Gyujin Park

CSE 6740 HOMEWORK #2

1. Questions

Question 1. 1.
(a)

$$
\begin{aligned}
&\sum_{z\in Z} p(z)p(x|z) \\
&= \sum_{z\in Z} \left(\Pi_{k=1}^{K} \pi_k^{z_k}\right) \left(N(x|\mu_k, \Sigma_k)^{z_k}\right) \\
&= \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)
\end{aligned}
$$

(b)

$$
\begin{aligned}
z_k^i &= p(z_k^i = 1|x_i) \\
&= \frac{p(z_k^i = 1)p(x_i|z_k^i = 1}{\sum_{j=1}^{K} p(z_j^i = 1)p(x_i|z_j^i = 1)} \\
&= \frac{\pi_k N(x_i|\mu_k \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j, \Sigma_j)}
\end{aligned}
$$

(c)

$$
L = \Pi_{i=1}^{N} p(x_i|\pi, \mu,)
$$

$$
\begin{aligned}
l &= logL \\
&= \sum_{i=1}^{N} \log p(x_i|\pi, \mu, \Sigma) \\
&= \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)
\end{aligned}
$$

$$
\frac{dl}{d\mu_k} = \sum_{i=1}^{N} z_k^i \Sigma_k^{-1}(x_i - \mu_k) = 0
$$

$$\rightarrow \sum_{i=1}^{N} z_k^i (x_i - \mu_k) = 0$$
$$\sum_{i=1}^{N} z_k^i x_i = \mu_k \sum_{i=1}^{N} z_k^i$$
$$\therefore \mu_k = \frac{1}{N_k} \sum_{i=1}^{N} z_k^i x_i (N_k = \sum_{i=1}^{N} z_k^i$$

Similarly, by setting the derivative of l with respect to Σ_k to be zero, we can get

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} z_k^i (x_i - \mu_k)(x_i - \mu_k)^T$$

Finally, by using a Lagrange multiplier and maximizing $l + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$, we can get

$$\pi_k = \frac{N_k}{N}$$

(e)
E-step

$$z^i = argmax_k (x^i - \mu_k)_k^{-1} (x^i - \mu_k)$$

M-step

$$\mu_k = \frac{\sum_i \delta(z^i, k) x^i}{\sum_i \delta(z^i, k)},$$
$$\Sigma_k = \frac{\sum_i \delta(z^i, k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i \delta(z^i, k)},$$

where

$$\delta(z^i, k) = 1 \ if \ z^i = k$$
$$\delta(z^i, k) = 0 \ if \ z^i \neq k$$

Question 2. 2.
(a)

$$L = \Pi p(x^i|\theta)$$

$$\begin{aligned} l &= \log L \\ &= \sum \log p(x^i|\theta) \\ &= \sum \log \left(\theta^{x^i}(1-\theta)^{1-x^i}\right) \\ &= \sum x^i \log\theta + (1-x^i)\log(1-\theta) \end{aligned}$$

$$\frac{dl}{d\theta} = \sum \frac{x^i}{\theta} - \frac{1-x^i}{1-\theta} = 0$$

$$\frac{1}{\theta}\sum x^i = \frac{1}{1-\theta}\sum(1-x^i)$$

$$\therefore \theta = \frac{1}{m}\sum x^i$$

(b)

$$L = \Pi p(x^i|\mu,\sigma)$$

$$\begin{aligned} l &= \log L \\ &= \sum \log p(x^i|\mu,) \\ &= \sum \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x^i-\mu)^2} \end{aligned}$$

$$l = \sum -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2\sigma^2}(x^i-\mu)^2$$

$$\frac{dl}{d\mu} = \sum \frac{1}{2\sigma^2}2(x^i-\mu) = 0$$

$$\to \sum(x^i-\mu) = 0$$

$$\therefore \mu = \frac{1}{m}\sum x^i$$

$$\frac{dl}{d\sigma} = \sum -\frac{1}{\sigma} + \sigma^{-3}(x^i-\mu)^2 = 0$$

$$\to \frac{m}{\sigma} = \sigma^{-3}\sum(x^i-\mu)^2$$

$$\therefore \sigma^2 = \frac{1}{m}\sum(x^i-\mu)^2$$

(c)

$$p(x) = \sum_{j=1}^{n} \frac{nc_j}{m} I(x \in B_j)$$

Conditions: $p(x) \geq 0$, $\int_0^1 p(x)dx = 1$ Proof.

$$\begin{aligned}
&\int_0^1 p(x)dx \\
&= \int_0^1 \sum \frac{nc_j}{m} I(x \in B_j)dx \\
&= \sum \int_0^1 \frac{nc_j}{m} I(x \in B_j)dx \\
&= \sum \int_{\frac{j-1}{n}}^{\frac{j}{n}} \frac{nc_j}{m} dx \\
&= \sum \frac{c_j}{m} \\
&= 1
\end{aligned}$$

(d)-1
Parametric models: Models which can be described by a fixed number of parameters
e.g. Bernoulli, Gaussian
Non-parametric models: Models where parameters are flexible and can vary
e.g. Histogram, Kernel density estimator

(d)-2

$$\begin{aligned}
&K(u) \geq 0 \\
&\int K(u)du - 1 \\
&\int uK(u)du - 0 \\
&\int u^2 K(u)du < \infty
\end{aligned}$$

(d)-3
1) There are too many bins so that memory requirement is a lot.
2) It is statistically not the best

(d)-4

$$
\begin{aligned}
&1.K(u) = \frac{1}{\sqrt{2\pi}}e^{-\dfrac{u^2}{2}} \\
&2.K(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1) \\
&3.K(u) = \frac{\pi}{4}cos(\frac{\pi}{2}u)I(|u| \leq 1) \\
&4.K(u) = \frac{5}{8}(1-u^4)I(|u| \leq 1)
\end{aligned}
$$

(d)-5
Nonparametric models place very mild assumptions on the data distribution and provide good models for complex data.
Parametric models rely on very strong(simplistic) distributional assumptions.

Nonparametric models(not histrograms) requires storing and computing with the entire data set.
Parametric models, once fitted, are much more efficient in terms of storage and computation.

Question 3. (a)

$$\begin{aligned} H(Z|X) &= -\sum_{x,z} p_{X,Z}(x,z)\log p_{X,Z}(z|x) \\ &= -\sum_{x,z} p_{X,Y}(x,z-x)\log p_{X,Y}(z-x|x) \\ &= -\sum_{x,y} p_{X,Y}(x,y)\log p_{X,Y}(y|x) \\ &= H(Y|X) \end{aligned}$$

If X and Y are independent,

$$H(Z) \geq H(Z|X) = H(Y|X) = H(Y)$$
$$\therefore H(Z) \geq H(Y)$$
$$H(Z) \geq H(Z|Y) = H(X|Y) = H(X)$$
$$\therefore H(Z) \geq H(X)$$

(b)
Let X be a random variable such that $X = 0$ for $1/2$ probability and $X = 1$ for $1/2$ probability.
Then let Y be a random variable such that $Y = 1 - X$. As a consequence, $Z = 1$.

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$
$$H(Y) = 1$$
$$H(Z) = -1\log 1 = 0$$
$$\therefore H(X) > H(Z) \; and \; H(Y) > H(Z),$$

which constitutes a desirable example.

Question 4.

mycluster.py

mycluster.py > ...

```python
    n_d = T.shape[0]
    n_w = T.shape[1]
    n_c = K

    # initialize
    pi = np.random.rand(n_c)
    pi /= np.sum(pi)
    mu = np.random.rand(n_w, n_c)

    for _ in range(num_iters):
        # E step
        gamma = np.zeros((n_d, n_c))

        for i in range(n_d):
            d = 0
            for c in range(n_c):
                d += pi[c] * np.product(np.power(mu[:, c], T[i, :]))

            for c in range(n_c):
                gamma[i][c] = pi[c] * np.product(np.power(mu[:, c], T[i, :])) / d

        # M step
        X = np.matmul(gamma.T, T)

        for c in range(n_c):
            for j in range(n_w):
                mu[j][c] = X[c][j] / np.sum(X[c, :])

        pi = np.sum(gamma, axis = 0)

    idx = np.argmax(gamma, axis=1)
    idx += 1
    # raise NotImplementedError
    return idx
```

```
/mnt/c/U/gyujin/De/6740hw  python homework2.py
accuracy 0.8625
/mnt/c/U/gyujin/De/6740hw                    ✔  32s ⧗  10:58:42 ⊙
```