INF554 - Machine Learning I

Lab 3: SOLUTIONS

Question 1

One reasonable way to choose the number of clusters is by considering the plot we have just created and observing after which number k of centroids the objective function of the K-means algorithm is not significantly improved by the addition of further centroids. In the plot we have obtained here this would lead us to choose K equal to 4 centroids in our K-means implementation. Choosing the number of centroids on the basis of a plot of the objective function of the K-means algorithm is often referred to as the "Elbow-method".

Question 2

As we have observed in our implementation the K-means algorithm is sensitive to it's initialisation. Hence, if we initialise several centroids in a cluster then there is a non-negligible chance that we end up fitting several centroids to a single cluster.

Question 3

1)

We begin by calculating the likelihood $L(x, z; \theta)$.

$$\begin{aligned} L(x, z; \theta) &= P(x, z|\theta) \\ &= P(x|z, \theta)P(z|\theta) && (1) \\ &= \prod_{i=1}^{n} P(x_i|z_i, \theta)P(z_i|\theta) && (2) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{K} \phi(x_i|\mu_j, \Sigma_j)^{\mathbb{I}(z_i=j)} w_j^{\mathbb{I}(z_i=j)}, && (3) \end{aligned}$$

where in (1) we use Bayes Theorem, in (2) we assume our data points are independent and in (3) we use the model definition. Hence, we are able to proceed to calculated the log likelihood $l(x, z; \theta)$.

$$\begin{aligned} \Rightarrow l(x, z; \theta) &= \log(L(x, z; \theta)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K} \mathbb{I}(z_i = j) \left(\log(\phi(x_i|\mu_j, \Sigma_j)) + \log(w_j)\right). \end{aligned}$$

2)

Now we are able to calculate the expected log likelihood over the distribution $P(z|x, \theta_t)$.

$$\mathbb{E}_{z|x,\theta_t}\left[l(x,z;\theta)\right] = \mathbb{E}_{z|x,\theta_t}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\mathbb{I}(z_i=j)\left(\log(\phi(x_i|\mu_j,\Sigma_j))+\log(w_j)\right)\right]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{K}\left(\log(\phi(x_i|\mu_j,\Sigma_j))+\log(w_j)\right)\mathbb{E}_{z|x,\theta_t}\left[\mathbb{I}(z_i=j)\right].$$

Note that the expectation of an indicator function is eqaul to the probability of the event in the indicator function.

$$\begin{aligned}
\mathbb{E}_{z|x,\theta_t}\left[\mathbb{I}(z_i=j)\right] &= P(z_i=j|x,\theta_t) \\
&= \frac{P(z_i=j,x|\theta_t)}{P(x|\theta_t)} && (4)\\
&= \frac{P(x|z_i=j,\theta_t)P(z_i=j|\theta_t)}{P(x|\theta_t)} && (5)\\
&= \frac{\phi(x_i|\mu_j^t,\Sigma_j^t)w_j}{\sum_{j=1}^{K} w_j\phi(x_i|\mu_j^t,\Sigma_j^t)} && (6)\\
&\equiv \gamma_{ij}(\theta_t).
\end{aligned}$$

Here we used Bayes theorem in (4) and in (5) and the model definition in (6). As discussed in the lab and lecture $\gamma_{ij}(\theta_t)$ is called the responsibility of the normal distribution j for data point x_i.

$$\begin{aligned}
\Rightarrow \mathbb{E}_{z|x,\theta_t}\left[l(x,z;\theta)\right] &= \sum_{i=1}^{n}\sum_{j=1}^{K}\left(\log(\phi(x_i|\mu_j,\Sigma_j))+\log(w_j)\right)\gamma_{ik}(\theta_t) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{K}\gamma_{ij}(\theta_t)\left(\log(w_j)+\log\left(\frac{1}{\sqrt{(2\pi)^d\det(\Sigma_j)}}\right)-\frac{1}{2}(x_i-\mu_j)^T\Sigma_k^{-1}(x_i-\mu_j)\right).
\end{aligned}$$

3)

In order to find the parameters optimising the expected log likelihood we differentiate it and set it eqaul to 0 beginning with μ_k.

$$\begin{aligned}
\frac{\partial}{\partial\mu_k}\mathbb{E}_{z|x,\theta_t}\left[l(x,z;\theta)\right] &= \sum_{i=1}^{n}\gamma_{ik}(\theta_t)\frac{\partial}{\partial\mu_k}\left(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\gamma_{ik}(\theta_t)2\Sigma_k^{-1}(x_i-\mu_k).
\end{aligned}$$

Setting this expression equal to 0 we obtain,

$$\begin{aligned}
0 &= \frac{\partial}{\partial\mu_k}\mathbb{E}_{z|x,\theta_t}\left[l(x,z;\theta)\right]. \\
\Rightarrow 0 &= \Sigma_k^{-1}\sum_{i=1}^{n}\gamma_{ik}(\theta_t)(x_i-\hat{\mu}_k). && (7)\\
\Rightarrow \hat{\mu}_k &= \frac{\sum_{i=1}^{n}\gamma_{ik}(\theta_t)x_i}{\sum_{i=1}^{n}\gamma_{ik}(\theta_t)}.
\end{aligned}$$

In (7) we utilised the fact that Σ_k^{-1} is independent of the index we are summing over and left multiplied Σ_k to elminiate it.

We proceed analogously for Σ_k by first taking the derivative of the expected log likelihood.

$$\begin{aligned}\frac{\partial}{\partial \Sigma_k}\mathbb{E}_{z|x,\theta_t}\left[l(x,z;\theta)\right] &= \sum_{i=1}^{n}\gamma_{ik}(\theta_t)\left(\frac{\partial}{\partial \Sigma_k}\left(-\frac{1}{2}\log\left(\det(\Sigma_k)\right)\right)+\frac{\partial}{\partial \Sigma_k}\left(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right)\right)\\ &= \sum_{i=1}^{n}\gamma_{ik}(\theta_t)\left(-\frac{1}{2}\Sigma_k^{-1}+\frac{1}{2}\Sigma_k^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\Sigma_k^{-1}\right).\end{aligned}$$

Setting this expression equal to 0 we obtain,

$$\begin{aligned}0 &= \sum_{i=1}^{n}\gamma_{ik}(\theta_t)\left(-\frac{1}{2}\hat{\Sigma}_k^{-1}+\frac{1}{2}\hat{\Sigma}_k^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\hat{\Sigma}_k^{-1}\right). \qquad (8)\\ \Rightarrow 0 &= \sum_{i=1}^{n}\gamma_{ik}(\theta_t)\left(-\hat{\Sigma}_k+(x_i-\mu_k)(x_i-\mu_k)^T\right).\\ \Rightarrow \hat{\Sigma}_k &= \frac{\sum_{i=1}^{n}\gamma_{ik}(\theta_t)(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}\gamma_{ik}(\theta_t)}.\end{aligned}$$

Here we left and right multiplied Σ_k in (8).

For the optimisation of w_k we need to use a Lagrange multiplier to appropriately take the constraint $\sum_{j=1}^{K} w_j = 1$ into account.

$$\frac{\partial}{\partial w_k}\left(\mathbb{E}_{z|x,\theta_t}\left[l(x,z;\theta)\right]-\lambda\left(\sum_{j=1}^{K}w_j-1\right)\right) = \sum_{i=1}^{n}\frac{\gamma_{ik}(\theta_t)}{w_k}-\lambda.$$

Setting this to 0 we obtain,

$$\begin{aligned}0 &= \sum_{i=1}^{n}\frac{\gamma_{ik}(\theta_t)}{\hat{w}_k}-\lambda\\ \Rightarrow \hat{w}_k\lambda &= \sum_{i=1}^{n}\gamma_{ik}(\theta_t).\end{aligned}$$

Now summing over k to find λ we obtain,

$$\begin{aligned}\sum_{k=1}^{K}\hat{w}_k\lambda &= \sum_{j=1}^{K}\sum_{i=1}^{n}\gamma_{ik}(\theta_t).\\ \Rightarrow \lambda &= \sum_{i=1}^{n}\sum_{j=1}^{K}\gamma_{ik}(\theta_t).\\ \Rightarrow \lambda &= \sum_{i=1}^{n}1.\\ \Rightarrow \lambda &= n.\\ \Rightarrow \hat{w}_k &= \frac{\sum_{i=1}^{n}\gamma_{ik}(\theta_t)}{n}.\end{aligned}$$