Introduction to Machine Learning, Spring 2025
Homework 7

(Due June 8, 2025 at 11:59pm (CST))

May 19, 2025

1. Please write your solutions in English.
2. Submit your solutions to the course Gradescope.
3. If you want to submit a handwritten version, scan it clearly.
4. Late homeworks submitted within 3 days of the due date will be marked down 25% each day cumulatively Homeworks submitted more than 3 days after the due date will not be accepted unless there is a valid reason, such as a medical or family emergency.
5. You are required to follow ShanghaiTech's academic honesty policies. You are allowed to discuss problems with other students, but you must write up your solutions by yourselves. You are not allowed to copy materials from other students or from online or published resources. Violating academic honesty can result in serious penalties.

1. [15 points] [Expectation Maximization Algorithm]

Consider a probabilistic model in which we collectively denote the observed variables by X and all of the hidden variables by Z. The joint distribution $p(X, Z|\theta)$ is parameterized by θ. Our goal is to maximize the likelihood function given by

$$p(X|\theta)$$

(a) Given an arbitrary distribution q, show that the log-likelihood of X is [5 points]

$$\log p(X|\theta) = \mathbb{E}_{Z\sim q}\left[\log \frac{p(X, Z|\theta)}{q(Z)}\right] + KL\left(q(Z)\|p(Z|X, \theta)\right)$$

(b) Next let's consider the expectation step. First show the evidence lower bound (ELBO) is a lower bound of the log-likelihood. [5 points]

$$\log p(X|\theta) \geq \mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log \frac{p(X, Z|\theta)}{p(Z|X, \theta^{(t-1)})}\right]$$

where $\theta^{(t-1)}$ is the parameter estimated in the previous iteration.

(c) We want to maximize the ELBO, $\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log \frac{p(X, Z|\theta)}{p(Z|X, \theta^{(t-1)})}\right]$ since maximizing $p(X|\theta)$ is hard. EM algorithm defines $Q(\theta|\theta^{(t-1)}) = \mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log p(X, Z|\theta)\right]$. The M-step is given by:

$$\theta^{(t)} \leftarrow \operatorname*{argmax}_{\theta} Q(\theta|\theta^{(t-1)})$$

Show that maximizing $Q(\theta|\theta^{(t-1)})$ and maximizing the ELBO is equivalent. [5 points] Formally,

$$\operatorname*{argmax}_{\theta} Q(\theta|\theta^{(t-1)}) = \operatorname*{argmax}_{\theta} \mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log \frac{p(X, Z|\theta)}{p(Z|X, \theta^{(t-1)})}\right]$$

Solution

(a) With Bayes' Rule, we can get that

$$p(X|\theta) = \frac{p(X, Z|\theta)}{p(Z|X, \theta)} = \frac{p(X, Z|\theta)}{q(Z)}\frac{q(Z)}{p(Z|X, \theta)}$$

So the log-likelihood of X is

$$\log p(X|\theta) = \log\left[\frac{p(X, Z|\theta)}{q(Z)}\frac{q(Z)}{p(Z|X, \theta)}\right] = \log \frac{p(X, Z|\theta)}{q(Z)} + \log \frac{q(Z)}{p(Z|X, \theta)}$$

Take the expectation of Z with respect to $q(Z)$ to the both side, we can get that

$$\mathbb{E}_{Z\sim q}\left[\log p(X|\theta)\right] = \mathbb{E}_{Z\sim q}\left[\log \frac{p(X, Z|\theta)}{q(Z)} + \log \frac{q(Z)}{p(Z|X, \theta)}\right]$$

With the linearty of expectation:

$$\mathbb{E}_{Z\sim q}\left[\log p(X|\theta)\right] = \mathbb{E}_{Z\sim q}\left[\log \frac{p(X, Z|\theta)}{q(Z)}\right] + \mathbb{E}_{Z\sim q}\left[\log \frac{q(Z)}{p(Z|X, \theta)}\right]$$

For $\mathbb{E}_{Z\sim q}\left[\log p(X|\theta)\right]$, we can get that it has nothing with Z, so

$$\mathbb{E}_{Z\sim q}\left[\log p(X|\theta)\right] = \int q(z)\log p(X|\theta)\mathrm{d}z = \log p(X|\theta)\int q(z)\mathrm{d}z = \log p(X|\theta)$$

For $\mathbb{E}_{Z\sim q}\left[\log \frac{q(Z)}{p(Z|X, \theta)}\right]$, accoding to the definition of KL divergence:

$$KL(p\|q) = \int p(z)\log \frac{p(z)}{q(z)}\mathrm{d}z$$

we can get that:

$$\mathbb{E}_{Z\sim q}\left[\log\frac{q(z)}{p(z|X,\theta)}\right]=\int q(z)\log\frac{q(z)}{p(z|X,\theta)}\mathrm{d}z=KL(q(Z)\|p(Z|X,\theta))$$

So above all, we have proved that

$$\log p(X|\theta)=\mathbb{E}_{Z\sim q}\left[\log\frac{p(X,Z|\theta)}{q(Z)}\right]+KL(q(Z)\|p(Z|X,\theta))$$

(b) For the log-likelihood:

$$\begin{aligned}\log p(X|\theta)&=\log\int p(X,z|\theta)\mathrm{d}z\\&=\log\int p(z|X,\theta^{(t-1)})\frac{p(X,z|\theta)}{p(z|X,\theta^{(t-1)})}\mathrm{d}z\\&=\log\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t-1)})}\right]\end{aligned}$$

Since log is a concave function, with Jensen's inequality, we have $\log\mathbb{E}(X)\geq\mathbb{E}(\log X)$, so

$$\log\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t-1)})}\right]\geq\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t-1)})}\right]$$

So above all, we have proved that the ELBO is that

$$\log p(X|\theta)\geq\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t-1)})}\right]$$

(c)

$$\begin{aligned}&\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t-1)})}\right]\\&=\int p(z|X,\theta^{(t-1)})\log\frac{p(X,z|\theta)}{p(z|X,\theta^{(t-1)})}\mathrm{d}z\\&=\int p(z|X,\theta^{(t-1)})(\log p(X,z|\theta)-\log p(z|X,\theta^{(t-1)}))\mathrm{d}z\\&=\int p(z|X,\theta^{(t-1)})\log p(X,z|\theta)\mathrm{d}z-\int p(z|X,\theta^{(t-1)})\log p(z|X,\theta^{(t-1)})\mathrm{d}z\end{aligned}$$

Since $Q(\theta|\theta^{(t-1)})=\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log p(X,Z|\theta)\right]$. So we have

$$\int p(z|X,\theta^{(t-1)})\log p(X,z|\theta)\mathrm{d}z=\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log p(X,Z|\theta)\right]=Q(\theta|\theta^{(t-1)})$$

And with the definition of entropy: $H(X)=-\int p(x)\log p(x)\mathrm{d}x$, we can get that

$$-\int p(z|X,\theta^{(t-1)})\log p(z|X,\theta^{(t-1)})\mathrm{d}z=H(Z|X,\theta^{(t-1)})$$

So

$$\begin{aligned}&\int p(z|X,\theta^{(t-1)})\log p(X,z|\theta)\mathrm{d}z-\int p(z|X,\theta^{(t-1)})\log p(z|X,\theta^{(t-1)})\mathrm{d}z\\&=Q(\theta|\theta^{(t-1)})+H(Z|X,\theta^{(t-1)})\end{aligned}$$

Since $H(Z|X,\theta^{(t-1)})$ is a constant of θ, so we can get that

$$\operatorname*{argmax}_\theta\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t-1)})}\right]=\operatorname*{argmax}_\theta Q(\theta|\theta^{(t-1)})+H(Z|X,\theta^{(t-1)})=\operatorname*{argmax}_\theta Q(\theta|\theta^{(t-1)})$$

So above all, we have proved that

$$\operatorname*{argmax}_\theta Q(\theta|\theta^{(t-1)})=\operatorname*{argmax}_\theta\mathbb{E}_{Z|X,\theta^{(t-1)}}\left[\log\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t-1)})}\right]$$

2. [15 points] [Performing K-Means by Hand]

Let's do k-means! To initialize the centroids, we use the k-means++ algorithm. And then use Euclidean distance to cluster the 8 data points into $k = 3$ clusters. The coordinates of the data points are:

$$x^{(1)} = (2, 8),\ x^{(2)} = (2, 5),\ x^{(3)} = (1, 2),\ x^{(4)} = (5, 8),$$
$$x^{(5)} = (7, 3),\ x^{(6)} = (6, 4),\ x^{(7)} = (8, 4),\ x^{(8)} = (4, 7).$$

Suppose that initially the first cluster centers is $x^{(1)}$.
To ensure consistent results, please use random numbers in the order shown in the table below. When selecting a center, arrange it in ascending order of sequence number. For example, when the normalized weights of 5 nodes are 0.2, 0.1, 0.3, 0.3, and 0.1, if the random number is 0.3, the selected node is the third one. Note that you don't necessarily need to use all of them.

0.6	0.2	0.5	0.9	0.3

(a) Perform the k-means++ algorithm to initialize other centers and report the coordinates of the resulting centroids. [5 points]

(b) Calculate the loss function

$$Q(r, c) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} r_{ij} \left\| x^{(i)} - c_j \right\|_2^2$$

where $r_{ij} = 1$ if $x^{(i)}$ belongs to the j-th cluster and 0 otherwise. [5 points]

(c) How many more iterations after initialization are needed to converge? [3 points] Calculate the loss after it converged. [2 points]

Solution

(a) We can calculate the other points' Euclidean distance to $x^{(1)}$ is $D\left(x^{(i)}\right)$, and the probability of selecting $x^{(i)}$ as the next center is $p\left(x^{(i)}\right)$, which is proportional to $D\left(x^{(i)}\right)^2$.
So the $D^2\left(x^{(i)}\right)$ and $p\left(x^{(i)}\right)$ are shown in the table below.

point	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$
$D^2\left(x^{(i)}\right)$	0	9	37	9	50	32	52	5
$p\left(x^{(i)}\right)$	0	0.05	0.19	0.05	0.26	0.16	0.27	0.03

We randomly sample a point. The random number is 0.6, and since $\sum_{i=1}^{5} p\left(x^{(i)}\right) = 0.55 < 0.6$, $\sum_{i=1}^{6} p\left(x^{(i)}\right) = 0.71 > 0.6$, so we choose $x^{(6)}$ as the second class center.

2. Then, we need to choose the third center.
Suppose that for the i-th point $x^{(i)}$, the Euclidean distance for it to $x^{(1)}$ is $D_1\left(x^{(i)}\right)$, the Euclidean distance for it to $x^{(6)}$ is $D_2\left(x^{(i)}\right)$.
So the Euclidean distance to the closest center $D\left(x^{(i)}\right) = \min\left(D_1\left(x^{(i)}\right), D_2\left(x^{(i)}\right)\right)$.
So the $D_1^2\left(x^{(i)}\right), D_2^2\left(x^{(i)}\right), D^2\left(x^{(i)}\right)$ and $p\left(x^{(i)}\right)$ are shown in the table below.

point	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$
$D_1^2\left(x^{(i)}\right)$	0	9	37	9	50	32	52	5
$D_2^2\left(x^{(i)}\right)$	32	17	29	17	2	0	4	13
$D^2\left(x^{(i)}\right)$	0	9	29	9	2	0	4	5
$p\left(x^{(i)}\right)$	0	0.16	0.50	0.16	0.03	0	0.07	0.09

We randomly sample a point. The random number is 0.2, and since $\sum_{i=1}^{2} p\left(x^{(i)}\right) = 0.16 < 0.2$, $\sum_{i=1}^{3} p\left(x^{(i)}\right) = 0.76 > 0.2$, so we choose $x^{(3)}$ as the third class center.

So above all, the initialized centers are:

$$c_1 = x^{(1)} = (2, 8), c_2 = x^{(3)} = (1, 2), c_3 = x^{(6)} = (6, 4)$$

(b) The centers after initialization are:

$$c_1 = x^{(1)} = (2, 8), c_2 = x^{(3)} = (1, 2), c_3 = x^{(6)} = (6, 4)$$

And $x^{(1)}, x^{(2)}, x^{(4)}, x^{(8)}$ belong to c_1, $x^{(3)}$ belong to c_2, $x^{(5)}, x^{(6)}, x^{(7)}$ belong to c_3.
So the loss is

$$Q(r, c) = \frac{1}{8} \sum_{i=1}^{8} \sum_{j=1}^{3} r_{ij} \left\| x^{(i)} - c_j \right\|^2 = \frac{(0 + 9 + 9 + 5) + (0) + (2 + 0 + 4)}{8} = \frac{29}{8}$$

So above all, the loss is $Q(r, c) = \frac{29}{8}$.
(c) For the 1-st iteration, we have:

$$\begin{aligned} c_1 &= \frac{1}{4}(x^{(1)} + x^{(2)} + x^{(4)} + x^{(8)}) = \left(\frac{13}{4}, 7\right) \\ c_2 &= x^{(3)} = (1, 2) \\ c_3 &= \frac{1}{3}(x^{(5)} + x^{(6)} + x^{(7)}) = \left(7, \frac{11}{3}\right) \end{aligned}$$

Then we calculate the Euclidean distance for each point to each center:

$$D_1^2\left(x^{(i)}\right) = \left\| x^{(i)} - c_1 \right\|^2, D_2^2\left(x^{(i)}\right) = \left\| x^{(i)} - c_2 \right\|^2, D_3^2\left(x^{(i)}\right) = \left\| x^{(i)} - c_3 \right\|^2$$

The row c_j means that the of the corresponding point is to the center c_j.
i.e. for the point $x^{(i)}$, the distance to the center c_j has the smallest Euclidean distance among all centers.

point	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$
center c_j	c_1	c_1	c_2	c_1	c_3	c_3	c_3	c_1

And we can see that the center c_j for each point is the same as the previous iteration.
So it converged. i.e. it only needs 1 iteration to converge.
And we can calculate the Euclidean distance for each point to their center:

$$\begin{aligned} D^2\left(x^{(1)}\right) &= \frac{41}{16}, D^2\left(x^{(2)}\right) = \frac{89}{16}, D^2\left(x^{(3)}\right) = 0 \\ D^2\left(x^{(4)}\right) &= \frac{65}{16}, D^2\left(x^{(5)}\right) = \frac{4}{9}, D^2\left(x^{(6)}\right) = \frac{10}{9} \\ D^2\left(x^{(7)}\right) &= \frac{10}{9}, D^2\left(x^{(8)}\right) = \frac{9}{16} \end{aligned}$$

So the loss after it converged is:

$$Q(r, c) = \frac{1}{8} \sum_{i=1}^{8} \sum_{j=1}^{3} r_{ij} \left\| x^{(i)} - c_j \right\|^2 = \frac{\left(\frac{41}{16} + \frac{89}{16} + \frac{65}{16} + \frac{9}{16}\right) + 0 + \left(\frac{4}{9} + \frac{10}{9} + \frac{10}{9}\right)}{8} = \frac{185}{96}$$

So above all, 1 iteration is needed to converge, and the loss after it converged is $Q(r, c) = \frac{185}{96}$.