CMPUT 466 Assignment 4

Khac Nguyen Nguyen

CCID: khacnguy

1 Name and CCID

Above

2 K-Means Clustering

1. Assign each data point to the closest centroid.
   centroid (2,4): (2,4), (3,3), (3,4)
   centroid (5,8): (5,8), (7,7), (6,9)

2. Recalculate the centroids of the clusters.
   centroid (2,4) $\rightarrow$ (7/3, 11/3)
   centroid (5,8) $\rightarrow$ (6, 8)

3. Assign each data point to the closest centroid.
   centroid (7/3, 11/3): (2,4), (3,3), (3,4)
   centroid (6,8): (5,8), (7,7), (6,9)
   Since there is no change in the assignment of points on clusters. This is the final cluster: centroid (7/3, 11/3): (2,4), (3,3), (3,4)
   centroid (6,8): (5,8), (7,7), (6,9)

3 Derivation of EM

Assuming z can only have an integer value between 1 and K and let X has the value $\frac{p(x_i,z_i=k;\theta)}{q(z_i=k|x_i)}$ with probability $q(z_i = k|x_i)$. We have

$$
\begin{aligned}
l_i(\theta, x_i) &= \log p(x_i;\theta) \\
&= \log \underbrace{\sum_{k=1}^{K} q(z_i = k|x_i) \frac{p(x_i, z_i = k;\theta)}{q(z_i = k|x_i)}}_{E_q[X]} \\
&\geq \sum_{k=1}^{K} q(z_i = k|x_i) \log\left(\frac{p(x_i, z_i = k;\theta)}{q(z_i = k|x_i)}\right) \\
&= \sum_{k=1}^{K} q(z_i = k|x_i)[\log(p(x_i, z_i = k;\theta)) - \log(q(z_i = k|x_i))] \\
&= \sum_{k=1}^{K} q(z_i = k|x_i) \log(p(x_i, z_i = k;\theta)) - \sum_{k=1}^{K} q(z_i = k|x_i) \log(q(z_i = k|x_i))
\end{aligned}
$$

since log is a concave function

4 Formulation of EM

4.1 Expectation step

To maximize F w.r.t q, we want to find q so that the inequality turn into an equality. From the Jensen's inequality, the equal sign holds when $\frac{p(x_i,z_i;\theta)}{q^(z_i=k|x_i)}$ is constant w.r.t k. Thus let

$$q^{t+1}(z_i = k|x_i) = p(z_i = k|x_i;\theta^t)$$

as

$$\frac{p(x_i, z_i;\theta^{t+1})}{p(z_i = k|x_i;\theta^t)} = p(x_i;\theta^t)$$

4.2 Maximization step

Using the new q value, now we need to maximize l and hence $E_q(L_c(X, z;\theta))$ w.r.t θ, thus

$$\theta^{t+1} = \arg\max_{\theta} F(q^{t+1}, \theta) = \arg\max_{\theta} E_{q^{t+1}}(L_c(X, z;\theta))$$

5 EM for MoG

5.1 Expectation Step

$$
\begin{aligned}
q^{t+1}(z_i = k|x_i) &= p(z_i = k|x_i, \theta^t) \\
&= \frac{p(X_i = x_i|z_i = k, \theta^t)p(Z_i = k|\theta^t)}{p(X_i = x_i|\theta^t)} \\
&= \frac{\mathcal{N}(x_i|\mu_k^t, \Sigma_k^t)\alpha_k^t}{\sum_{j=1}^K p(X_i = x_i|z_i = j, \theta^t)p(z_i = j)} \\
&= \frac{\mathcal{N}(x_i|\mu_k^t, \Sigma_k^t)\alpha_k^t}{\sum_{j=1}^K \mathcal{N}(x_i|\mu_j^t, \Sigma_j^t)\alpha_j^t}
\end{aligned}
$$

where

$$
\mathcal{N}(x_i|\mu_k^t, \Sigma_k^t) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k^t|}} \exp\left(-\frac{1}{2}(x_i - \mu_k^t)^T(\Sigma_k^t)^{-1}(x_i - \mu_k^t)\right)
$$

5.2 Maxmization Step

Taking derivative of F and set to 0, we can solve for μ, Σ, α, though intuitively we can understand them as follow:

$$
\mu_k^{t+1} = \frac{\sum_{i=1}^N x_i q_{ki}^{t+1}}{\sum_{i=1}^N q_{ki}^{t+1}}
$$

The mean change to one with weights, where the mean will move more towards point with higher probability.

$$
\Sigma_k^{t+1} = \frac{\sum_{i=1}^N q_{ki}^{t+1}(x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T}{\sum_{i=1}^N q_{ki}^{t+1}}
$$

The new variance based on the new mean.

$$
\alpha_k^{t+1} = \frac{\sum_{i=1}^N q_{ki}^{t+1}}{N}
$$

The new "average" "probability" based on the new q-value calculated on the expectation step.

References

[1] https://math.stackexchange.com/questions/628386/when-jensens-inequality-is-equality