-1}(Z), \text{ where } m_{nk}^\top Z = \phi_{nk} \\ \pi_k^{(t+1)} &= \frac{\alpha_k - 1 + q_k^{(t)}}{\beta} \\ \text{where } \beta &= \sum_k (\alpha_k - 1 + q_k)\end{aligned}\tag{11}$$

where $\eta(\theta) = \nabla g(\theta)$ and $\eta^{-1}(\cdot)$ denotes its functional inverse; $m_{nk} = m_0 + \mathbf{1}q_k^{(t)}$ and $\phi_{nk} = \phi_0 + \sum_i Q_{ik}^{(t)} \phi(x_i)$.

The θ update can be rewritten using the Kronecker triple product formula

$$AXB = C \Leftrightarrow (B^\top \otimes A) \text{vec}(X) = \text{vec}(C)$$

to obtain

$$\theta_k^{(t+1)} = \eta^{-1}((\mathbf{I}_p \otimes m_{nk}^\top)^+ \phi_{nk})\tag{12}$$

where p is the dimension of θ and X^+ is the pseudoinverse of X .

Recall the Multivariate Normal distribution from homework 5 (http://alex.smola.org/teaching/10-701-15/homework/hw5_sol.pdf). The distribution of Multivariate Normal $\mathcal{N}(\mu, \Sigma)$ is given by

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)\tag{13}$$

where $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$ is a symmetric positive definite $d \times d$ matrix.

The conjugate prior for Multivariate Normal Distribution can be parametrized as the Normal Inverse Wishart Distribution $\mathcal{NIW}(\mu_0, \kappa_0, \Sigma_0, \nu_0)$. The distribution is given by:

$$\begin{aligned}p(\mu, \Sigma; \mu_0, \kappa_0, \Sigma_0, \nu_0) &= \mathcal{N}(\mu|\mu_0, \Sigma/\kappa_0) \mathcal{W}^{-1}(\Sigma|\Sigma_0, \nu_0) \\ &= \frac{\kappa_0^{\frac{d}{2}} |\Sigma_0|^{\frac{\nu_0}{2}} |\Sigma|^{-\frac{\nu_0+d+2}{2}}}{2^{\frac{(\nu_0+1)d}{2}} \pi^{\frac{d}{2}} \Gamma_d(\frac{\nu_0}{2})} e^{-\frac{\kappa_0}{2}(\mu - \mu_0)^\top \Sigma^{-1}(\mu - \mu_0) - \frac{1}{2} \text{tr}(\Sigma_0 \Sigma^{-1})}\end{aligned}\tag{14}$$

Homework 9 Solutions

Now, derive the Expectation-Maximization rules for the mixture of Multivariate Normal, $\mathcal{N}(\mu_k, \Sigma_k)$ for $k = 1, \dots, K$ and with the shared prior $\mathcal{N}\mathcal{I}\mathcal{W}(\mu_0, \kappa_0, \Sigma_0, \nu_0)$.

Question 3

Multivariate Normal

Write down update rules on E-step and M-step respectively for the Multivariate Gaussian with mixture parameters $\theta_k = (\mu_k, \Sigma_k)$. You can use the solutions from Question 1 and Question 2 and please refer to HW5 for Multivariate Gaussians.

Answer

E step:

$$Q_{ik}^{(t)} = \frac{\exp\left[-\frac{1}{2}(x_i - \mu_k^{(t)})^\top \Sigma_k^{(t)-1}(x_i - \mu_k^{(t)})\right] \pi_k^{(t)}}{\sum_{j=1}^k \exp\left[-\frac{1}{2}(x_i - \mu_j^{(t)})^\top \Sigma_j^{(t)-1}(x_i - \mu_j^{(t)})\right] \pi_j^{(t)}} \quad (15)$$

M step: First,

$$\pi_k^{(t+1)} = \frac{\alpha_k - 1 + \sum_i Q_{ik}^{(t)}}{\sum_{j=1}^k [\alpha_j - 1 + \sum_i Q_{ij}^{(t)}]}. \quad (16)$$

Now, in natural parameter space, $\theta_1 = \Sigma^{-1}\mu$ and $\theta_2 = -\frac{1}{2}\Sigma^{-1}$. Thus,

$$\begin{aligned} \Sigma &= -\frac{1}{2}\theta_2^{-1} \\ \mu &= -\frac{1}{2}\theta_2^{-1}\theta_1 \end{aligned} \quad (17)$$

and

$$g(\theta) = \left[\begin{array}{c} -\frac{1}{4}\theta_1^\top \theta_2^{-1} \theta_1 \\ \frac{d}{2} \log(2\pi) - \frac{1}{2} \log | -2\theta_2 | \end{array} \right]. \quad (18)$$

So

$$\frac{\partial g}{\partial \theta_1} = \begin{bmatrix} -\frac{1}{2}\theta_2^{-1}\theta_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \end{bmatrix} \quad \frac{\partial g}{\partial \theta_2} = \begin{bmatrix} \frac{1}{4}\theta_2^{-1}\theta_1\theta_1^\top\theta_2^{-1} \\ -\frac{1}{2}\theta_2^{-1} \end{bmatrix} = \begin{bmatrix} \mu\mu^\top \\ \Sigma \end{bmatrix}. \quad (19)$$

The derivative $\frac{\partial g_1}{\partial \theta_2}$ comes from the identity $\frac{\partial \text{tr}(X^{-1}A)}{\partial X} = -(X^{-1})^\top A(X^{-1})^\top$ and the invariance of the trace operator under cyclic permutation; see 'Wikipedia/Matrix calculus'. Also recall that

$$\phi_0 = \begin{pmatrix} \kappa_0 \mu_0 \\ \Sigma_0 + \kappa_0 \mu_0 \mu_0^\top \end{pmatrix} \quad m_0 = \begin{pmatrix} \kappa_0 \\ \nu_0 + d + 2 \end{pmatrix} \quad \phi(x) = \begin{pmatrix} x \\ xx^\top \end{pmatrix}. \quad (20)$$

(continued)

Homework 9 Solutions

Answer

(continued from previous page)

Combining all of this with (11) we see that

$$\begin{aligned}\frac{\partial F}{\partial \theta_1} &= \kappa_0 \mu_0 + \sum_i Q_{ik}^{(t)} x_i - (\kappa_0 + q_k^{(t)}) \mu_k^{(t+1)} = 0 \\ \frac{\partial F}{\partial \theta_2} &= \Sigma_0 + \kappa_0 \mu_0 \mu_0^\top + \sum_i Q_{ik}^{(t)} x_i x_i^\top - (\kappa_0 + q_k^{(t)}) \mu_k^{(t+1)} \mu_k^{(t+1)\top} - (\nu_0 + d + 2 + q_k^{(t)}) \Sigma_k^{(t+1)} = 0\end{aligned}\tag{21}$$

and hence

$$\begin{aligned}\mu_k^{(t+1)} &= \frac{\kappa_0 \mu_0 + \sum_i Q_{ik}^{(t)} x_i}{\kappa_0 + q_k^{(t)}} \\ \Sigma_k^{(t+1)} &= \frac{\Sigma_0 + \kappa_0 \mu_0 \mu_0^\top + \sum_i Q_{ik}^{(t)} x_i x_i^\top - (\kappa_0 + q_k^{(t)}) \mu_k^{(t+1)} \mu_k^{(t+1)\top}}{\nu_0 + d + 2 + q_k^{(t)}}.\end{aligned}\tag{22}$$

Homework 9 Solutions

2 Gibbs Sampling

In the last sub-problem we used EM for inference and now we turn to Gibbs Sampling, another popular method. Gibbs sampling is a variety of MCMC sampling in which we cycle through all our latent random variables, resampling each conditioned on the currently sampled values of all other random variables.

Question 4

Write out the simple Gibbs sampler for the mixture model, i.e. derive the following conditional probabilities:

- $p(z_i | \text{rest})$
- $p(\theta_k | \text{rest})$
- $p(\pi | \text{rest})$

in terms of $\alpha, m_0, \phi_0, \phi(\cdot), g(\cdot), h(\cdot, \cdot)$. An useful notation would be $n_k = |\{i : z_i = k\}|$ and $\phi_k = \phi_0 + \sum_{i:z_i=k} \phi(x_i)$. *Hint: Update equations from HW5 might come in handy!*

Answer

$$\begin{aligned} P(z_i | x_i, x_{-i}, z_{-i}, \pi, \theta) &= \frac{P(x_i | z_i, z_{-i}, x_{-i}, \pi, \theta) P(z_i | x_{-i}, z_{-i}, \theta, \pi)}{P(x_i | x_{-i}, z_{-i}, \pi, \theta)} \\ &= \frac{f(x_i; \theta_{z_i}) \pi_{z_i}}{\sum_k f(x_i; \theta_k) \pi_k} \end{aligned} \quad (23)$$

where $f(x; \theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$;

$$\begin{aligned} P(\theta_k | \mathbf{X}, \mathbf{Z}, \pi, \phi_0, m_0, \alpha) &= P(\theta_k | \{x_i : z_i = k\}, \phi_0, m_0) \\ &= \exp \left[\left\langle \begin{pmatrix} \phi_k \\ m_k \end{pmatrix}, \begin{pmatrix} \theta \\ -g(\theta) \end{pmatrix} \right\rangle - h(m_k, \phi_k) \right] \end{aligned} \quad (24)$$

where $m_k = m_0 + |\{i : z_i = k\}| = m_0 + n_k$;

$$\begin{aligned} P(\pi | \mathbf{X}, \mathbf{Z}, \theta, \phi_0, m_0, \alpha) &= P(\pi | \mathbf{Z}, \alpha) \\ &\propto P(\mathbf{Z} | \pi) P(\pi | \alpha) \\ &= \exp \left[\left\langle \begin{pmatrix} \vdots \\ |\{i : z_i = k\}| \\ \vdots \end{pmatrix}_{k=1 \dots K}, \log \pi \right\rangle \right] \exp[\langle \alpha - 1, \log \pi \rangle - \log \mathbf{B}(\alpha)] \\ &\propto \exp \left[\left\langle (n_1 - 1 \quad \dots \quad n_K - 1)^\top, \log \pi \right\rangle \right] \\ &\propto \text{Dirichlet}([\dots n_k \dots]^\top) \end{aligned} \quad (25)$$

and because proportional distributions are equal,

$$P(\pi | \text{rest}) = \exp[\langle \vec{n}_k - 1, \log \pi \rangle - \mathbf{B}(\vec{n}_k)]. \quad (26)$$

Homework 9 Solutions

However, we can do a better job by collapsing out θ_k and π . In general also, collapsing helps the chain mix faster (a consequence of Rao-Blackwell theorem).

Question 5

1. After collapsing out θ_k and π , what would be the Markov blanket for a given z_i ?
2. Using this information derive the conditional probability $p(z_i | \text{rest})$.
Hint: HW5 Posterior predictive might come in handy!

Answer

The Markov blanket for z_i becomes $x_{-i}, z_{-i}, x_i, \phi_0, m_0$, and α in this case. We obtain

$$\begin{aligned}
 P(z_i | \text{rest}) &= \frac{P(x_i | z_i, \{x_j : z_j = z_i\}, m_0, \phi_0) P(z_i | z_{-i}, \alpha)}{P(x_i | x_{-i}, m_0, \phi_0)} \\
 &= \frac{\exp[h(m_{z_i}, \phi_{z_i}) - h(m_{z_i} - 1, \phi_{z_i} - \phi(x_i))] \exp[\log \mathbf{B}(\vec{n}_k) - \log \mathbf{B}(\vec{n}_k - \vec{e}_{z_i})]}{\sum_{j=1}^K \exp[h(m_j, \phi_j) - h(m_j - 1, \phi_j - \phi(x_i))] \exp[\log \mathbf{B}(\vec{n}_k) - \log \mathbf{B}(\vec{n}_k - \vec{e}_j)]} \\
 &= \frac{\exp[h(m_{z_i}, \phi_{z_i}) - h(m_{z_i} - 1, \phi_{z_i} - \phi(x_i))] \frac{n_{z_i} - 1}{(\sum_k n_k) - 1}}{\sum_{j=1}^K \exp[h(m_j, \phi_j) - h(m_j - 1, \phi_j - \phi(x_i))] \frac{n_j - 1}{(\sum_k n_k) - 1}} \\
 &= \frac{\exp[h(m_{z_i}, \phi_{z_i}) - h(m_{z_i} - 1, \phi_{z_i} - \phi(x_i))] (n_{z_i} - 1)}{\sum_{j=1}^K \exp[h(m_j, \phi_j) - h(m_j - 1, \phi_j - \phi(x_i))] (n_j - 1)} \tag{27}
 \end{aligned}$$

Note that $m_k - m_0 = n_k - \alpha_k$. From here on out, $\tilde{n}_k = |\{i : z_i = k\}|$ so that $m_k = m_0 + \tilde{n}_k$ and $n_k = \tilde{n}_k + \alpha$.

Homework 9 Solutions

Thus, we need to take care of only two invariants, namely n_k and ϕ_k per component in the inference procedure and a neat program can be written.

Question 6

Your task is to complete the generic program given in Algorithm 1 for Gibbs sampling of mixture models for T iterations.

Algorithm 1 Collapsed Gibbs sampling for mixture models

- 1: Initialize z randomly and evaluate initial counts \tilde{n}_k and statistics ϕ_k .
- 2: $t \leftarrow 0$
- 3: **while** $t \leq T$ **do**
- 4: **for** $i = 1 \rightarrow N$ **do**
- 5: Remove datum from current component and update statistics: $\tilde{n}_{z_i} \leftarrow \tilde{n}_{z_i} - 1, \phi_{z_i} \leftarrow \phi_{z_i} - \phi(x_i)$
- 6: Sample z_i using the PMF stored in
 $p[k] \leftarrow (\alpha + \tilde{n}_k - 1) \exp(h(m_0 + \tilde{n}_k + 1, \phi_k + \phi(x_i)) - h(m_0 + \tilde{n}_k, \phi_k)); p \leftarrow p / \text{sum}(p);$
- 7: Add datum to the new component and update statistics: $\tilde{n}_{z_i} \leftarrow \tilde{n}_{z_i} + 1, \phi_{z_i} \leftarrow \phi_{z_i} + \phi(x_i)$
- 8: **end for**
- 9: $t \leftarrow t + 1$
- 10: **end while**

Homework 9 Solutions

Last but not the least, we derive Gibbs sampler for the Multivariate Gaussian with a Normal Inverse Wishart. Follow the notations of previous sub-problem or HW5.

Question 7

Write out the Gibbs Sampler, i.e. $p(z_i | \text{rest})$ explicitly in terms of data and hyper-parameters $\alpha, \mu_0, \kappa_0, \Sigma_0, \nu_0$. What would be the posterior means of mixture parameters $\theta_k \Rightarrow \{\mu_k, \Sigma_k\}$ given a sample \mathbf{Z} obtained from our favourite collapsed Gibbs sampling algorithm?

Answer

First, apply the downdate equations in the following order:

$$\begin{aligned} \kappa_{z_i} &\leftarrow \kappa_{z_i} - 1, & \nu_{z_i} &\leftarrow \nu_{z_i} - 1 \\ \mu_{z_i} &\leftarrow \frac{(\kappa_{z_i} - 1)\mu_{z_i}}{\kappa_{z_i}} - x_i \\ \Sigma_{z_i} &\leftarrow \Sigma_{z_i} - \frac{\kappa_{z_i}}{\kappa_{z_i} - 1} (x_i - \mu_{z_i})(x_i - \mu_{z_i})^\top \end{aligned} \quad (28)$$

Directly using Q8 from HW 5, update z_i by sampling from the distribution:

$$\begin{aligned} P(z_i = k | \text{rest}) &= \frac{P(x_i | z_i = k, x_{-i})(n_k - 1)}{\sum_j P(x_i | z_i = j, x_{-i})(n_j - 1)} \\ &= \frac{t_{\nu_k - d + 1} \left(x_i \mid \mu_k, \frac{(\kappa_k + 1)\Sigma_k}{\kappa_k(\nu_k - d + 1)} \right) (n_k - 1)}{\sum_{j=1}^K t_{\nu_j - d + 1} \left(x_i \mid \mu_j, \frac{(\kappa_j + 1)\Sigma_j}{\kappa_j(\nu_j - d + 1)} \right) (n_j - 1)} \end{aligned} \quad (29)$$

then apply the update equations in the following order:

$$\begin{aligned} \Sigma_{z_i} &\leftarrow \Sigma_{z_i} + \frac{\kappa_{z_i}}{\kappa_{z_i} - 1} (x_i - \mu_{z_i})(x_i - \mu_{z_i})^\top \\ \mu_{z_i} &\leftarrow \frac{\kappa_{z_i}\mu_{z_i} + x_i}{\kappa_{z_i} + 1} \\ \nu_{z_i} &\leftarrow \nu_{z_i} + 1, & \kappa_{z_i} &\leftarrow \kappa_{z_i} + 1 \end{aligned} \quad (30)$$

At the end of the procedure, we obtain

$$\begin{aligned} \hat{\mu}_k &= \frac{\kappa_0\mu_0 + \sum_{i:z_i=k} x_i}{\kappa_0 + \tilde{n}_k} \\ \hat{\Sigma}_k &= \frac{\Sigma_0 + \sum_{i:z_i=k} x_i x_i^\top + \kappa_0\mu_0\mu_0^\top - (\kappa_0 + \tilde{n}_k)\mu_k\mu_k^\top}{\nu_0 + \tilde{n}_k - d - 1} \end{aligned} \quad (31)$$

where again $\tilde{n}_k = |\{i : z_i = k\}|$.

Homework 9 Solutions

Hopefully, this homework threw light on the commonly used finite mixture models. In all our examples so far, we have chosen some fixed K to be the dimension of π , which effectively enforces a fixed number of clusters. However, we would like to point out a major drawback of this model: the number of clusters present in the data needs to be known apriori, which is not the case often in practice. A novel workaround is to allow unboundedly many clusters, i.e. take $K \rightarrow \infty$. This yields a Bayesian nonparametric model known as the Dirichlet Process mixture. Extending our samplers to sample from a mixture of potentially infinitely many components requires only minor changes to the code; see e.g. section 3 of this nice paper http://www.stat.columbia.edu/npbayes/papers/neal_sampling.pdf.