Short Assignment 3

This is an individual assignment.

For the analytical problems, you can write your answers in markdown cells with $\LaTeX$ or `push` a single pdf to your repository with all handwritten answers.

- *Always* show complete work and justify your answers.

Exercise 1 (3.5 points)

Use the Expectation-Maximization (EM) algorithm to solve for the parameters of an Rayleigh Mixture Model (with K components) given a set of training data $\mathbf{X} = \{x_i\}_{i=1}^N$, where $x_i \geq 0, \forall i$. Recall the form of the Rayleigh probability density function is defined as:

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

Answer the following questions:

1. (0.5 points) **Assuming your data is i.i.d., write down the observed data likelihood, $\mathcal{L}^0$.**

$$\mathcal{L}^0 = \prod_{i=1}^N \sum_{k=1}^K \pi_k P\left(x|\sigma_k\right)$$

```
In [ ]:
```

1. (0.5 points) **Introduce the latent hidden variables $\mathbf{z}$ and write down the complete data likelihood, $\mathcal{L}^c$.**

z_i =label of distribution from which x_i was drown from

$$\mathcal{L}^c = \prod_{i=1}^{N} \pi_{z_i} P\left(x|\sigma_{z_i}\right)$$

In [ ]:

1. (0.5 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 (and simplified) log-terms.**

$$\begin{aligned} Q(\Theta, \Theta^t) &= \sum_{z_i=1}^{K} \ln(\mathcal{L}^c) P(\mathbf{z}_i|\mathbf{x}_i, \Theta^t) \\ &= \sum_{z_i=1}^{K} \ln\left(\prod_{i=1}^{N} \pi_{z_i} P\left(x|\sigma_{z_i}\right)\right) P(\mathbf{z}_i|\mathbf{x}_i, \Theta^t) \\ &= \sum_{k=1}^{K} \ln\left(\prod_{i=1}^{N} \pi_{z_i} P\left(x|\sigma_{z_i}\right)\right) P(\mathbf{z}_i = k|\mathbf{x}_i, \Theta^t) \\ &= \sum_{z_i=1}^{K} \sum_{i=1}^{N} \left(\ln(\pi_k) + \ln(P(x_i|\sigma_{z_i}))\right) C_{ik} \\ &= \sum_{z_i=1}^{K} \sum_{i=1}^{N} \left(\ln(\pi_k) + \ln(x_i) - 2\ln(\sigma_k) - \frac{x_i^2}{2\sigma_k^2}\right) C_{ik} \end{aligned}$$

C_{ik} is the expected value of the latent variable.

In [ ]:

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

$$\frac{\partial Q}{\partial \sigma_k} = \sum_{z_i=1}^{K} \sum_{i=1}^{N} \left(\frac{1}{\sigma_k} - \frac{x_i^2}{\sigma_k^3}\right) C_{ik} = 0$$

$$\sigma_k^{t+1} = \sqrt{\frac{\sum_{i=1}^{N} C_{ik} x_i^2}{\sum_{i=1}^{N} 2C_{ik}}}$$

In [ ]:

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

$$\frac{\partial Q}{\partial \pi_k} = \sum_{z_i=1}^{K}\sum_{i=1}^{N}\frac{1}{\pi_k}C_{ik} = 0$$

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

In [ ]:

Exercise 2 (3.5 points)

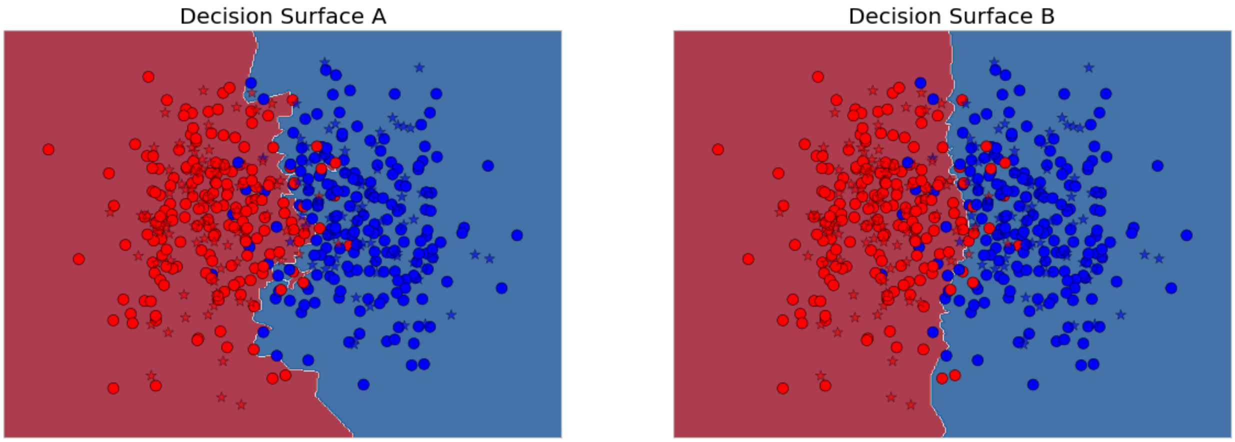
For each of the following statements state whether True (T) or False (F) and justify your answers.

1. (0.5 points) **The figure below shows two decision surfaces obtained from training a K-Nearest Neighbors classifier with the training set (circles) and evaluate prediction on test set (stars). One was obtained for $k = 20$ and the other for $k = 3$. Decision surface A used $k = 3$.**

In [1]:

```
from IPython.display import Image
Image('knn-performances.png',width=900)
```

Out[1]:



True. In surface, the boundry is more jagged, which means it is more influenced by the nearby data points and the classifier becomes more sensitive to local variations and noise in the data. In surface B, the classifier considers a larger neighborhood of data points for making predictions, which means it has a larger k.

In [ ]:

1. (0.5 points) **When performing clustering with the K-Means algorithm using Mahalanobis distance or with Gaussian Mixture Models with full covariance matrices, we should be able to learn elliptical-shaped clusters.**

True. Both of them have covariance matrices, which can adjust for correlations between variables to learn elliptical-shaped clusters.

In [ ]:

1. (0.5 points) **Consider the following confusion matrix:**

	Predicted C_1	**Predicted C_2**	**Predicted C_3**
True C_1	96	3	1
True C_2	3	42	5
True C_3	6	2	42

The False Positive Rate (FPR) for class C_1 is $\frac{5}{150}$, for class C_2 is $\frac{6}{150}$ and for class C_3 is $\frac{9}{100}$.

False.For class C_1 is $\frac{3}{3+42+2}$, for class C_2 is $\frac{6}{96+6+1}$ and for class C_3 is $\frac{9}{96+3+5}$

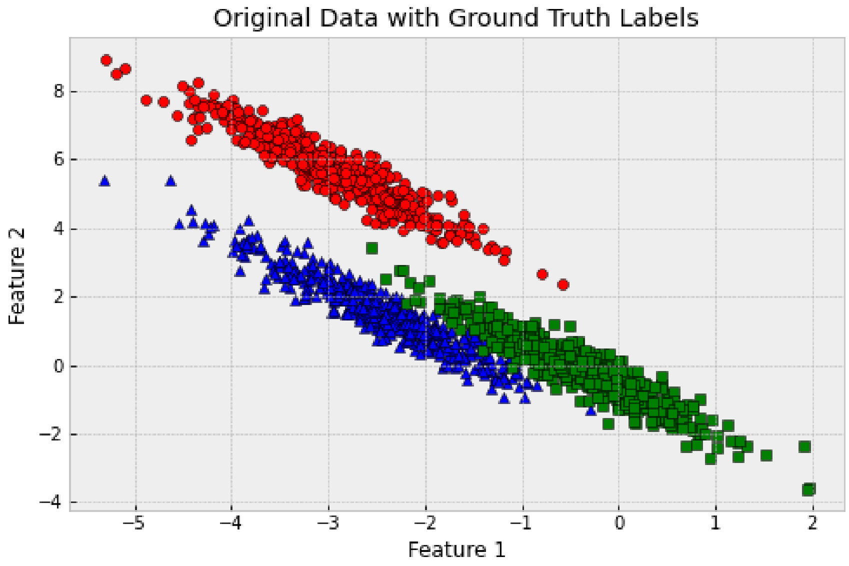
In [ ]:

1. (0.5 points) **Suppose that you will run three different clustering algorithms and hope to arrive at the ground truth labels shown in the figure above. Between the options (1) K-Means with Euclidean distance, (2) Gaussian Mixture Models with full covariance or (3) Gaussian Mixture Models with isotropic covariance matrix, option (1) is the best choice for this dataset.**

In [2]:

```
Image('3_class_dataset.png',width=500)
```

Out[2]:



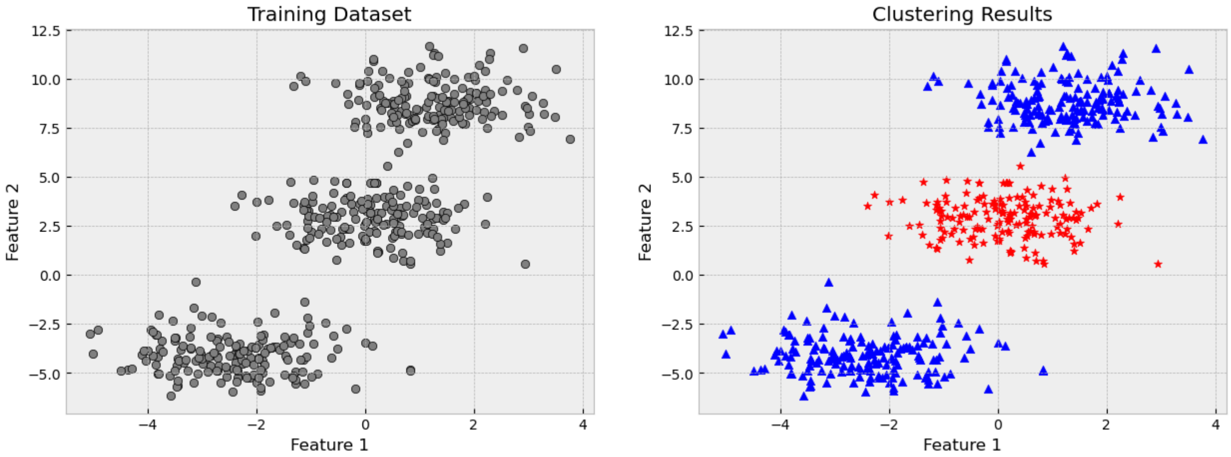
False. To learn the clusters in the figure the target algorithm should be able to learn elliptical-shaped clusters, which can only be implemented by option (2) because it uses full covariance.

```
In [ ]:
```

1. (0.5 points) **Consider the training dataset (left) and the clustering results with 2 clusters (right) depicted in the figure below. Both GMM and the standard K-Means clustering algorithms could have produced the clustering results for $K = 2$.**

```
In [3]: Image('clustering_results.png', width=900)
```

Out[3]:



False. Generally I think it is impossible for K-means to converge to a solution where the top cluster and the bottom cluster are merged into a single cluster, but it can happen depend on the initialization of cluster centroids. Here I can't tell with 100% confidence without the initialization and the number of E-M algorithm iterations.

```
In [ ]:
```

1. (0.5 points) **The K-Means clustering algorithm can handle clusters with varying densities, non-convex clusters and imbalanced cluster sizes.**

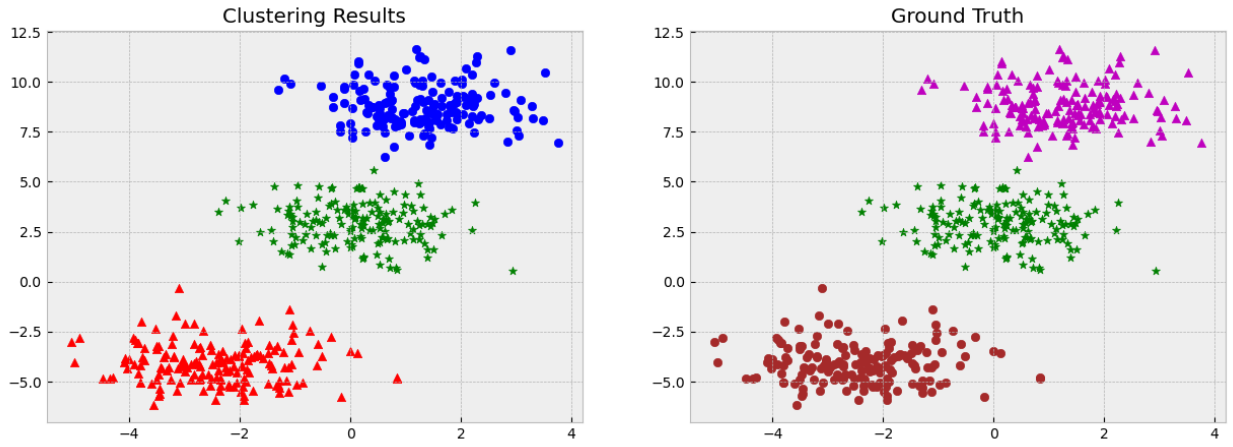
False. K-Means assumes that clusters have equal variance along all dimensions, which may not be appropriate for datasets with varying densities, and when coping with non-convex clusters it may incorrectly assign data points to the wrong cluster if the true clusters have complex shapes. DBSCAN and GMMs are more suitable for these tasks.

```
In [ ]:
```

1. (0.5 points) **The rand index of the clustering results (left) and the respective ground truth (right) as depicted below would be 1.**

```
In [4]: Image('clustering_ground_truth.png', width=900)
```

Out[4]:



True. The clustering results and the ground truth are identical.

In [ ]:

Exercise 3 (2 points)

Consider a centroid-based clustering algorithm with the following objective function:

$$J(\Theta, \mathbf{U}) = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik}^{m} d^2(x_i, \theta_k), \quad \text{such that } \ 0 \leq u_{ik} \leq 1$$

where $d^2(x_i, \theta_k)$ is the distance of sample x_i to cluster centroid θ_k, u_{ik} is the membership value of sample x_i in cluster with centroid θ_k in the range $[0, 1]$, and m is a scalar with $m > 1$.

1. (1 point) **Discuss whether we need to include the constraint $\sum_{k=1}^{K} u_{ik} = 1$. Explain why or why not?**

In K-Means algorithms, where each point is hard-assigned to a single cluster, the sum of membership values across clusters should be 1. However, for fuzzy clustering algorithms, each data point can belong to multiple clusters with varying degrees of membership, the normalization constraint may not be necessary.

In [ ]:

1. (1 point) **Propose a term to be added onto $J(\Theta, \mathbf{U})$ to ensure that the membership value for each sample x_i is close to 1 for at least one of the cluster groups k, while leaving outliers with low membership values.**

$$J(\Theta, \mathbf{U}) = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik}^{m} d^2(x_i, \theta_k) + \lambda \sum_{i=1}^{N} \|\mathbf{u}\|_1$$

In [ ]:

```
```

On-Time (1 point)

Submit your assignment before the deadline.

Submit Your Solution

Confirm that you've successfully completed the assignment.

Along with the Notebook, include a PDF of the notebook with your solutions.

`add` and `commit` the final version of your work, and `push` your code to your GitHub repository.

Submit the URL of your GitHub Repository as your assignment submission on Canvas.
