CSE 6740 - HW 2

Manvitha Kalicheti
gtID: 903838438

October 20, 2022

Collaborators

Keerthan Ramnath, Aishwarya Vijaykumar Sheelvant

References

Pattern Recognition and Machine Learning by Christopher Bishop, Class slides
For debugging and finding different methods: geeksforgeeks.org, stackoverflow.com, towardsdatascience.com

1

(a)

$$\begin{aligned}
p\left(z^{(i)}\right) &= \pi_1^{z_1} \ldots..\pi_k^{z_k} = \pi_1^0 \pi_i^{z_i} \ldots \pi_k^0 \\
&= \pi_i^{z_i} = \pi_i \\
p\left(x \mid z^i\right) &= \mathcal{N}\left(x \mid \mu_t^i, \Sigma_i\right) \\
p(x) &= \sum_{z \in Z} p(z) p(x \mid z) \longrightarrow (2) \\
&= \pi_1 \mathcal{N}\left(x \mid \mu_1, \Sigma_1\right) + \pi_2 \mathcal{N}\left(x / \mu_2, \Sigma_2\right) \\
&\quad + \pi_k \mathcal{N}\left(x \mid \mu_k, \Sigma_k\right) \\
&= \sum_{i=1}^{k} \pi_i \mathcal{N}\left(x \mid \mu_k, \Sigma_k\right) \longrightarrow (1)
\end{aligned}$$

$$\boxed{1 \equiv 2}$$

(b)

Using Bayes rule,

$$\begin{aligned}
P(z \mid x) &= \frac{P(x \mid z) P(z)}{P(x)} = \frac{P(x, z)}{\sum_{z'} P\left(x, z'\right)} \\
P\left(z_k^i \mid x_i\right) &= \frac{p\left(x_i \mid z_k^i\right) p\left(z_k^i\right)}{P\left(x_i\right)} \\
P\left(z_k^i \mid x_i\right) &= \frac{\mathcal{N}\left(x \mid \mu_i, \Sigma_i\right) \pi^i}{\sum_j \pi_j \mathcal{N}\left(x \mid \mu_j, \Sigma_j\right)}
\end{aligned}$$

(c)

Objective function:

$$\begin{aligned}
&= E_{q(z^1,z^2,\ldots z^n|x^i)}[\log \Pi_{i=1}^n p(x^i, z^i|\theta)] \\
&= E_{q(z^1,z^2,\ldots z^n|x^i)}[\log \Pi_{i=1}^n \pi_{z^i}\mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})] \\
&= E_{q(z^1,z^2,\ldots z^n|x^i)}[\sum_{i=1}^n \log \pi_{z^i} + \log \frac{1}{(2\pi)^{d/2}|\Sigma_{z^i}|^{1/2}} \exp[-\frac{1}{2}(x^i-\mu_{z^i})^T\Sigma_{z^i}^{-1}(x^i-\mu_{z^i})]] \\
&= \sum_{i=1}^n E_{q(z^i|x^i)}[\log \pi_{z^i} - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{z^i}| - \frac{1}{2}(x^i-\mu_{z^i})^T\Sigma_{z^i}^{-1}(x^i-\mu_{z^i})]
\end{aligned}$$

Put $\tau_k^i = p(z_k^i = k|x^i)$ and simplify:

$f(\theta) = \sum_{i=1}^n \sum_{k=1}^K \tau_k^i[\log \pi_k - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{z^i}| - \frac{1}{2}(x^i-\mu_{z^i})^T\Sigma_{z^i}^{-1}(x^i-\mu_{z^i})]$

Adding Lagrangian constraints:

$L = \sum_{i=1}^n \sum_{k=1}^K \tau_k^i[\log \pi_k - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{z^i}| - \frac{1}{2}(x^i-\mu_{z^i})^T\Sigma_{z^i}^{-1}(x^i-\mu_{z^i})] + \lambda(1 - \sim_{k=1}^K \pi_k)$

Setting derivative of this wrt π_k to 0:

$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^n \frac{\tau_k^i}{\pi_k} - \lambda = 0 \implies \pi_k = \frac{\sum_{i=1}^n \tau_k^i}{\lambda}$

We know that $\sum_{k=1}^K \pi_k^i = 1$.

$\sigma_{k=1}^K \frac{1}{\lambda}\sum_{i=1}^n \tau_k^i = 1 \implies \frac{1}{\lambda}n = 1 \implies \lambda = n \implies \boxed{\pi_k = \frac{\sum_{i=1}^n \tau_k^i}{n}}$

Setting derivative of L wrt μ_k to 0:

$\frac{\partial L}{\partial \mu_k} = \sum_{i=1}^n \tau_k^i \sum_k^{-1}(x^i - \mu_k) = 0 \implies \boxed{\mu_k = \frac{\sum_{i=1}^n \tau_k^i x^i}{\sum_{i=1}^n \tau_k^i}}$

Setting derivative of L wrt Σ_k to 0:

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

Using $\frac{\partial}{\partial \Sigma_k^{-1}}\log|\Sigma_k^{-1}| = \Sigma_k^T$ and $\frac{\partial}{\partial \Sigma_k^{-1}} tr(\Sigma_k^{-1}(x^i-\mu_k)^T(x^i-\mu_k)) = (x^i-\mu_k)(x^i-\mu_k)^T$ from PRML Appendix.

(d)

With latent variable z, the log likelihood function:

$$l(x;\theta) = \log \sum_{z \in Z} p(x, z|\theta)$$

Consider a $q(z)$ such that $q(z) \geq 0$ and $\sum_{z \in Z} q(z) = 1$.

$$l(x;\theta) = \log \sum_{z \in Z} q(z)\frac{p(x,z|\theta)}{q(z)}$$

Applying Jensen's inequality for convex functions here gives us

$$l(x;\theta) \geq \sum_{z \in Z} q(z) \log \frac{p(x,z|\theta)}{q(z)}$$

Let us call $L = (q,\theta) \sum_{z \in Z} q(z) \log \frac{p(x,z|\theta)}{q(z)}$.

From PRML 9.4, $l(x;\theta)$ is non decreasing in each iteration of EM. In the E step, $L(q,\theta)$ is maximised keeping θ fixed. This gives us $q(z) = p(z|x,\theta)$. Using this $q(z)$, we get that the log likelihood $l(x;\theta)$ equals its lower bound $L(q,\theta)$.

In the M step, $L(q,\theta)$ is maximised keeping $q(z)$ fixed. Let us call the θ we obtain from solving this optimization problem θ'. As we are maximising $L(q,\theta)$, we know that $L(q,\theta') \geq L(q,\theta)$. From the E step, $l(x;\theta) = L(q,\theta)$. Applying Jensen's inequality again gives us $l(x;\theta') \geq L(q,\theta')$. Putting these inequalities together gives is

$$l(x,\theta') \geq l(x;\theta)$$

This shows us that the log likelihood function will be nondecreasing in each iteration of this algorithm, which implies that the EM algorithm converges.

(e)

From PRML 9.3.2

Consider a GMM with a common covariance matrix of ϵI for all the mixtures.

$\tau_k^i = \frac{\pi_k \exp\{-||x_i - \mu_k||^2/2\epsilon\}}{\sum_j \pi_j \exp\{-||x_i - \mu_j||^2/2\epsilon\}}$

If we consider the limit $\epsilon \longrightarrow 0$, we see that in the denominator the term for which $||x_n - \mu_j||^2$ is smallest will go to zero most slowly, and hence the responsibilities τ_k for the data point will all go to zero except τ_j which will go to unity. This is the hard assignment of data points to clusters, aka, the E step.

In the M step, as $\epsilon \longrightarrow 0$, the expected complete-data log likelihood becomes

$E_Z[\ln p(X,Z|\mu,\Sigma,\pi)] \longrightarrow -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 + const.$

Maximising this expression is equivalent to minimising the cost expression for K-means given by $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$. This involves calculating the partial derivative of the expression and setting it to 0. This is the center recomputation step in K-means.

2

(a)

$$
\begin{aligned}
L(\theta \mid D) &= \prod_{i=1}^{m} P\left(x_i \mid \theta\right) \\
&= \left[\theta^{x_1}(1-\theta)^{1-x_1}\right]\left[\theta^{x_2}(1-\theta)^{1-x_2}\right]_{\ldots} \cdot \left[\theta^{x_m}(1-\theta)^{m}\right. \\
&= \theta^{x_1+x_2+\ldots x_m} \quad (1-\theta)^{m-(x_1+x_2+\ldots x_m)}
\end{aligned}
$$

Applying log

$$\log(L(\theta \mid D) = \frac{m\bar{x}}{\theta} + \frac{m-m\bar{x}}{1-\theta}(-1)$$

Setting this derivative to 0

$$
\begin{aligned}
\bar{x}(1-\theta) &= (1-\bar{x})\theta \\
\bar{x}-\bar{x}\theta &= \theta-\bar{x}\theta \\
\theta = \bar{x} &= \boxed{\frac{\sum_{i=1}^{m} x_i}{m}}
\end{aligned}
$$

(b)

$$
\begin{aligned}
L(\mu,\sigma^2|D) &= \Pi_{x_i\in D}P(x_i|\mu,\sigma^2) \\
&= \left(\frac{1}{2\sigma^2}e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2}\right)\left(\frac{1}{2\sigma^2}e^{-\frac{1}{2\sigma^2}(x_2-\mu)^2}\right)\ldots\left(\frac{1}{2\sigma^2}e^{-\frac{1}{2\sigma^2}(x_m-\mu)^2}\right)
\end{aligned}
$$

Applying log

$$\log(L(\mu,\sigma^2|D)) = m\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \frac{1}{2\sigma^2}\left[(x_1-\mu)^2+(x_2-\mu)^2+\ldots(x_m-\mu)^2\right]$$

Setting derivative of this wrt σ^2 to 0

$$\frac{\partial}{\partial\sigma^2}\log(L(\mu,\sigma^2|D)) = \frac{-m}{2\sigma^2} + \frac{1}{2\sigma^4}\left[(x_1-\mu)^2+(x_2-\mu)^2+\ldots(x_m-\mu)^2\right] = 0$$

$$\boxed{\sigma^2 = \frac{1}{m}\sum_{x_i\in D}(x_i-\mu)^2} \longrightarrow \textit{variance}$$

Setting derivative of the log likelihood wrt μ to 0

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\log(L(\mu,\sigma^2|D)) &= \frac{1}{\sigma^2}\left[(x_1-\mu)+(x_2-\mu)+\ldots(x_m-\mu)\right] \\
m\mu &= \sum_{i=1}^{m} x_i
\end{aligned}
$$

$$\boxed{\mu = \frac{\sum_{i=1}^{m} x_i}{m}} \longrightarrow \textit{mean}$$

(c)

$p(x) = \frac{\text{number of observations within } B_k}{\text{number of observations}}\frac{1}{\text{length of the bin}} = \sum_{j=1}^{n}\frac{nc_j}{m}I(x\in B_j)$

Requirement for density $p(x)$:

$p(x) \geq 0$ and $\int_\Omega p(x)dx = 1$

For histogram, clearly $p(x) \geq 0$.

$$\begin{aligned}
\int_{[0,1)} p(x)dx &= \int_{[0,1)} \sum_{j=1}^{n} \frac{nc_j}{m} I(x \in B_j)dx \\
&= \sum_{j=1}^{n} \int_{\left[\frac{j-1}{n}, \frac{j}{n}\right)} \frac{nc_j}{m} dx \\
&= \sum_{j=1}^{n} \frac{c_j}{m} = 1
\end{aligned}$$

$\therefore$, $\int_\Omega p(x)dx = 1$ for histogram.

(d)

From lecture slides 6

1

Parametric models are those which can be described by a fixed number of parameters. eg. Bernoulli distribution, Gaussian Distribution

Non parametric models are those which cannot be described by a fixed number of parameters. eg. Histogram, Kernel density estimator

2

A smoothing kernel function should have the following 4 properties:

$K(u) \geq 0$

$\int K(u)du = 1$

$\int uK(u)du = 0$

$\int u^2K(u)du < \infty$

3

Curse of dimensionality: When we are working with high dimensional data, i.e. $n^d > m$, we will require a lot of bins but most bins will be empty.

Further, the output depends on where we put the bins and the number of bins.

Kernel Density Estimators are more flexible because we can vary the bandwidth and use kernels of different shapes and sizes. Information is also not lost due to bin choices like in histogram.

4

Gaussian Kernel: $K(u) = \frac{1}{2\pi}e^{-\frac{u^2}{2}}$

Epanechnikov Kernel: $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$

Cosine Kernel: $K(u) = \frac{\pi}{4}cos\left(\frac{\pi}{2}u\right)I(|u| \leq 1)$

Top hat Kernel: $K(u;h) = \frac{1}{2h}for|u| < h$

5

Non parametric 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.

Non parametric models (not histograms) require storing and computing with the entire data set. Parametric models, once fitted, are much more efficient in terms of storage and computation.

3

(a)

$$\begin{aligned}
H(Z|X) &= \sum_x \sum_z p(z,x) \log p(z|x) \\
&= \sum_x p(x) \sum_z p(z|x) \log p(z|x) \\
&= \sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
&= H(Y|X)
\end{aligned}$$

Similarly, $H(Z|Y) = H(X|Y)$.

If X, Y are independent, $H(Y|X) = H(Y)$ and $H(X|Y) = H(X)$

$$\begin{aligned}
& H(Z) \geq H(Z|X) \\
\implies & H(Z) \geq H(Y|X) \\
\implies & H(Z) \geq H(Y)
\end{aligned}$$

$$\begin{aligned}
& H(Z) \geq H(Z|Y) \\
\implies & H(Z) \geq H(X|Y) \\
\implies & H(Z) \geq H(X)
\end{aligned}$$

(b)

Let k be some constant. Suppose X and Y are dependent such that $P(Y = k - x)|X = x) = 1$. Then $P(z = k) = 1$ which means $H(Z) = 0$.

Here, $H(X) = H(Y) > 0$. $\implies H(X) > H(Z)$ and $H(Y) > H(Z)$.

(c)

X and Y are independent, and these two random variables never sum up to the same value.

$Z = X + Y = f(X, Y)$

$$\begin{aligned} H(Z) &= H(f(X,Y)) \\ &\leq H(X,Y) \\ &= H(X) + H(Y|X) \\ &= H(X) + H(Y) \end{aligned}$$

IF $f(X, Y)$ is such that only one pair of (x, y) maps to one value of z, then $H(f(X,Y)) = H(X,Y)$ and if X, Y are independent then $H(Y|X) = H(Y)$. Hence, $H(Z) = H(X) + H(Y)$.

4

I used np.random.rand (uniform distribution between 0 and 1) to initialise the μs followed by normalising and $\frac{1}{n_c}$ to initialise all the πs as we need them to sum up to 1. Here, $n_c = 4$. Hence, $\pi_c = 0.25$. The algorithm converges when the π values stop changing.

I set a different seed value for each iteration and performed 10 iterations.

Maximum accuracy: 85.75%

Minimum accuracy: 64.00%

Mean accuracy: 76.83%

A more detailed table is below:

Seed	**Iterations till convergence**	**Accuracy**
10	144	82.25
1	205	78.25
3	44	66.75
34	48	73.00
67	69	85.50
45	80	79.75
33	42	75.50
56	50	64.00
23	82	85.75
12	56	77.50

Clearly, the accuracy greatly depends on the initialisation.