CS181 Practice Questions: Expectation Maximization and HMMs

1. **Expectation and Maximization (Bishop 9.15)**

   Show that if we maximize the expected complete-data log likelihood function (Bishop Eq. (9.55)) for a mixture of Bernoulli distributions with respect to μ_k , we obtain the M step equation (9.59).

   We calculate derivatives of 9.55, set them to zero, and solve for μ_{ki}:

$$\frac{d}{d\mu_{ki}} E[\ln p(X, Z|\mu, \pi)] = \sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}}\right))$$

$$= \frac{\sum_n \gamma(z_{nk}) x_{ni} - \sum_n \gamma(z_{nk}) \mu_{ki}}{\mu_{ki}(1 - \mu_{ki})}$$

   We set equal to zero, solve and get what (9.59):

$$\mu_{ki} = \frac{\sum_n \gamma(z_{nk}) x_{ni}}{\sum_n \gamma(z_{nk})}$$

2. **E and M (Bishop 9.20)**

Show that maximization of the expected complete-data log likelihood function (Bishop Eq. (9.62)) for the Bayesian linear regression model leads to the M step re-estimation result Eq. (9.63) for α.

We take the derivatives of (9.62) w.r.t. α:

$$\frac{d}{d\alpha} E[\ln p(t, x|\alpha, \beta)] = \frac{M}{2\alpha} - \frac{1}{2} E[w^T w]$$

If you set equal to zero and rearrange, you get (9.63).

3. **E and M**

Show that if we maximize the expected complete-data log-likelihood function given in eq. 9.55 for a mixture of Bernoulli's with respect to μ_k, we obtain the M-step equation 9.59.

Calculate the derivatives of 9.55 and set them to 0. We see that

$$0 = \frac{\partial}{\partial \mu_{ki}} \mathbb{E}[\ln p(X, Z \mid \mu, \pi)] = \frac{\sum_n \gamma(z_{n,k}) x_{n,i} - \sum_n \gamma(z_{n,k}) \mu_{k,i}}{\mu_{k,i}(1 - \mu_{k,i})}$$

If we solve for μ_{ki} we get

$$\frac{\sum_n \gamma(z_{n,k}) x_{n,i}}{\sum_n \gamma(z_{n,k})}$$

which is equivalent to 9.59.

4. **E and M**

Show that if we maximize the expected complete-data log-likelihood function given in eq. 9.55 for a mixture of Bernoulli's with respect to the mixing coefficients π_k, using a Lagrange multiplier to enforce to summation constraint (they must sum to 1), we obtain the M-step equation 9.60.

First, we introduce the Lagrange multiplier to enforce $\sum_k \pi_k = 1$. Then, our objective becomes

$$\mathbb{E}[\ln p(X, Z \mid \mu, \pi)] + \lambda \sum_k (\pi_k - 1)$$

if we differentiate this w.r.t π_k, we get

$$\sum_{n=1}^{N} \gamma(z_{n,k}) \frac{1}{\pi_k} + \lambda = \frac{N_k}{\pi_k} + \lambda$$

If we set this to zero, wee see that $N_k = -\pi_k \lambda$. If we sum this over all k's, because of the summation constraint, we have that $\pi_k = N_k / N$ which is equivalent to 9.60.

5. **Bernoulli Mixtures**

Consider the joint distribution of latent and observed variables for the Bernoulli distribution obtained by forming the product of the $p(\boldsymbol{x} \mid z, \mu)$ given by 9.52 and $p(z \mid \pi)$ given by 9.53. Show that if we marginalize this joint distribution with respect to z (i.e., sum over all possible choices for z), we obtain 9.59.

The product of the distributions gives us

$$\prod_{k=1}^{K} (p(\boldsymbol{x} \mid \mu_k)\pi_k)^{z_k}$$

And if we marginalize out the z's, we get

$$\sum_{z} \prod_{k=1}^{K} (p(\boldsymbol{x} \mid \mu_k)\pi_k)^{z_k} = \sum_{j=1}^{K} \prod_{k=1}^{K} (p(\boldsymbol{x} \mid \mu_k)\pi_k)^{I_{j,k}} = \sum_{j=1}^{K} \pi_j p(\boldsymbol{x} \mid \mu_j),$$

the desired result.

6. **When to Use HMMs (CMU)**

For each of the following scenarios, is it appropriate to use a hidden markov model to model the dataset? Why or why not.

(a) Stock market price data

(b) Recommendations on a database of movie reviews (like the book reviews from the first practical)

(c) Daily precipitation data in Boston

(d) Optical character recognition

(a) Stock market price data Yes, stock market data is time-dependent.

(b) Recommendations on a database of movie reviews (like the book reviews from the first practical) No, we don't expect user preferences to change much over time.

(c) Daily precipitation data in Boston Yes, precipitation today is very likely to affect the chance of precipitation tomorrow.

(d) Optical character recognition, where we are identifying words Yes, word recognition is very dependent upon the sequence of characters.

7. **E-M For HMM's (Bishop 13.6)**

Show that if any elements of the parameters π (start probability) or A (transition probability) for a hidden Markov model are initialized to 0, then those elements will remain 0 in all subsequent updates of the EM algorithm.

Suppose a particular element π_k of π has been initialized to 0. In the first E-step, the quantity $\alpha(z_{1k})$ is given by

$$\alpha(z_{ik}) = \pi_k p(\boldsymbol{x}_1 \mid \phi_k) = 0$$

where we have defined α as in Bishop eq. 13.34. Then, $\gamma(z_{1k})$ will also be zero, so in the next M-step the new value of π_k will again be 0. Since this is true for any EM cycle, this will remain zero throughout.

Now, suppose that A_{jk} is initialized to 0. Since $A_{jk} = p(z_{nk} \mid z_{n-1,j})$, we see that $\eta(z_{n-1,j}, z_P n, k) = 0$ (by eq. 13.43). In the subsequent M-step the new value of A_{jk} is given by 13.19 (which sums over the $\eta(z_{n-1,j,z_{nk}}))$ must also be 0, this it will remain 0 for all subsequent update steps.

8. **E-M For HMM's (Bishop 13.5)**

Verify the M-step equation 13.18 (the update rule for π_k) for the initial state probabilities of the hidden Markov model by maximization of the expected complete-data log likelihood function (given in eq. 13.17), using Lagrange multipliers to enforce the summation constraint on the components of π.

Our summation constraints is simply that $\sum_k \pi_k = 1$. Since we are only maximizing with respect to π_k, we can omit the terms from $Q(\theta, \theta_{old})$ which are independent of π and add a Lagrange multiplier term to enforce the constraint, giving the following objective

$$\sum_{k=1}^{K} \gamma(z_{1k} \ln \pi_k + \lambda(\sum_{k=1}^{K} -1)$$

If we take the derivative with respect to π_k and set the result equal to 0, we get

$$0 = \gamma(z_{1k})\frac{1}{\pi_k} + \lambda$$

Now, if we multiply through by π_k and sum over all k, we get

$$\lambda = -\sum_{k=1}^{K} \gamma(z_{1k})$$

Now, if we substitute this value into the previous equation, we get that

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}$$