SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Homework #7

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *December 18th, 2019*

Homework Submission Instructions. Please write up your responses to the following problems clearly and concisely. We require you to write up your responses with A4 paper. You are allowed and encouraged to work together. You may discuss the homework to understand the problem and reach a solution in groups. However, *each student must write down the solution independently*. You must understand the solution well enough in order to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

- **Written Homeworks.** All calculation problems **MUST** be written on single-sided A4 paper. You should bring and hand in it before class on the day of the deadline. Submitting the scan or photo version on Sakai will **NOT** be accepted.
- **Coding Homeworks.** All coding assignments will be done in Jupyter Notebooks. We will provide a `.ipynb` template for each assignment. Your final submission will be a `.ipynb` file with your answers and explanations (you should know how to write in Markdown or LaTeX). Make sure that all packages you need are imported at the beginning of the program, and your `.ipynb` file should **work step-by-step without any error**.

Question 1

Consider a density model given by a mixture distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k)$$

and suppose that we partition the vector $\mathbf{x}$ into two parts so that $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$. Show that the conditional density $p(\mathbf{x}_b|\mathbf{x}_a)$ is itself a mixture distribution and find expressions for the mixing coefficients and for the component densities.

Answer. For the mixture model the joint distribution can be written

$$p(\mathbf{x}_a, \mathbf{x}_b) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_a, \mathbf{x}_b|k).$$

We can find the conditional density $p(\mathbf{x}_a|\mathbf{x}_b)$ by making use of the relation

$$p(\mathbf{x}_a|\mathbf{x}_b) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_a)}.$$

For mixture model the marginal density of $\mathbf{x}_a$ is given by

$$p(\mathbf{x}_a) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_a|k)$$

where

$$p(\mathbf{x}_a|k) = \int p(\mathbf{x}_a, \mathbf{x}_b|k) \mathrm{d}\mathbf{x}_b.$$

Thus we can write the conditional density in the form

$$p(\mathbf{x}_b|\mathbf{x}_a) = \frac{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_a, \mathbf{x}_b|k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_a|j)}.$$

Now we decompose the number using

$$p(\mathbf{x}_a, \mathbf{x}_b|k) = p(\mathbf{x}_b|\mathbf{x}_a, k) p(\mathbf{x}_a|k)$$

which allows us finally to write the conditional density as a mixture model of the form

$$p(\mathbf{x}_b|\mathbf{x}_a) = \sum_{k=1}^{K} \lambda_k p(\mathbf{x}_b|\mathbf{x}_a, k)$$

where the mixture coefficients are given by

$$\lambda_k \equiv p(k|\mathbf{x}_a) = \frac{\pi_k p(\mathbf{x}_a|k)}{\sum_j \pi_j p(\mathbf{x}_a|j)}.$$

Question 2

Consider a Bernoulli mixture model as discussed in Section 9.3.3, together with a prior distribution $p(\mu_k|a_k, b_k)$ over each of the parameter vectors μ_k given by the beta distribution (2.13), and a Dirichlet prior $p(\pi|\alpha)$ given by (2.38). Derive the EM algorithm for maximizing the posterior probability $p(\mu, \pi|X)$

Answer. When dealing withMAP estimation for a general mixture model, we know that the E-step will remain unchanged. In the M-step we maximize

$$\mathcal{Q}(\theta, \theta^{old}) + \ln p(\theta)$$

which in the case of the given model becomes,

$$\begin{aligned}
&\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})]\} \\
&+ \sum_{j=1}^{N} \sum_{i'=1}^{D} \{(a_i - 1) \ln \mu_{ji'} + (b_j - 1) \ln(1 - \mu_{ji'})\} + \sum_{l=1}^{K} (\alpha_l - 1) \ln \pi_l \quad (1)
\end{aligned}$$

where we have used (9.55), (2.13) and (2.38), and we have dropped terms independent of $\{\mu_k\}$ and π. Note that we have assumed that each parameter μ_{ki} has the same prior for each i, but this can differ for different components k. Differenting (1) w.r.t. μ_{ki} yields

$$\sum_{n=1}^{N} \gamma(z_{nk})\{\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\} + \frac{a_k}{\mu_{ki}} - \frac{1-b_k}{1-\mu_{ki}}$$
$$= \frac{N_k\bar{x}_{ki} + a - 1}{\mu_{ki}} - \frac{N_k - N_k\bar{x}_{ki} + b - 1}{1-\mu_{ki}}$$

where N_k is given by (9.57) and $\bar{x}_{ki}$ is the i^{th} element of $\bar{x}$ defined in (9.58). Setting this equal to zero and rearranging, we get

$$\mu_{ki} = \frac{N_k\bar{x}_{ki} + a - 1}{N_k + a - 1 + b - 1}. \quad (2)$$

Note that if $a_k = b_k = 1$ for all k, this reduces to the standard maximum likelihood result. Also, as N becomes large, (2) will approach the maximum likelihood result. When maximizing w.r.t. π_k, we need to enforce the constraint $\sum_k \pi_k - 1$, which we do by adding a Lagrange multiplier term to (1). Dropping terms independent of π we are left with

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\ln\pi_k + \sum_{l=1}^{K}(\alpha_l - 1)\ln\pi_l + \lambda(\sum_{j=1}^{K}\pi_j - 1).$$

Differentiating this w.r.t. π_k, we get

$$\frac{N_k + \alpha_k - 1}{\pi_k} + \lambda$$

and setting this equal to zero and rearranging, we have

$$N_k + \alpha_k - 1 = -\lambda\pi_k.$$

Summing both sides over k, using $\sum_k \pi_k = 1$, we see that $-\lambda = N + \alpha_0 - K$, where α_0 is given by (2.39), and thus

$$\pi_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K} \quad (3)$$

Also in this case, if $\alpha_k = 1$ for all k, we recover the maximum likelihood result exactly. Similarly, as N gets large, (3) will approach the maximum likelihood result.

Question 3

Consider the incremental form of EM algorithm for a mixture of Gaussians, in which the responsibilities are recomputed only for a specific data point $\mathbf{x}_m$. Starting from the M-step formular (9.17) and (9.18), derive the results (9.78) and (9.79) for updating the component means.

Answer. From (9.18) we get

$$N_k^{old} = \sum_n \gamma^{old}(z_{nk}).$$

We get N_k^{new} by recomputing the responsibilities, $\gamma(z_{mk})$, for a specific data point, $\mathbf{x}_m$, yielding

$$N_k^{new} = \sum_{n \neq m} \gamma^{old}(z_{nk}) + \gamma^{new}(z_{mk}).$$

Combining this with N_k^{old}, we get (9.79). Similarly, from (9.17) we have

$$\mu_k^{old} = \frac{1}{N_k^{old}} \sum_n \gamma^{old}(z_{nk})\mathbf{x}_n.$$

and recomputing the responsibilities, $\gamma(z_{mk})$, we get

$$\begin{aligned}
\mu_k^{new} &= \frac{1}{N_k^{new}}(\sum_{n \neq m} \gamma^{old}(z_{nk})\mathbf{x}_n + \gamma^{new}(z_{mk})\mathbf{x}_m) \\
&= \frac{1}{N_k^{new}}(N_k^{old}\mu_k^{old} - \gamma^{old}(z_{mk})\mathbf{x}_m + \gamma^{new}(z_{mk})\mathbf{x}_m) \\
&= \frac{1}{N_k^{new}}((N_k^{new} - \gamma^{new}(z_{mk}) + \gamma^{old}(z_{mk}))\mu_k^{old} - \gamma^{old}(z_{mk})\mathbf{x}_m + \gamma^{new}(z_{mk})\mathbf{x}_m) \\
&= \mu_k^{old} + (\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}})(\mathbf{x}_m - \mu_k^{old}),
\end{aligned}$$

where we have used (9.79).

Question 4

Consider the K-means algorithm discussed in book Section 9.1. Show that as a consequence of there being a finite number of possible assignments for the set of discrete indicator variables r_{nk}, and that for each such assignment there is a unique optimum for the μ_k, the K-means algorithm must converge after a finite number of iterations.

Answer. Since both the E-step and the M-step minimise the distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}||\mathrm{x}_n - \mu_k||^2 \tag{1}$$

the algorithm will never change from a particular assignment of data points to prototypes, unless the new assignment has a lower value for (1).

Since there is a finite number of possible assignments, each with a corresponding unique minimum of (1) w.r.t. the prototypes, $\{\mu_k\}$, the K-means algorithm will converge after a finite number of steps, when no re-assignment of data points to prototypes will result in a decrease of (1). When no-reassignment takes place, there also will not be any change in $\{\mu_k\}$.

Question 5

Imagine a class where the probability that a student gets an A grade is P (A) = 1/2, a B grade $P(B) = \mu$, a C grade $P(C) = 2\mu$, and a D grade $P(D) = 1/2 - 3\mu$. We are told that c students get a C and d students get a D. We dont know how many students got exactly an A or exactly a B. But we do know that h students got either an a or b. Therefore, a and b are unknown values where a + b = h. Our goal is to use expectation maximization to obtain a maximum likelihood estimate of μ.

(a) Expectation step: Compute the expected values of a and b given μ.

(b) Maximization step: Given the expected values of a and b, compute the maximum likelihood estimate of μ.

Hint. Compute the MLE of μ assuming unobserved variables are replaced by their expectation.

Answer. (a)

$$\hat{a} = \frac{1/2}{1/2 + \mu} h$$

$$\hat{b} = \frac{\mu}{1/2 + \mu} h$$

(b)

$$\hat{\mu} = \frac{h - a + c}{6(h - a + c + d)}$$

Question 6

Assume each data point $X_i \in \mathbb{R}^+ (i = 1 \ldots n)$ is drawn from the following process:

$$Z_i \sim Multinomial(\pi_1, \pi_2, \ldots, \pi_K)$$

$$X_i \sim Gamma(2, \beta_{Z_i})$$

The probability density function of $Gamma(2, \beta)$ is $P(X = x) = \beta^2 x e^{-\beta x}$

(a) Assume $K = 3$ and $\beta_1 = 1, \beta_2 = 2, \beta_3 = 4$. Whats $P(Z = 1|X = 1)$?

(b) Describe the E-step: compute $P(Z = k|X = x)$ for each $X = x$. Write an equation for each value being computed.

Hint. $P(Z = 1|X = 1) = \frac{P(X=1|Z=1)P(Z=1)}{\sum_{k=1} P(X=1|Z=k)P(Z=k)}$

Answer. (a)

$$P(Z=1|X=1) \propto P(X=1|Z=1)P(Z=1) = \pi_1 e^{-1}$$

$$P(Z=2|X=1) \propto P(X=1|Z=2)P(Z=2) = \pi_2 4e^{-2}$$

$$P(Z=3|X=1) \propto P(X=1|Z=3)P(Z=3) = \pi_3 16e^{-4}$$

$$P(Z=1|X=1) = \frac{\pi_1 e^{-1}}{(\pi_1 e^{-1} + \pi_2 4e^{-2} + \pi_3 16e^{-4})}$$

(b) For each $X = x$,

$$P(Z=k|X=x) = \frac{P(X=x|Z=k)P(Z=k)}{\sum_{k'} P(X=x|Z=k')P(Z=k')} = \frac{\beta_k^2 x e^{-\beta_k x} \pi_k}{\sum_{k'} \beta_{k'}^2 x e^{-\beta_{k'} x} \pi_{k'}}$$