Assignment 5

Laurence R Freeman

December 2022 - Feb 2023

1 Question 1: [70 marks] Mean-field learning

$\mathcal{F}(q,\Theta) = \sum_n^N \left\langle log\ p(s^{(n)}, x^{(n)}|\Theta)_{q(s^{(n)})} \right\rangle - \left\langle log\ q(s^{(n)})_{q(s^{(n)})} \right\rangle$

$= \sum_n^N \left\langle log\ p(x^{(n)}|s^{(n)},\Theta)_{q(s^{(n)})} \right\rangle + \sum_n^N \left\langle log\ p(s^{(n)}|\Theta)_{q(s^{(n)})} \right\rangle - \sum_n^N \left\langle log\ q(s^{(n)})_{q(s^{(n)})} \right\rangle$

$= log(\ (2\pi\sigma^2)^{-D/2} exp[-\frac{1}{2\sigma^2}(x^{(n)} - \sum_{i=1}^K s_i^{(n)}\mu_i)^T(x^{(n)} - \sum_{i=1}^K s_i^{(n)}\mu_i)]\ )_{q(s^{(n)})}$

$+ log(\prod_{i=1}^K \pi_i^{s_i^{(n)}}(1-\pi_i)^{(1-s_i^{(n)})})_{q(s^{(n)})} - log(\prod_{i=1}^K \lambda_i^{s_i^{(n)}}(1-\lambda_i)^{(1-s_i^{(n)})})_{q(s^{(n)})}$

$= \ -\frac{D}{2}log(2\pi) - Dlog(\sigma)$

$-\frac{1}{2\sigma^2}\left[\sum_{n=1}^N x^{(n)T}x^{(n)} + \sum_{i,j}^K \mu_i^T\mu_j \sum_{n=1}^N \left\langle s_i^{(n)}s_j^{(n)} \right\rangle_{q(s^{(n)})} - 2\sum_i \mu_i^T \sum_{n=1}^N \left\langle s_i^{(n)} \right\rangle_{q(s^{(n)})} x^{(n)}\right]$

$+ \sum_i^K \left\langle s_i^{(n)} \right\rangle\ log\frac{\pi_i}{\lambda_{in}} + (1 - \left\langle s_i^{(n)} \right\rangle) log\frac{(1-\pi_i)}{(1-\lambda_{in})}$

Where the expectation of s_i is λ_i and the expectation of $\langle s_i s_j \rangle$ is $\lambda_i * \lambda_j$ when i!=j and λ_i when i=j. Then, setting this equation to zero and taking the derivative w.r.t to a single λ_{in} we get:

$$\frac{\delta\mathcal{F}}{\delta\lambda_i} = log\frac{\pi_i}{1-\pi_i} - log\frac{\lambda_i}{1-\lambda_i} - \frac{\mu_i^T\mu_i - 2\mu_i^T x^{(n)} + 2\sum_{i,j:i\neq j}^K \mu_i^T\mu_j\lambda_{jn}}{2\sigma^2}$$

$$\lambda_{in} = f(log\frac{\pi_i}{1-\pi_i} + \frac{1}{\sigma^2}(y - \sum_{j\neq i}^K \lambda_{jn}\mu_j)^T\mu_i - \frac{1}{2\sigma^2}\mu_i^T\mu_i)$$

Thought process

One of the most popular forms of variational inference (i.e: the existence of an approximation gap between proposed and true posterior) is called the mean field approximation in which the posterior is of a fully factorised form:

$$q(x) = \prod_{i=1}^{D} q_i(x_i)$$

In this case our goal here is to solve the reverse KL problem:

$$Argmin_{q_1 \dots q_D} \; KL(q \;||\; p)$$

When computing the likelihood for a given data point, a common approach is to leave out the factor corresponding to that data point and calculate the likelihood using all other factors in the hope that this approximates the true log likelihood (where $\tilde{p}$ is the unnormalised posterior):

$$log \; q_j(x_j) = E_{-q_j}[log \; \tilde{p}(x)] + const$$

Given that we are using Jensen's inequality and working with maximising a lower bound. Let us first derive the free energy formula.

$$\mathcal{F}(q, \Theta) = \sum_{n}^{N} \Big\langle log \; p(s^{(n)}, x^{(n)}|\Theta)_{q(s^{(n)})} \Big\rangle - \Big\langle log \; q(s^{(n)})_{q(s^{(n)})} \Big\rangle$$

Noting that:

$$P(A, B|C) = P(A|B, C)P(B|C)$$

We can reorder the free energy as:

$$\mathcal{F}(q, \Theta) = \sum_{n}^{N} \Big\langle log \; p(x^{(n)}|s^{(n)}, \Theta)_{q(s^{(n)})} \Big\rangle + \Big\langle log \; p(s^{(n)}|\Theta) \Big\rangle - \Big\langle log \; q(s^{(n)})_{q(s^{(n)})} \Big\rangle$$

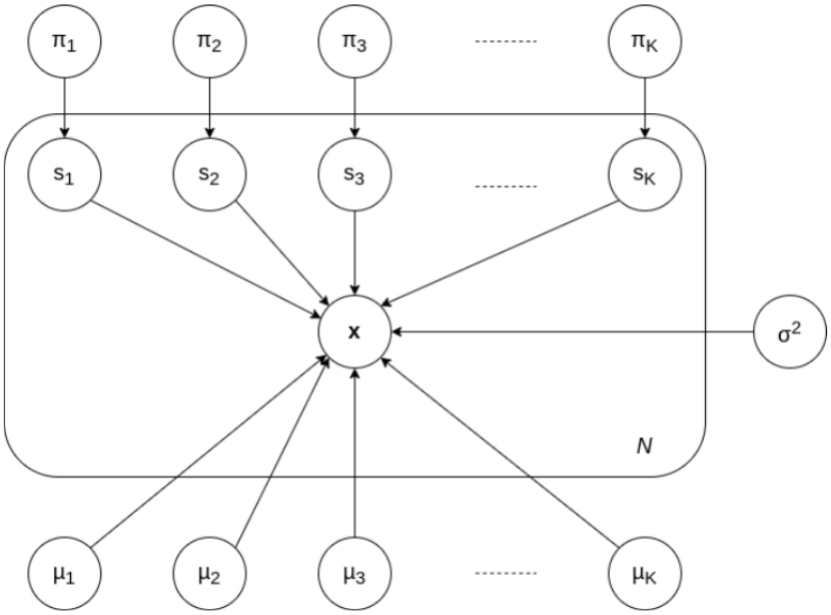
Given that these three components are provided to us in the homework. Such that:

$$q_n(s^{(n)}) = \prod_{i=1}^{K} \lambda_{in}^{s_i^{(n)}} (1 - \lambda_{in})^{(1 - s_i^{(n)})}$$

$$p(s|\pi) = \prod_{i=1}^{K} \pi_i^{s_i} (1 - \pi_i)^{(1 - s_i)}$$

$$p(x|s_1, \dots s_K, \mu, \sigma^2) = \mathcal{N}(\sum_i s_i \mu_i, \sigma^2 I)$$

Also noting that the model can be represented as a DAG as a helpful guide when deriving the math:



We can now substitute each component into the free energy equation to produce:

$$\mathcal{F}(q,\Theta) = log((2\pi\sigma^2)^{-D/2} exp[-\frac{1}{2\sigma^2}(x^{(n)} - \sum_{i=1}^{K} s_i^{(n)}\mu_i)^T(x^{(n)} - \sum_{i=1}^{K} s_i^{(n)}\mu_i)])$$

$$+ log(\prod_{i=1}^{K} \pi_i^{s_i^{(n)}}(1-\pi_i)^{(1-s_i^{(n)})}]) - log(\prod_{i=1}^{K} \lambda_i^{s_i^{(n)})}(1-\lambda_i)^{(1-s_i^{(n)})}])$$

Applying the log and expanding out the brackets we get:

$$\mathcal{F}(q,\Theta) = \; -\frac{D}{2}log(2\pi) - Dlog(\sigma)$$

$$-\frac{1}{2\sigma^2}\left[\sum_{n=1}^{N} x^{(n)T}x^{(n)} + \sum_{i,j}^{K}\mu_i^T\mu_j\sum_{n=1}^{N}\left\langle s_i^{(n)}s_j^{(n)}\right\rangle_{q(s^{(n)})} - 2\sum_i \mu_i^T\sum_{n=1}^{N}\left\langle s_i^{(n)}\right\rangle_{q(s^{(n)})} x^{(n)}\right]$$

$$+\sum_{i}^{K}\left\langle s_i^{(n)}\right\rangle \; log\frac{\pi_i}{\lambda_{in}} + (1-\left\langle s_i^{(n)}\right\rangle)log\frac{(1-\pi_i)}{(1-\lambda_{in})}$$

Given that the expectation of a Bernoulli is it's probability term. And that given a output of a Bernoulli is a random variable and thus this is why we are working with expectations. The expectation of s_i is λ_i and the expectation of $\langle s_i s_j\rangle$ is $\lambda_i * \lambda_j$ when i!=j and λ_i when i=j. And the expectation of s_i^2 is just λ_i. Focusing on that there are two conditions contained within this sum product sum within the brackets. Let us split out those conditions:

$$-\frac{1}{2\sigma^2}\left[\sum_{n=1}^{N} x^{(n)T}x^{(n)} + \sum_{i,j\;i\neq j}^{K}\mu_i^T\mu_j\sum_{n=1}^{N}\lambda_{in}\lambda_{jn} + \sum_{i,j\;i=j}^{K}\mu_i^2\sum_{n=1}^{N}\lambda_{in} \; - 2\sum_{i}^{K}\mu_i^T\sum_{n=1}^{N}\lambda_{in}\; x^{(n)}\right]$$

Here we have also replaced the s_i^n values with λ_{in}. Then the derivative of the first breakout term for a single lambda is the below, where the 2 is because you are cycling through pairs of index's for i and j. Because $(i*j)*\lambda == (j*i)*\lambda$ so it's doubled. The sum over N disspears because the derivative is for a single $lambda_{in}$:

$$\frac{\partial}{\partial\lambda_{in}} = 2\sum_{i,j,i\neq j}^{K}\mu_i^T\mu_j\lambda_{jn}$$

Where λ_i disappears because it is treated as a 1 as we are doing a derivative w.r.t to that term. Also note λ_j remains because we are not doing a derivative w.r.t to that term so it's a constant. The derivative of the second breakout term becomes:

$$\frac{\partial}{\partial\lambda_{in}} = \mu_i^2$$

Continuing calculating the derivatives of single terms we look at terms outside of the bracket:

$$\sum_{i}^{K} \left\langle s_i^{(n)} \right\rangle \ log \frac{\pi_i}{\lambda_{in}}$$

$$\Rightarrow \sum_{i}^{K} \lambda_{in} \ log(\pi_i) - \sum_{i}^{K} \lambda_{in} \ log(\lambda_{in})$$

$$\frac{\partial}{\partial \lambda_{in}} = log(\pi_i) - (1 + log(\lambda_{in}))$$

Where the first term λ_{in} disappears because derivative w.r.t to that term and the log remains because it's a constant. The second term uses the product rule of differentiation. Now let us calculate the derivative of the next log term:

$$+(1 - \left\langle s_i^{(n)} \right\rangle) log \frac{(1 - \pi_i)}{(1 - \lambda_{in})}$$

$$\Rightarrow + \ (1 - \lambda_{in}) log(1 - \pi_i) - (1 - \lambda_{in}) log(1 - \lambda_{in})$$

$$\frac{\partial}{\partial \lambda_{in}} = - \ log(1 - \pi_i) + (1 - log(1 - \lambda_{in}))$$

Now placing these components together and simplifying we get:

$$\frac{\delta \mathcal{F}}{\delta \lambda_i} = log \frac{\pi_i}{1 - \pi_i} - log \frac{\lambda_i}{1 - \lambda_i} - \frac{\mu_i^T \mu_i - 2 \mu_i^T x^{(n)} + 2 \sum_{i,j:i \neq j}^{K} \mu_i^T \mu_j \lambda_{jn}}{2 \sigma^2}$$

$$\frac{\delta \mathcal{F}}{\delta \lambda_i} = log \frac{\pi_i}{1 - \pi_i} - log \frac{\lambda_i}{1 - \lambda_i} + \frac{1}{\sigma^2} (x - \sum_{j \neq i} \lambda_{jn} \mu_j)^T \mu_i - \frac{1}{2 \sigma^2} \mu_i^T \mu_i$$

Rearranging the λ_{in} to one side by setting the derivative to zero allows us to formulate:

$$\lambda_{in} = f(log \frac{\pi_i}{1 - \pi_i} + \frac{1}{\sigma^2} (y - \sum_{j \neq i}^{K} \lambda_{jn} \mu_j)^T \mu_i - \frac{1}{2 \sigma^2} \mu_i^T \mu_i)$$

Where f is the logistic sigmoid function. $f(x) = 1/(1 + exp(x))$. Thus we have a formula for computing the free energy in the expectation step. Also a formula for calculating each λ_{in} using all other factors as is typical in mean field learning.