CSE 6740B HW2

Che-Ting,Meng

September 29th, 2022

1 EM for Mixture of Gaussian

1.1 (a)

For $p(z) = \prod \pi_k^{z_k}$, $p(x|z) = \prod N(x|\mu_k, \Sigma_k)^{z_k}$, z(i)=[0,0,....1($i_{th}$ element),...0],

p(z_i)=$\pi_k^{1^0} * \pi_k^{2^0} * ... * \pi_k^{i\,1}$($i_{th}$ element)$*... * \pi_k^{k^0} = \pi_k^i$,

$p(x|z_i) = N(x|\mu_1, \Sigma_1)^0 * ... * N(x|\mu_i, \Sigma_i)^1$($i_{th}$ element)$*... * N(x|\mu_k, \Sigma_k)^0 = N(x|\mu_i, \Sigma_i)$,

$$\begin{aligned}\therefore p(x) &= \Sigma p(z)p(x|z) \\ &= \prod \pi_k^{z_1} * \prod N(x|\mu_k, \Sigma_k)^{z_1} + \prod {\pi_k}^{z_2} * \prod N(x|\mu_k, \Sigma_k)^{z_2} + ... + \prod \pi_k^{z_k} * \prod N(x|\mu_k, \Sigma_k)^{z_k} \\ &= \pi_k^1 * N(x|\mu_1, \Sigma_1) + \pi_k^2 * N(x|\mu_2, \Sigma_2) + ... + \pi_k^k * N(x|\mu_k, \Sigma_k) \\ &= \Sigma \pi_k N(x|\mu_k, \Sigma_k)\end{aligned}$$

1.2 (b)

$\tau_k = p(z_k^i = 1|x_i) = \frac{p(x_i|z_k^i=1)*p(z_k^i=1)}{p(x_i)} = \frac{N(x_i|\mu_i,\Sigma_i)*\pi_k^i}{\Sigma \pi_k N(x_i|\mu_i,\Sigma_i)}$

1.3 (c)

The maximum likelihood function:

$$\begin{aligned}
f(\theta) &= \log \prod E_{p(z_i|x_i,\theta)} p(z_i, x_i|\theta) \\
&= \Sigma E_{p(z_i|x_i,\theta)} \log p(z_i, x_i|\theta) \\
&= \Sigma E_{p(z_i|x_i,\theta)} \log \pi_k^i N(x_i|\mu_i, \Sigma_i) \\
&= \Sigma E_{p(z_i|x_i,\theta)} [\log \pi_k^i - (x^i - \mu_i)^T \Sigma_{zi}^{-1} (x^i - \mu_i) - \frac{1}{2} \log |\Sigma_{zi}| + c] \\
&= \Sigma\Sigma \tau_k^i [\log \pi_k^i - (x^i - \mu_i)^T \Sigma_k^{-1} (x^i - \mu_i) - \frac{1}{2} \log |\Sigma_{zi}| + c]
\end{aligned}$$

Forming Lagrangian for constrain $\Sigma\pi_k = 1$:

$$L = \Sigma\Sigma \tau_k^i [\log \pi_k^i - (x^i - \mu_i)^T \Sigma_k^{-1} (x^i - \mu_i) - \tfrac{1}{2} \log |\Sigma_{zi}| + c] + \lambda(1 - \Sigma\pi_k)$$

Maximum at partial derivatives equals 0:

$$\begin{aligned}
\frac{\partial L}{\partial \pi_k} &= \Sigma \frac{\tau_k^i}{\pi_k} - \lambda = 0 \\
\pi_k &= \frac{1}{\lambda} \Sigma \tau_k^i
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L}{\partial \Sigma_k} &= -\frac{\Sigma \tau_k^i (x^i - \mu_i)(x^i - \mu_i)^T}{\Sigma_k^2} - \frac{\Sigma \tau_k^i}{\Sigma_k} = 0 \\
\Sigma_k &= \frac{\Sigma \tau_k^i (x^i - \mu_i)(x^i - \mu_i)^T}{\Sigma \tau_k^i}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L}{\partial \mu_k} &= 2\Sigma \tau_k (x_i - \mu_k) = 0 \\
\Sigma \tau_k * x_i &= \Sigma \tau_k \mu_k \\
\mu_k &= \frac{\Sigma x_i \tau_k}{\Sigma \tau_k}
\end{aligned}$$

1.4 (d)

With log likelihood as a concave function, we have Jensen's inequality in EM algorithm:

$$\log \Sigma_z q(z) \frac{p(x,z|\theta)}{q(z)} \geqslant \Sigma q(z) \log \frac{p(x,z|\theta)}{q(z)}$$

This assures that each of our updated expectations is under the true possibility distribution function. In order to further ensure non-decreasing in each iteration, we find that using posterior of z: $q(z) = p(z|x, \theta^t)$, makes the specific scenario where the inequality is equal, making our expectations maximizing over tight lower bounds. That is what assures every iteration is non-decreasing.

$$\begin{aligned}
E(f(x)) &= f(q(z), \theta) \\
&= f(p(z|x, \theta^t), \theta^t) \\
&= \Sigma p(z|x, \theta^t) \log \frac{p(x, z|\theta^t)}{p(z|x, \theta^t)} \\
&= \Sigma p(z|x, \theta^t) \log p(x|\theta^t) \\
&= \log p(x|\theta^t) \\
&= \log \Sigma p(x, z|\theta^t) \\
&= f(E(x))
\end{aligned}$$

1.5 (e)

E-Step: In the E-step of K-means, we use a hard assignment (only 1 or 0 for probability) of classes determined by the euclidean distances between points and each centroid instead of computing the posterior probability τ_k using a possibility distribution π_k.

M-Step: We subsequently calculate the MLE with respect to μ_k and Σ_k with partial derivative equals 0 and find the new parameter values to update.

2 Density Estimation

2.1 (a)

$$\begin{aligned} F(\theta) &= \log \sqcap \theta^x (1-\theta)^{1-x} \\ &= \Sigma[x \log \theta + \log(1-\theta) - x \log(1-\theta)] \\ &= n \log(1-\theta) + \Sigma[x \log \theta - x \log(1-\theta)] \end{aligned}$$

$$\begin{aligned} \frac{\partial F}{\partial \theta} = \frac{n}{\theta - 1} + \frac{\Sigma x}{\theta} - \frac{\Sigma x}{\theta - 1} &= 0 \\ n\theta + \Sigma x \theta - \Sigma x - \Sigma x \theta &= 0 \\ \theta &= \frac{\Sigma x}{n} \end{aligned}$$

2.2 (b)

$\mu = \frac{\Sigma x}{n}$

$\Sigma^2 = \frac{\Sigma (x_i - \mu)^2}{n}$

2.3 (c)

$p(x) = \Sigma \frac{nC_j}{m} I(x \in B_j)$

The valid condition is $\Sigma p(x) = 1$.

$$\begin{aligned}\int p(x)dx &= \int \Sigma \frac{nC_j}{m} I(x \in B_j)dx \\ &= \Sigma \int \frac{nC_j}{m} dx \\ &= \Sigma \frac{nC_j}{m} \int dx \\ &= \Sigma \frac{nC_j}{m} * \frac{1}{n} \\ &= \Sigma \frac{C_j}{m} \\ &= 1\end{aligned}$$

2.4 (d)

2.4.1 (d-1)

Parametric models are probability density functions that has a set number of parameters whereas non-parametric models have an unset number of parameters.

Parametric models: Gaussian distribution, Bernoulli distribution.
non-Parametric models: Histogram, Kernel density estimator.

2.4.2 (d-2)

1. Centered at 0;
2. Symmetric;
3. Has finite support;
4. Area under the curve equals 1.

2.4.3 (d-3)

When dealing with high dimensional data, histogram can have empty bins and has statistically higher inaccuracy.

KDE requires less memory, doesn't need any training (fit to data points directly) and is statistically better in terms of accuracy.

2.4.4 (d-4)

Gaussian, tophat, epanechnikov, cosine.

2.4.5 (d-5)

Firstly, they have different numbers of parameters, parametric models are fixed while non-parametric models are not.
Secondly, parametric models make strong assumptions while non-parametric models are arbitrary to the data points, making the statistical guarantee for parametric models dependant on whether the data type is a match with the assumptions while the statistical guarantee for non-parametric models dependant on the algorithm itself inherently.

3 Information Theory

3.1 (a)

Under $Z = X + Y$, the probabilities of Y and Z are equal when X is given.

$$\begin{aligned} H(Z|X) &= \Sigma\Sigma p(z,x)\log_2 p(z|x) \\ &= \Sigma p(x)\Sigma p(z|x)\log_2 p(z|x) \\ &= \Sigma p(x)\Sigma p(y|x)\log_2 p(y|x) \\ &= H(Y|X) \end{aligned}$$

If X,Y are independent, then $H(Y|X) = H(Y)$.
$\therefore H(Z) \geqslant H(Z|X) = H(Y|X) = H(Y)$. Similarly, $H(Z) \geqslant H(Y)$

3.2 (b)

In $Z = X + Y$, the condition sustains if $H(Y = -x_i|X = x_i) = 1$, so that $P(Z = 0) = 1, H(Z) = 0$. The logic is to make Z inevitably a constant to eliminate its uncertainty.
For example, rolling a regular dice and let X be the upper side, Y be the bottom side, then $Z = X+Y$ will inevitability result in $Z = 7$, having $H(X) = H(Y) > 0 > H(Z)$.

3.3 (c)

As proved above. if X and Y are independent, $H(Y|X) = H(Y)$.

$$H(Z) = H(f(X,Y)) \leqslant H(X,Y) = H(X) + H(Y|X) = H(X) + H(Y)$$

For H(f(X,Y))=H(X,Y), then the uncertainty of Z should equal to the uncertainty to the joint distribution of X and Y, meaning that every z_i should have a specific corresponding set of x_i, y_i.
Hence, if f(X,Y) is a bijection function, then H(f(X,Y))=H(X,Y). Consequently, H(Z)=H(X)+H(Y).

4 Programming Report: Text Clustering

I implemented initialization by summing up each column of the T matrix (total number of one specific word) and adding it with a normal distribution of word count with range between 0 and average word count in one document to create four different modified word count columns for each cluster. Then, I divide the modified text count matrix with the total sum of the itself to normalize it.
The results fluctuates around 93 to 96 accuracy. I have found that reducing the iteration numbers produces similar results, showing convergence before 1000 iterations. At least with this initialization method, 700 iterations shows confidence in convergence.