Aprendizagem 2022

Homework IV – Group 019

Diogo Gaspar 99207, Rafael Oliveira 99311

Part I: Pen and paper

Given the bivariate observations $\left\{\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}-1\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}\right\}$, and the following multivariate Gaussian mixture:

$$\mu_1 = \begin{bmatrix}2\\2\end{bmatrix}, \quad \mu_2 = \begin{bmatrix}0\\0\end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}, \quad \pi_1 = \pi_2 = 0.5$$

1. **Perform one epoch of the EM clustering algorithm and determine the new parameters.**

   As a side note, we'll be using the k_1 and k_2 notation to represent clusters 1 and 2 - with that, we'll say that $\pi_1 = P(C = k_1)$, with analogous notation for π_2.

   EM-Clustering, being an unsupervised learning algorithm intending to calculate the probability of a sample belonging to a certain cluster, is a method that iteratively updates the parameters of the model until convergence is reached (for a given definition of convergence). Here, we'll perform exactly one epoch of the algorithm, which means we'll be going through two steps:

   - **E-step:** Here, we're aiming to calculate the **posterior probability** of each sample belonging to each cluster. In order to perform this calculation, we'll be using **Bayes' rule**, of course, to decompose the posterior probability into the product of the **likelihood** and the **prior probability** of the sample belonging to the cluster. Let's try, then, to assign each sample to the cluster that maximizes the posterior probability.

     For starters, we must first note that the likelihood of a sample belonging to a cluster is given by the **multivariate Gaussian distribution**, which can be written as (considering $d = 2$):

     $$P(x_i \mid C = k_n) \sim \mathcal{N}(x_i; \mu_n, \Sigma_n) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma_n}} \exp\left(-\frac{1}{2}(x - \mu_n)^T \Sigma_n^{-1} (x - \mu_n)\right)$$

     Moreover, in this step we'll use teal to denote the priors and purple to denote the likelihoods.

As a given, we have that the priors are (for every sample, of course):

$$P(C = k_1) = P(C = k_2) = 0.5$$

Regarding x_1, we have:

$$\begin{aligned} P(x_1 \mid C = k_1) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_1}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_1^{-1}(x_1 - \mu_1)\right) \\ &= \frac{1}{\sqrt{(2\pi)^2 \det \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}\right)^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}\right)\right) \\ &= 0.0658407 \end{aligned}$$

$$\begin{aligned} P(x_1 \mid C = k_2) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_2}} \exp\left(-\frac{1}{2}(x_1 - \mu_2)^T \Sigma_2^{-1}(x_1 - \mu_2)\right) \\ &= \frac{1}{\sqrt{(2\pi)^2 \det \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)\right) \\ &= 0.0227993 \end{aligned}$$

The (normalized) posteriors can be computed as follows:

$$\begin{aligned} P(C = k_1 \mid x_1) &= \frac{P(C = k_1)P(x_1 \mid C = k_1)}{P(C = k_1)P(x_1 \mid C = k_1) + P(C = k_2)P(x_1 \mid C = k_2)} \\ &= \frac{0.5 \cdot 0.0658407}{0.5 \cdot 0.0658407 + 0.5 \cdot 0.0227993} \\ &= 0.742788 \end{aligned}$$

$$\begin{aligned} P(C = k_2 \mid x_1) &= \frac{P(C = k_2)P(x_1 \mid C = k_2)}{P(C = k_1)P(x_1 \mid C = k_1) + P(C = k_2)P(x_1 \mid C = k_2)} \\ &= \frac{0.5 \cdot 0.0227993}{0.5 \cdot 0.0658407 + 0.5 \cdot 0.0227993} \\ &= 0.257212 \end{aligned}$$

Note that, with the aid of the total probability law, we can say that $P(C = k_1 \mid x_1) + P(C = k_2 \mid x_1) = 1$; going forward, we'll calculate the normalized posterior for k_2 utilizing this fact.

We can now repeat the same process for x_2:

$$
\begin{aligned}
P(x_2 \mid C = k_1) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_1}} \exp\left(-\frac{1}{2}(x_2 - \mu_1)^T \Sigma_1^{-1} (x_2 - \mu_1)\right) \\
&= \frac{1}{\sqrt{(2\pi)^2 \det \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} -1 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}\right)^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \left(\begin{bmatrix} -1 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}\right)\right) \\
&= 0.00891057
\end{aligned}
$$

$$
\begin{aligned}
P(x_2 \mid C = k_2) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_2}} \exp\left(-\frac{1}{2}(x_2 - \mu_2)^T \Sigma_2^{-1} (x_2 - \mu_2)\right) \\
&= \frac{1}{\sqrt{(2\pi)^2 \det \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} -1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \left(\begin{bmatrix} -1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)\right) \\
&= 0.0482662
\end{aligned}
$$

The (normalized) posteriors can be computed as follows:

$$
\begin{aligned}
P(C = k_1 \mid x_2) &= \frac{P(C = k_1) P(x_2 \mid C = k_1)}{P(C = k_1) P(x_2 \mid C = k_1) + P(C = k_2) P(x_2 \mid C = k_2)} \\
&= \frac{0.5 \cdot 0.00891057}{0.5 \cdot 0.00891057 + 0.5 \cdot 0.0482662} \\
&= 0.155843
\end{aligned}
$$

Like stated above, using the total probability law, we can say that $P(C = k_1 \mid x_2) + P(C = k_2 \mid x_2) = 1$; therefore, $P(C = k_2 \mid x_2) = 1 - P(C = k_1 \mid x_2) = 0.844157$.
Finally, repeating the same process for x_3:

$$
\begin{aligned}
P(x_2 \mid C = k_1) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_1}} \exp\left(-\frac{1}{2}(x_3 - \mu_1)^T \Sigma_1^{-1} (x_3 - \mu_1)\right) \\
&= \frac{1}{\sqrt{(2\pi)^2 \det \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}\right)^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}\right)\right) \\
&= 0.0338038
\end{aligned}
$$

$$
\begin{aligned}
P(x_3 \mid C = k_2) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_2}} \exp\left(-\frac{1}{2}(x_3 - \mu_2)^T \Sigma_2^{-1} (x_3 - \mu_2)\right) \\
&= \frac{1}{\sqrt{(2\pi)^2 \det \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)\right) \\
&= 0.061975
\end{aligned}
$$

The (normalized) posteriors can be computed as follows:

$$
\begin{aligned}
P(C = k_1 \mid x_3) &= \frac{P(C = k_1) P(x_3 \mid C = k_1)}{P(C = k_1) P(x_3 \mid C = k_1) + P(C = k_2) P(x_3 \mid C = k_2)} \\
&= \frac{0.5 \cdot 0.0338038}{0.5 \cdot 0.0338038 + 0.5 \cdot 0.061975} \\
&= 0.352936
\end{aligned}
$$

$$
P(C = k_2 \mid x_3) = 1 - P(C = k_1 \mid x_3) = 0.647064
$$

- **M-Step: Having calculated the posteriors, we can now update the parameters of the cluster-defining distributions.**

2. **Given the updated parameters computed in previous question:**

(a) **Perform a hard assignment of observations to clusters under a MAP assumption.**

(b) **Compute the silhouette of the larger cluster using the Euclidean distance.**

Part II: Programming and critical analysis

The code utilized to answer the following questions is available in this report's appendix.

Recall the `pd_speech.arff` dataset from earlier homeworks, centered on the Parkinson diagnosis from speech features. For the following exercises, normalize the data using `sklearn`'s `MinMaxScaler`.

3. **Using `sklearn`, apply k-means clustering fully unsupervisedly (without targets) on the normalized data with $k = 3$ and three different seeds (using random $\in \{0, 1, 2\}$). Assess the silhouette and purity of the produced solutions.**

4. **What is causing the non-determinism?**

5. **Using a scatter plot, visualize side-by-side the labeled data using as labels: i) the original Parkinson diagnoses, and ii) the previously learned $k = 3$ clusters (`random = 0`). To this end, select the two most informative features as axes and color observations according to their label. For feature selection, select the two input variables with highest variance on the MinMax normalized data.**

6. **The fraction of variance explained by a principal component is the ratio between the variance of that component (i.e., its eigenvalue) and total variance (i.e., sum of all eigenvalues). How many principal components are necessary to explain more than 80% of variability?**

Appendix