CSE/ISYE 6740 Homework 2 (Le Song)

deadline: 10/17

1 EM for Mixture of Gaussians

(a) Answer:

Since

$$p(z^i) = \prod_{k=1}^{K} \pi_i^{z_k^i} = \pi_1^{z^1} ... \pi_i^{z^i} ... \pi_k^{z^k} = \pi_1^0 ... \pi_i^1 ... \pi_k^0 = \pi_i$$

$$p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{Zk} = N(x|\mu_1, \Sigma_1)^0 ... N(x|\mu_i, \Sigma_i)^1 ... N(x|\mu_K, \Sigma_K)^0 = N(x|\mu_i, \Sigma_i)$$

Then

$$p(x) = \prod_{z \in Z} p(z)p(x|z) = p(z^1)p(x|z^1) + ... + p(z^1)p(x|z^i) + ... + p(z^K)p(x|z^K)$$

$$= \pi_1 N(x|\mu_1, \Sigma_1) + ... + \pi_i N(x|\mu_i, \Sigma_i) + ... + \pi_1 N(x|\mu_K, \Sigma_K) = \prod_{k=1}^{K} \pi_k N(X|\mu_K, \Sigma_K)$$

(b) Answer:

According to Bayes' rule

$$P(z|x) = \frac{P(x|z)p(z)}{P(x)},$$

$$p(z_k^n|x_n) = \frac{p(x_n|z_n^k)p(z_k^n)}{p(x_n)} = \frac{\pi_k N(x|\mu_K, \Sigma_K)}{\Sigma_i \pi_i N(X|\mu_K, \Sigma_K)}$$

(c) Answer:

The objective function

$$f(\theta) = E_{q(z^1, z^2, ..., z^n|x^i)}[log \prod_{i=1}^{n} p(x^i, z^i|\theta)]$$

$$= E_{q(z^1, z^2, ..., z^n|x^i)}[log \prod_{i=1}^{n} \pi_{z^i} N(x|\mu_i, \Sigma_i)$$

$$= E_{q(z^1, z^2, ..., z^n|x^i)}[log \pi_{z^i} - (x^i - \mu_k)^T \Sigma_k^{-1}(x^i - \mu_k) - \frac{1}{2} log|\Sigma_{z^i}| + c]$$

$$= \sum_{i=1}^{N} E_{p(z^i|x^i)}[log \pi_k - (x^i - \mu_k)^T \Sigma_k^{-1}(x^i - \mu_k) - \frac{1}{2} log|\Sigma_{z^i}| + c],$$

Since

$$\tau_k^i = p(z_k^i = k|x^i),$$

$$f(\theta) = \sum_{i=1}^{N} \sum_{K=1}^{K} \tau_k^i [log \pi_k - (x^i - \mu_k)^T \Sigma_k^{-1}(x^i - \mu_k) - \frac{1}{2} log|\Sigma_{z^i}| + c]$$

Form Lagrangian,

$$L = \sum_{i=1}^{N} \sum_{K=1}^{K} \tau_k^i [log \pi_k - (x^i - \mu_k)^T \Sigma_k^{-1}(x^i - \mu_k) - \frac{1}{2} log|\Sigma_{z^i}| + c] + \lambda(1 - \sum_{K=1}^{K} \pi_k)$$

Let partial derivative

$$\frac{\partial L}{\partial \Sigma_k^{-1}} = \sum_{i=1}^{N} \tau_k^i \Sigma_k^T - \sum_{i=1}^{N} \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^T = 0$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^T}{\sum_{i=1}^{N} \tau_k^i}$$

(d) Answer:
Since all the mixture components are given by covariance $\Sigma_k = \epsilon I$, so the Gaussian mixture is:

$$N(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{\frac{1}{2}}} exp - \frac{1}{2\epsilon}||x - \mu_k||^2$$

Using Bayes' theorem, the posterior probability is:

$$\gamma(z_k^n) = \frac{\pi_k N(x^n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_j N(x^n|\mu_j, \Sigma_j)} = \frac{\pi_k exp[-||x^n - \mu_k||^2/2\epsilon]}{\sum_{j=1}^{K} \pi_j exp[-||x^n - \mu_j||^2/2\epsilon]}$$

As $\epsilon \to 0$, if $k = j, \gamma(z_k^n) = 1$; if $k \neq j, \gamma(z_k^n) = 0$.

Thus, the log likelihood function is:

$$l = \sum_{i=1}^{N} \sum_{K=1}^{K} \tau_k^i [log\pi_k - \frac{1}{2\epsilon}(x^i - \mu_k)^T (x^i - \mu_k) - \frac{n}{2} log\epsilon + c]$$

$$= -\frac{1}{2\epsilon} \sum_{i=1}^{N} \sum_{K=1}^{K} \gamma_k^i (x^i - \mu_k)^T (x^i - \mu_k) + c \ (\gamma_k^i = 0 \ or \ 1)$$

As a result, in the limit of $\epsilon \to 0$, maximizing the log-likelihood function for this model is equivalent to minimizing objective function $J = \sum_{i=1}^{N} \sum_{K=1}^{K} \gamma_n k ||x_n - \mu_k||^2$ in K-means. (Reference: Bishop-PRML 9.3.2)

2 Density Estimation

(a) What is the log-likelihood function?
Answer:
Suppose that there are M regions,

$$p(x) = \prod_{n=1}^{N} p(x^n) = \prod_{i=1}^{M} h_i^{n_i}$$

the log-likelihood function is:

$$log \ p(x) = log \prod_{i=1}^{M} h_i^{n_i} = \sum_{i=1}^{M} n_i log \ h_i$$

(b) Derive an expression for the maximum likelihood estimator for h_i.
Answer:
The constraint is

$$\sum_i h_i \Delta_i = 1$$

so Lagrange multiplier can be written as:

$$L = \sum_{i=1}^{M} n_i log h_i + \lambda(1 - \sum_i h_i \Delta_i)$$

Take partial derivative

$$\frac{\partial L}{\partial h_i} = \frac{n_i}{h_i} - \lambda \Delta_i = 0 \ \Rightarrow \ h_i = \frac{n_i}{\lambda \Delta_i}$$

$$\Rightarrow L = \sum_{i=1}^{M} n_i (log n_i - log\lambda - log\Delta_i) + \lambda(1 - \sum_i \frac{n_i}{\lambda})$$

$$\frac{\partial L}{\partial \lambda} = -\sum_i \frac{n_i}{\lambda} + 1 = 0,$$

$$\frac{N}{\lambda} = 1, \;\; \lambda = N$$

Thus the estimator for h_i is:

$$\hat{h}_i = \frac{n_i}{N\Delta_i}$$

(c) Mark T if it is always true, and F otherwise. Briefly explain why.

• Non-parametric density estimation usually does not have parameters.

Answer: F. "Non-paramatric" doesn't mean that there are no parameters; otherwise, it means that there are no fixed number of parameters.

• The Epanechnikov kernel is the optimal kernel function for all data.

Answer: F. The Epanechnikov kernel is optimal in a mean square error sense.So it's not optimal for all data.

• Histogram is an efficient way to estimate density for high-dimensional data.

Answer: F. If the total number of bins in histogram is larger than the number of sample, then most of the bins will be empty, which leads to inefficiency.

• Parametric density estimation assumes the shaple of probability density.

Answer: T. The parametric density estimation assumes the shpe of probability density by choosing diferrent parameters.

3 Information Theory

(a) Prove that $H(X,Y) \leq H(X) + H(Y)$

$$\begin{aligned}
H(X,Y) = -\sum_{x\in X}\sum y \in Y p(x,y) logp(x,y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y) log[p(x)p(y|x)] \\
= -\sum_{x\in X}\sum y \in Y p(x,y) logp(x) &- \sum_{x\in X}\sum_{y\in Y} p(x,y) logp(y|x) \\
= -\sum_{x\in X} logp(x) \sum_{y\in Y} p(x,y) &- \sum_{x\in X}\sum_{y\in Y} p(x,y) logp(y|x) \\
= -\sum_{x\in X} logp(x)p(x) &- \sum_{x\in X}\sum_{y\in Y} p(x,y) logp(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

Mutual information

$$\begin{aligned}
I(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(y|x)}{p(y)} \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y) log\ p(y|x) - \sum_{y\in Y} logp(y) \sum_{x\in X} p(x,y) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y) log\ p(y|x) - \sum_{y\in Y} logp(y)p(y) \\
&= -H(Y|X) + H(Y)
\end{aligned}$$

To combine these two equations,

$$H(X,Y) = H(X) + H(Y|X) = H(X) + H(Y) - I(X,Y)$$

$$\Rightarrow I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Because of non-negativity of mutual information $I(X;Y)$,

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0,$$

$$H(X,Y) \leq H(X) + H(Y)$$

(b) Show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

Answer: according to the solution of (a), we get that:

$$H(X,Y) = H(X) + H(Y|X)$$

$$I(X;Y) = H(Y) - H(Y|X)$$

So $I(X;Y) = H(Y) - H(X,Y) - H(X)) = H(X) + H(Y) - H(X,Y)$

(c) Under what conditions does $H(Z) = H(X) + H(Y)$
Answer:
Since $Z = X + Y,\ p(Z = z|X = x) = p(Y = z - x|X = x)$

$$H(Z|X) = -\sum_{x \in X} p(X = x) \sum_{z \in Z} p(Z|X = x) log P(Z|X = x)$$

$$= -\sum_{x \in X} p(X = x) \sum_{z \in Z} p(Y = z - x|X = x) log P(Y = z - x|X = x) = H(Y|X)$$

Similarly, we can get

$$H(Z|Y) = H(X|Y)$$

so $H(Z) = H(Z|X) + H(Z|Y) = H(Y|X) + H(X|Y)$
If X and Y is independent, $H(Y|X) + H(X|Y) = H(Y) + H(X)$,
then $H(Z) = H(X) + H(Y)$

4 Programming: Text Clustering

Answer:
I initialized $\mu_j c$ randomly by function $rand(n_w, 4)$ and I normalized it.
For π_c, I initialized it evenly by setting $\pi_c = [0.25, 0.25, 0.25, 0.25$.
My implementation stops at iteration = 1000.
Results: I run this algorithm for 10, 20 and 30 times. The average accuracy, maximum accuracy and minimum accuracy are summarized in the table below.

run	avg_acc	max_acc	min_acc
10	82.1250	88.7500	78
20	77.8500	86.2500	62.2500
30	75.8000	89.7500	51.2500