Homework 2 Part 1 - Solutions

Problem 1 (25 points)

Suppose you have reason to believe that your collected data $\mathbf{X} = \{x_i\}_{i=1}^N$, where $x_i \geq 0, \forall i$, can be modeled using a Mixture Model, in particular, a Rayleigh Mixture Model. Its data likelihood can be written as:

$$p(x) = \sum_{k=1}^{K} \pi_k f_k\left(x|\sigma_k\right)$$

where

$$\sum_{k=1}^{K} \pi_k = 1$$

and

$$f_k\left(x|\sigma_k\right) = \frac{x}{\sigma_k^2} e^{-x^2/(2\sigma_k^2)}$$

with $\sigma_k > 0$ and $0 \leq \pi_k \leq 1, \forall k$. Answer the following questions:

1. (3 points) **Assuming your data is i.i.d., write down the observed data likelihood, $\mathcal{L}^0$.**

The observed data likelihood is:

$$\mathcal{L}^0 = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \frac{x_i}{\sigma_k^2} e^{-x_i^2/(2\sigma_k^2)}$$

1. (3 points) **Can you introduce hidden latent variables Z to this problem? Describe precisely what they are.**

Yes, we can introduce the hidden latent variable Z, where z_i corresponds to the Rayleigh component from which sample x_i was drawn from. Moreover, since we have a total of K Rayleigh components, $z_i \in \{1, 2, \ldots, K\}$.

1. (4 points) **For the hidden variables you defined above, write down the complete data likelihood, $\mathcal{L}^c$.**

The complete data likelihood is given by:

$$\mathcal{L}^c = \prod_{i=1}^{N} \pi_{z_i} \frac{x_i}{\sigma_{z_i}^2} e^{-x_i^2/(2\sigma_{z_i}^2)}$$

1. (5 points) **Write down the EM optimization function, $Q(\Theta, \Theta^t)$, where $\Theta = \{\pi_k, \sigma_k\}_{k=1}^K$. Your final solution should contain the sum of simple log-terms.**

The EM optimization function is given by:

$$\begin{aligned}
Q(\Theta, \Theta^t) &= \mathbb{E}_z[\ln(\mathbf{L}^c)|\mathbf{X}, \Theta^t] \\
&= \sum_{z_i=1}^{K} \ln(\mathbf{L}^c) P(z_i|x_i, \Theta^t) \\
&= \sum_{k=1}^{K} \left[\sum_{i=1}^{N} \left(\ln(\pi_k) + \ln(x_i) - 2\ln(\sigma_k) - \frac{x_i^2}{2\sigma_k^2} \right) \right] P(z_i = k|x_i, \Theta^t) \\
&= \sum_{k=1}^{K} \left[\sum_{i=1}^{N} \left(\ln(\pi_k) + \ln(x_i) - 2\ln(\sigma_k) - \frac{x_i^2}{2\sigma_k^2} \right) \right] C_{ik}
\end{aligned}$$

1. (5 points) **Derive the update equations for the parameters σ_k.**

The solution for the parameter σ_k is:

$$\begin{aligned}
\frac{\partial Q(\Theta, \Theta^t)}{\partial \sigma_k} &= 0 \\
\sum_{i=1}^{N} \left(-\frac{2}{\sigma_k} + \frac{4\sigma_k x_i^2}{(2\sigma_k^2)^2} \right) C_{ik} &= 0 \\
\sum_{i=1}^{N} \left(-2\sigma_k^2 + x_i^2 \right) C_{ik} &= 0 \\
\sum_{i=1}^{N} 2\sigma_k^2 C_{ik} &= \sum_{i=1}^{N} x_i^2 C_{ik} \\
\sigma_k &= \sqrt{\frac{\sum_{i=1}^{N} x_i^2 C_{ik}}{\sum_{i=1}^{N} 2C_{ik}}}
\end{aligned}$$

1. (5 points) **Derive the update equations for the parameters π_k.**

In order to find the solution for π_k, we must add the constraint $\sum_{k=1}^{K} \pi_k = 1$ to the optimization function:

$$Q_\pi(\Theta, \Theta^t) = Q(\Theta, \Theta^t) - \lambda \left(\sum_{k=1}^{N} \pi_k - 1 \right)$$

The solution for π_k is then:

$$\frac{\partial Q_\pi(\Theta, \Theta^t)}{\partial \pi_k} = 0$$

$$\sum_{i=1}^{N} \frac{1}{\pi_k} C_{ik} - \lambda = 0$$

$$\pi_k = \frac{1}{\lambda} \sum_{i=1}^{N} C_{ik}$$

Since $\sum_{k=1}^{K} \pi_k = 1$, we find that: $\sum_{k=1}^{K} \frac{1}{\lambda} \sum_{i=1}^{N} C_{ik} = 1 \iff \lambda = N$. Hence, the solution for π_k is:

$$\pi_k = \frac{\sum_{i=1}^{N} C_{ik}}{N}$$

Problem 2 (7.5 points)

Consider the scenario where you have a dataset of training samples, x_i and its corresponding target labels, t_i, $D = \{(x_i, t_i)\}_{i=1}^{N}$. Consider the case where the feature space is at least 2-dimensional, that is, $x_i \in \mathbb{R}^D$, where $D > 1$.

In the process of the dataset D, you notice that some samples are missing.

Explain in words and show equations on how the EM algorithm can be used to perform density estimation of this dataset with missing samples.

Similarly to the example of censored data, we can introduce a latent hidden variable Z that will correspond to the true (unknown) missing value x_i. The EM algorithm will be allow us to perform standard density estimation while making an estimate for what the value of the missing sample could be.

Let samples $\{x_j\}_{j=m}^{N}$ be the samples with missing values.

The *observed* data likelihood as:

$$\mathcal{L}^0 = \prod_{i=1}^{m} f_X(x_i|\Theta) \prod_{j=m+1}^{N} \int_{-\infty}^{\infty} f_X(x_j|\Theta) dx_j$$

where Θ is the set of parameters of the PDF $f_X(x)$.

Let the hidden latent variables be defined as:

$$z_j : \text{true (unknown) value for the missing value } x_j$$

we can write the complete data likelihood:

$$\mathcal{L}^c = \prod_{i=1}^{m} f_X(x_i|\Theta) \prod_{j=m+1}^{N} f_X(z_j|\Theta) dz_j$$

We can know optimize the EM Q-function to iteratively solve for the hidden latent variables Z (E-STEP) and the parameters of the distribution Θ (M-STEP).

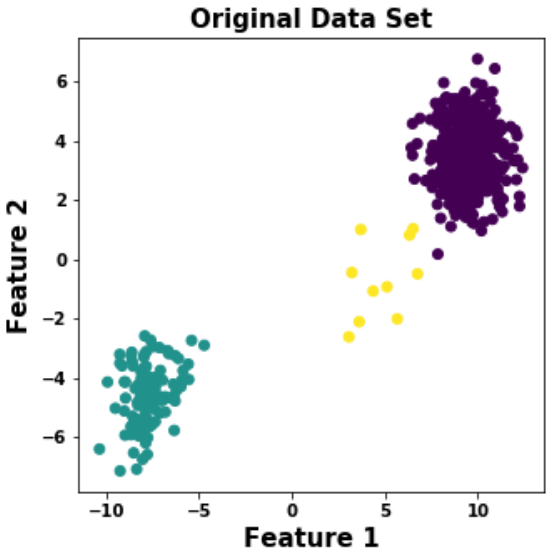
$$\begin{aligned} Q(\Theta, \Theta^t) &= E[\ln(\mathcal{L}^c)|X, \Theta^t] \\ &= \sum_{\mathbf{z}} \ln(\mathcal{L}^c) P(z|X, \Theta^t) \end{aligned}$$

Problem 3 (5 points)

Consider the following dataset:

In [2]:
```
from IPython.display import Image
Image('figures/Clustering.png',width=300)
```

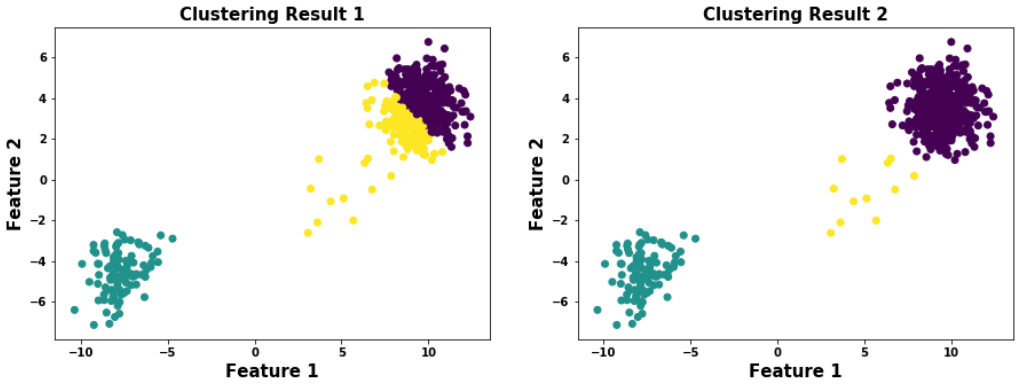
Out[2]:



Suppose that both k-Means and a Gaussian Mixture Model were applied to this data set. Form the results below, which clustering algorithm produced each plot? Justify your answer.

In [3]:
```
Image('figures/ClusteringResults.png',width=1000)
```

Out[3]:



"Clustering Result 1" (figure on the left) was produced by K-Means and "Clustering Result 2" (figure on the right) was produced by GMM.

K-Means tries to minimize the within-cluster distances, for this reason it will prefer to split up a much larger cluster into two than assigning a cluster with a small number of points where each point is is far away from the center. We can see this in the figure on the left, where the dense cluster was split into two by a harsh straight line. The straight line cutoff is also an indicator of K-Means as K-Means assumes the cluster are spherical.

On the other hand, the figure on the right shows a nice clustering result when compared with the original data. GMMs are powerful models as they are capable of modeling clusters regardless of their size.

Problem 4 (5 points)

Suppose you would like to learn the number of clusters needed for your data set while running k-means. To accomplish this, suppose you are able to devise an optimization strategy to minimize the k-means objective function with respect to the number clusters, in addition to the cluster centers and the crisp membership labels,

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} \|x_i - \theta_k\|_2^2$$

where $u_{ik} = \{0, 1\}$ **and** $\sum_{k=1}^{K} u_{ik} = 1$**.**

Would this be an effective approach to learning the number of clusters? Explain your reasoning.

No, this objective function will not be an effective approach to learning the number of clusters. Because when the number of clusters is the same as the number of data samples, $k = N$, each data point will be its own cluster centroid, and the objective function will be at its minimum value $J = 0$. But this is not a good clustering result.

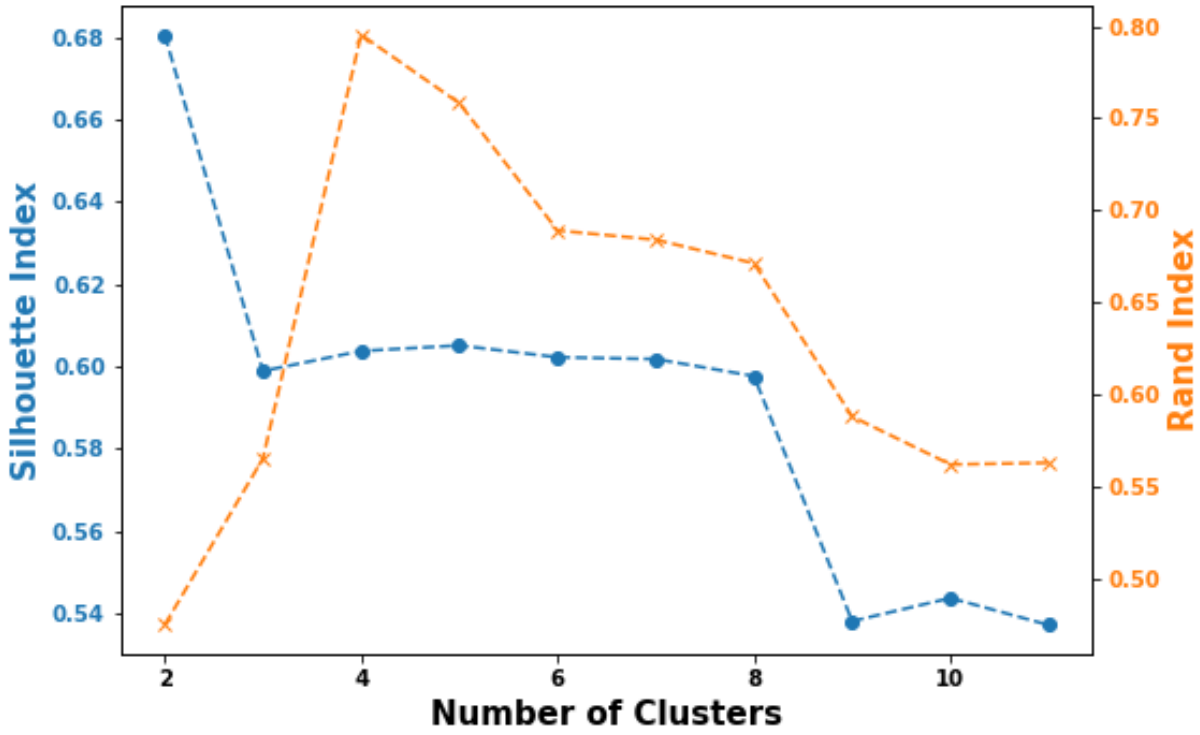
Alternatively, we should use cluster validity metrics to select the best value of k to partition the data into.

Problem 5 (5 points)

Suppose that you ran the k-Means algorithm and the GMM on the data for a given range for the number of clusters. You have computed the silhouette index for the k-Means clustering performance, and computed the rand index to compare the clustering results between k-Means and GMM. This is the plot you obtained:

In [4]:
```
Image('figures/ClusterValidity.png',width=700)
```

Out[4]:



1. **What should be the best choice for the number of clusters (for each index measure) based on the results shown in this figure? Explain your reasoning.**

2. **Do both measures (silhouette index and rand index) agree on the best number of clusters? Explain what this may indicate about the true underlying data clusters.**

1. The silhouette index is a type of internal criteria index. It *prefers* a clustering result for which the cluster are compact and far away from each other. The value of silhouette index varies from -1 and 1 and higher indicates better clustering results. In this figure, according to the silhouette index, the best choice for the number of clusters is $k = 2$ as the silhouette index is maximized.

The rand index is a type of external criteria index. It computes the ratio between number of agreements between clustering result and ground truth over the number of pairs in the data. The value of rand index varies from 0 and 1 and higher indicates higher consistency. In this figure, according to the rand index, the best choice for the number of clusters is $k = 4$ as the rand index is maximized.

1. Both indices do not agree on the *best choice* for the number of cluster, this may be an indicator that there is sufficient overlap between clusters.
