Homework 2 Part 1

This is an individual assignment.

Write your answers using markdown cells.

Exercise 1 (9 points)

Motivation:

> **Waste minimization is a critical aspect of sustainable production practices, particularly in the manufacturing of consumer goods. As the global demand for products continues to rise, so does the generation of waste, posing significant environmental and economic challenges. The production of consumer goods involves numerous stages, from raw material extraction to manufacturing, packaging, and distribution. At each step, there is the potential for waste generation, leading to resource depletion, pollution, and increased costs. Addressing this issue and implementing effective waste minimization strategies is essential to mitigate the environmental impact, conserve resources, and optimize production processes, all while maintaining profitability and meeting consumer expectations.**

Suppose your company is in the business of manufacturing consumer goods products and that you are interested in minimizing waste production. Instead of producing a large number of items at any given time, design a strategy to determine how many products should be manufactured at any given time to minimize waste production.

1. (2 points) **The dataset $\mathbf{X} = \{x_i\}_{i=1}^N$ is comprised of N samples x_i which represent the "number of products to be produced in a given interval of time". Which random variable (RV) best describes sample x_i? With that, define the observed data likelihood for the dataset.**

It can be described with Poission random variable:

$$L(\lambda | X) = \prod^{N} p(x_i | \lambda) = \prod^{N} \frac{e^{-\lambda} \lambda^{x}}{}$$

In [ ]:

```

```

2. (3 points) **Solve for the parameter of the observed data likelihood using the MLE approach.**

In [5]:

```
from IPython.display import Image
```

In [7]:
```
Image('1.png', width = 900)
```

Out[7]:

HW2

Ex1.Q2

We use Poisson Distribution for the likelihood:

$$p(x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$\mathcal{L}^{\circ} = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$\mathcal{L} = \ln \mathcal{L}^{\circ} = \sum_{i=1}^{N} \left(-\lambda + x_i \cdot \ln\lambda - \ln(x_i!)\right)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \Leftrightarrow \sum_{i=1}^{N}\left(-1 + \frac{x_i}{\lambda}\right) = 0$$

$$\lambda = \sum_{i=1}^{N} x_i$$

Q4

$$\mathcal{L}^{\circ} = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{x}}{x!}$$

prior: $P(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$

$$P(\lambda \mid X, \alpha, \beta) \propto \mathcal{L}^{\circ} \cdot \text{prior}$$

$$\ln P(\lambda \mid X, \alpha, \beta) \propto \sum_{i=1}^{N} \left(-\lambda + x_i \cdot \ln\lambda - \ln(x_i!)\right) + (\alpha-1)\cdot\ln\lambda - \beta\lambda$$

$$\frac{\partial \ln P(\lambda \mid X, \alpha, \beta)}{\partial \lambda} = 0 \Leftrightarrow$$

$$\sum_{i=1}^{N}\left(\frac{x_i}{\lambda} - 1\right) + \frac{\alpha-1}{\lambda} - \beta = 0 \Leftrightarrow \lambda = \frac{\sum_{i=1}^{N} x_i + \alpha - 1}{\beta + 1}$$

3. (1 point) **Can you introduce a prior probability for the parameter of the data likelihood such that it forms a conjugate prior relationship?**

Yes, we can use Gamma Distribution as prior so that the posterior will also be Gamma Distribution.

In [ ]:

```

```

4. (3 points) **Solve for the parameter of the observed data likelihood using the MAP approach for the prior probability on the parameter you selected in part 3.**

In [8]:

```python
Image('2.png', width = 900)
```

Out[8]:

Ex 2.

$$P(C_0|x) \propto P(x|C_0) \cdot P(C_0)$$

$$P(C_1|x) \propto P(x|C_1) \cdot P(C_1)$$

$$P(C_0) = \frac{100}{150} \qquad P(C_1) = \frac{50}{150}$$

$$P(C_0|x) \propto \frac{e^{-2} \cdot 2^3}{3!} \cdot \frac{1}{2} \cdot \frac{2}{3}$$

$$P(C_1|x) \propto \frac{10!}{3!\,7!}\, 0.5^3 \cdot 0.5^7 \cdot \frac{1}{2} \cdot \frac{1}{3}$$

$\because P(C_0|x) > P(C_1|x)$

$\therefore$ $x=3$ should assign to C_0

In [ ]:

```

```

Exercise 2 (8 points)

Consider a training set containing positive natural numbers - including zero - ($x \in \mathbb{N}_0$) for 2 classes, C_0 and C_1. The training set has 100 samples for class C_0 and 50 for C_1.

Suppose that you have reason to believe that samples belonging from C_0 are drawn from a Poisson random variable with parameter $\lambda > 0$, and samples belonging to C_1 are drawn from a Binomial random variable with parameters $n \in \mathbb{N}_0$ and $p \in [0, 1]$. In other words:

$$p(x|C_0) = \frac{\lambda^x e^{-\lambda}}{x!}$$
$$p(x|C_1) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient of "n choose x".

Consider $\lambda = 2$, $n = 10$ and $p = 0.5$. Consider the test point $x = 3$. Use the Naive Bayes classifier, to assign $x = 3$ to C_0 or C_1? Show your work.

In [10]:

```
Image('2.png', width = 900)
```

Out[10]:

下午11:02 10月1日周日 11%

5840no... Computer... computer-... 《计算机... Chapter 4... Chapter 5...

Ex 2.

$$P(C_0|x) \propto P(x|C_0)\cdot P(C_0)$$

$$P(C_1|x) \propto P(x|C_1)\cdot P(C_1)$$

$$P(C_0) = \frac{100}{150} \qquad P(C_1) = \frac{50}{150}$$

$$P(C_0|x) \propto \frac{e^{-2}\cdot 2^3}{3!}\cdot\frac{1}{2}\cdot\frac{2}{3}$$

$$P(C_1|x) \propto \frac{10!}{3!\,7!}\,0.5^3\cdot 0.5^7\cdot\frac{1}{2}\cdot\frac{1}{3}$$

$\because P(C_0|x) > P(C_1|x)$

$\therefore$ $x=3$ should assign to C_0

In [ ]:

In [ ]:

Exercise 3 (8 points)

Consider a Naive Bayes classifier for a binary classification task in a two-dimensional feature space with $P(\mathbf{x}|C_1) \sim G(\mu_1, \Sigma_1)$, $P(\mathbf{x}|C_2) \sim G(\mu_2, \Sigma_2)$ and $P(C_1) = P(C_2) = \frac{1}{2}$, where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is a sample observation, μ_i is the 2×1 mean vector for class C_i and Σ_i is the 2×2 covariance matrix for class C_i.

Let $g_i(x) = \ln(P(\mathbf{x}|C_i)P(C_i))$ be the discriminant function for class C_i. Then, let $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$. Using the Bayesian decision theory we find:

$$\begin{array}{lll} \textbf{Choose} & C_1 & \textbf{if} \;\; g(\mathbf{x}) > 0 \\ \textbf{Choose} & C_0 & \textbf{otherwise} \end{array}$$

1. (4 points) **Let $\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\Sigma_1 = \Sigma_2 = \mathbf{I}$. Find $g(\mathbf{x})$ and show all your work. What is the shape of the decision boundary?**.

In [11]:
```python
Image('3.png', width = 900)
```

Out[11]:

Ex3. Q1

$\because \Sigma_1 = \Sigma_2 = I \quad \therefore \sigma_1 = \sigma_2 = 1 \qquad d=2$

$$p(x|C_i) = (2\pi)^{-\frac{d}{2}} |\Sigma_i|^{-\frac{1}{2}} \operatorname{Exp}\left(-\frac{1}{2}(x_j-\mu_i)^T \Sigma_i^{-1}(x_j-\mu_i)\right)$$

$$g_1(x) = \ln(p(x|C_1)p(C_1))$$

$$= -\ln(2\pi) - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \cdot I \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \ln\left(\frac{1}{2}\right)$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}[x_1, x_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}(x_1^2 + x_2)^2$$

$$g_2(x) = -\ln(2\pi) - \frac{1}{2}[x_1-1, x_2-1]\begin{bmatrix} x_1-1 \\ x_2-1 \end{bmatrix} + \ln\left(\frac{1}{2}\right)$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}(x_1^2 + x_2^2 - 2x_1 - 2x_2 + 2)$$

$$g(x) = g_1(x) - g_2(x)$$

$$= \frac{1}{2}(-2x_1 - 2x_2 + 2)$$

$$= -x_1 - x_2 + 1$$

The decision boundry is a line.

In [ ]:

In [ ]:

2. (4 points) **Let** $\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}$ **and** $\Sigma_2 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$**. Find** $g(\mathbf{x})$ **and show all your work. What is the shape of the decision boundary?**.

In [12]: Image('4.png', width = 900)

Out[12]:

Q2

$$g_1(x) = -\ln(2\pi) - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}[x_1, x_2]\begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \ln\left(\frac{1}{2}\right)$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}\ln(3.75) - \frac{1}{2}(2x_1^2 + 2x_2^2 + x_1x_2)$$

$$g_2(x) = -\ln(2\pi) - \frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}[x_1-1, x_2-1]\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}\begin{bmatrix} x_1-1 \\ x_2-1 \end{bmatrix} + \ln\left(\frac{1}{2}\right)$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}\ln(9) - \frac{1}{2}[5x_1 - 5 + 4x_2 - 4,\ 4x_1 - 4 + 5x_2 - 5]\begin{bmatrix} x_1-1 \\ x_2-1 \end{bmatrix}$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}\ln(9) - \frac{1}{2}\left((5x_1 + 4x_2 - 9)(x_1 - 1) + (4x_1 + 5x_2 - 9)(x_2 - 1)\right)$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}\ln(9) - \frac{1}{2}(5x_1^2 + 4x_1x_2 - 9x_1 - 5x_1 - 4x_2 + 9 + 4x_1x_2 + 5x_2^2 - 9x_2 - 4x_1 - 5x_2 + 9)$$

$$= -\ln(2\pi) + \ln\left(\frac{1}{2}\right) - \frac{1}{2}\ln(9) - \frac{1}{2}(5x_1^2 + 5x_2^2 + 8x_1x_2 - 18x_1 - 18x_2 + 18)$$

$$g(x) = g_1(x) - g_2(x) = -\frac{1}{2}(-3x_1^2 - 3x_2^2 - 7x_1x_2 + 18x_1 + 18x_2 - 18)$$

It's a quadratic decision boundry in the 2-D feature space.

In [ ]:

```

```

In [ ]:

```

```

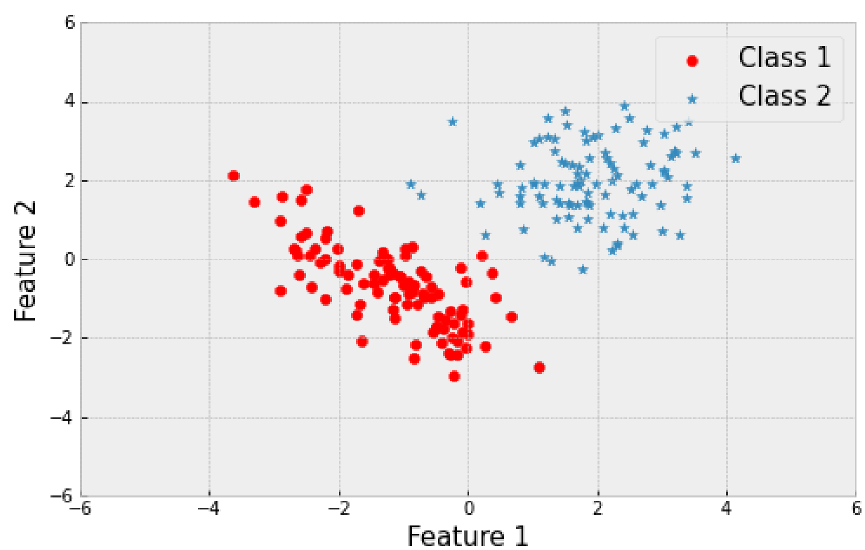
Exercise 4 (8 points)

Consider the a dataset composed of two classes (class 1 and class 2) in a 2-dimensional feature space. Each class has 100 samples. This is what the data looks like:

In [1]:

```
from IPython.display import Image
Image('figures/Dataset.png',width=500)
```

Out[1]:

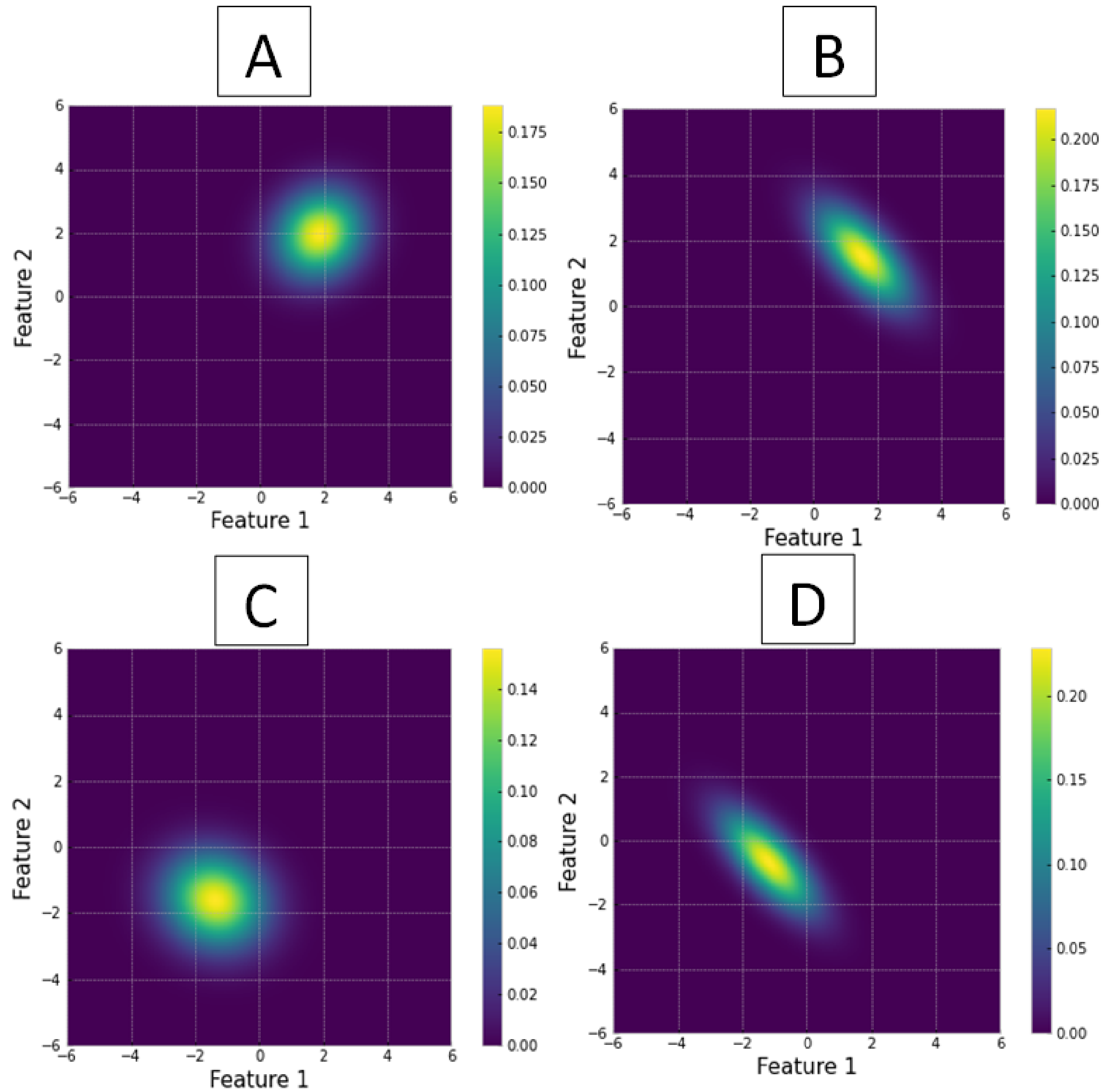


Answer the following questions:

1. (2 points) **The four plots below (A, B, C and D) represent contours plots of the data likelihood as modeled with a Gaussian distribution. from the four plots, select one that represents the data of class 1, and another that represents the data in class 2. Justify your answer.**

In [2]:
```
Image('figures/DataLikelihoods.png',width=600)
```

Out[2]:



Plot D represents class 1 and Plat A represents class 2. To get the answer, first estimate the central of each class, then compare its shape with what in the plot below.

In [ ]:
```

```

2. (2 points) **Suppose we want to implement the Naive Bayes Classifier on this training data. What is the prior probability for each class?**

It depends on how many samples each class have. Here both class 1 and class 2 have 100 samples so by way of relative frequency approach we know $p(C_1) = p(C_2) = \frac{1}{2}$

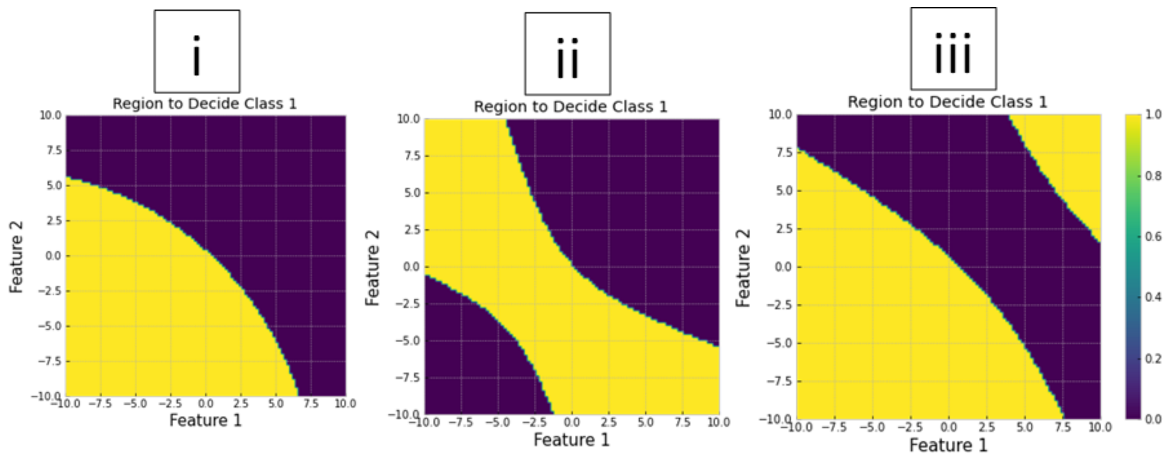
In [ ]:

3. (2 points) **Suppose that we implemented the Naive Bayes Classifier using the prior probabilities you computed in part (2). The three plots below (i, ii and iii) represent the decision surface for deciding class 1 (red class). Based on the plotting of the data, the data likelihoods selected from part (1) and the priors from part (2), which of the following plots corresponds to the true decision surface for deciding class 1? Justify your answer.**

In [3]:

```python
Image('figures/DecisionSurfaces.png',width=900)
```

Out[3]:



Plot i. Only plot i have 2 seperated regions, which suits the discriminant function(it won't be so complex because we only have 2 classes and both of them follow Gaussian Distribution).

In [ ]:

4. (2 points) **Using the Naive Bayes decision surface you selected in part (3), which class are the test points $x_1 = [-1, -1]$ and $x_2 = [-10, 10]$ going to be assigned as? How confident would you be in that decision? Justify your answer.**

x1: class 1 x2: class 2. I'm about 100% confident with the decisions based on the colors of the plot. (value of probability density can be refered on the right side of the plots).

In [ ]:

Exercise 5 (12 points)

For each statement below, state whether it is true (T) or false (F) and provide a justification for your answer.

1. (1 point) **Minimizing the Least Squares objective function with a weight decay term (regularizer) is equivalent to maximizing the data likelihood with a prior on the weights.**

False. It is equivalent to maximizing the data likelihood with a Gaussian prior on the weights.

In [ ]:

2. (1 point) **The observed data likelihood on N i.i.d. samples x_i (where $x_i \geq 0, \forall i$) with a Gamma distribution, $P(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, and a prior probability on parameter β as another Gamma distribution, $P(\beta|a, b) = \frac{b^a}{\Gamma(a)} \beta^{a-1} e^{-b\beta}$, form a Conjugate Prior relationship.**

True. They form a Conjugate Prior relationship because the prior and the likelihood are in a same category.

In [ ]:

3. (1 point) **In MLE estimation we are optimizing the posterior distribution.**

False. MLE's goal is to optimize the likelihood function.

In [ ]:

4. (1 point) **Consider $\theta_{\text{MLE}} = \arg_\theta \max P(X|\theta)$ and $\theta_{\text{MAP}} = \arg_\theta \max P(X|\theta)P(\theta)$. We have that $\theta_{\text{MLE}} = \theta_{\text{MAP}}$ for uniform-distributed priors.**

True. When $P(\theta)$ is uniform-distributed, it is equivalent to a constan so $\theta_{MLE} = \theta_{MAP}$.

In [ ]:

5. (1 point) **The EM algorithm guarantees convergence to the global optima solution.**

False. EM is used to find the maximum likelihood estimates of parameters in models with latent variables or missing data.

In [ ]:

6. (1 point) **The latent variables z in the EM algorithm are discrete random variables.**

True. Latent variables in EM are used to represent the missing information in the data.

In [ ]:

7. (1 point) **We can use the EM algorithm to optimize a Gamma mixture model, $P(x) = \sum_{k=1}^{K} \pi_k \Gamma(x|\nu_k, \theta_k)$ to a set of N i.i.d. samples $\{x_i\}_{i=1}^{N}$.**

True. EM is commonly applied to mixture models, including Gamma mixture models.

In [ ]:

8. (1 point) **The MLE solution for the parameter θ_k in the Gamma mixture model is $\theta_k = \frac{\nu_k \sum_{i=1}^{N} C_{ik}}{\sum_{i=1}^{N} x_i C_{ik}}$.**

In [ ]:
```
-----
```

In [ ]:

9. (1 point) **A Gaussian Mixture Model can be used to find cluster groups in the data.**

True. GMMs are widely used for clustering applications.

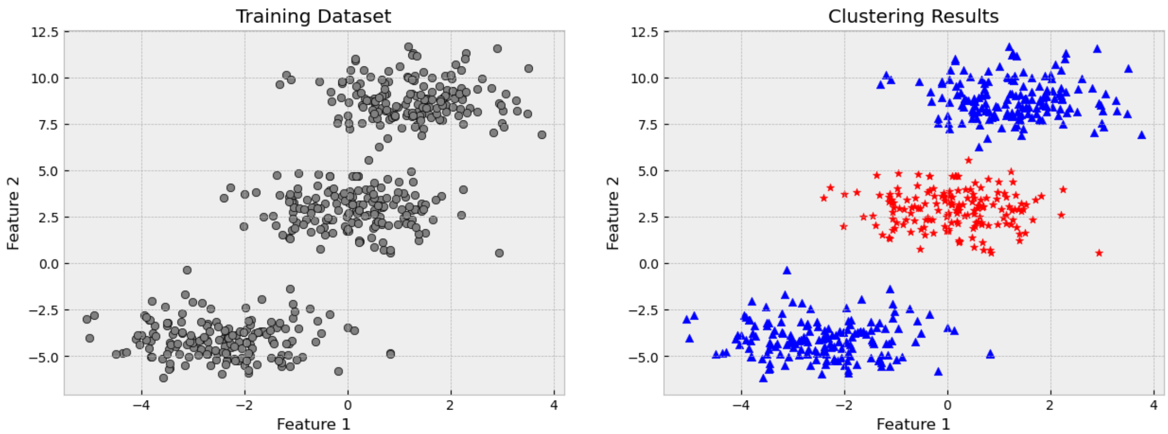
In [ ]:

10. (1 point) **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 [4]:

```
Image('figures/clustering_results.png', width=900)
```

Out[4]:



In [ ]:

```
-----
```

In [ ]:

```
```

11. (1 point) **The K-Means clustering algorithm can handle clusters with varying densities, non-convex clusters and imbalanced cluster sizes.**

In [ ]:

```
-----
```

In [ ]:

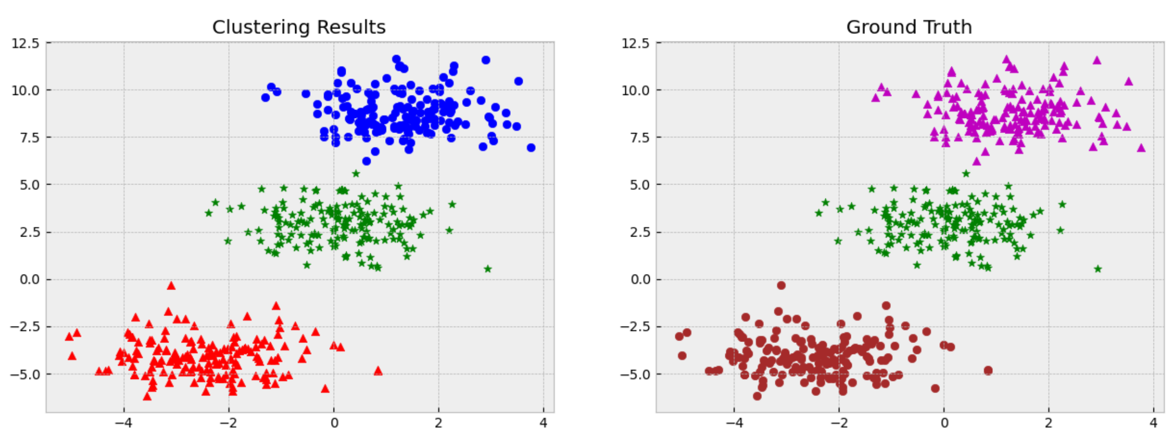
```
```

12. (1 point) **The rand index of the clustering results (left) and the respective ground truth (right) as depicted below would be 1.**

In [5]:

```
Image('figures/clustering_ground_truth.png', width=900)
```

Out[5]:



In [ ]:
```
-----
```

In [ ]:
```

```

On-Time (5 points)

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.
