Question 0 - The concept that I fully understand and enjoyed is k means clustering :

K-means is an unsupervised learning algorithm used to group data points into K clusters based on their similarities. It's a popular technique for tasks like customer segmentation, image recognition, and anomaly detection. The algorithm's goal is to minimize the within-cluster sum of squares (WCSS), which measures the distance between each data point and the centroid of its assigned cluster.In K-means clustering, data points are grouped into clusters based on their similarity. It is a type of centroid-based clustering, where each cluster is represented by a centroid. The goal is to minimize the within-cluster sum of squares (WCSS) by assigning data points to the nearest centroid.

The difference between intra-cluster and inter-cluster refers to the distances within a cluster and between different clusters, respectively. Intra-cluster distance measures the similarity or compactness of data points within the same cluster. Inter-cluster distance measures the dissimilarity or separation between different clusters. In K-means, the algorithm aims to minimize the intra-cluster distance while maximizing the inter-cluster distance to create well-separated and distinct clusters.

Here's a step-by-step breakdown of the K-means algorithm, including the mathematical expressions involved:

1. Initialization:

- Randomly select K data points from the dataset as initial cluster centroids.

- Let's denote these centroids as μ₁, μ₂, ..., μₖ.

2. Assignment:

- For each data point xᵢ in the dataset, calculate the Euclidean distance to each centroid.

- Assign xᵢ to the cluster with the nearest centroid.

- Let's denote the cluster assignment for xᵢ as cᵢ.

3. Update:

- After assigning all data points to clusters, update the centroids by calculating the mean of the data points in each cluster.

- The updated centroids are denoted as μ₁', μ₂', ..., μₖ'.

4. Iteration:

- Repeat steps 2 and 3 until convergence.

- Convergence occurs when the centroids no longer change significantly or a maximum number of iterations is reached.

The mathematical expressions used in K-means are as follows:

- Euclidean Distance:

- The Euclidean distance between two data points xᵢ and xⱼ is calculated as:

dist(xᵢ, xⱼ) = √(Σ(xᵢₖ - xⱼₖ)²), where k represents the dimensions of the data points.

- Within-Cluster Sum of Squares (WCSS):

- WCSS measures the sum of squared distances between each data point and its assigned centroid within a cluster.

- WCSS for a cluster cₖ is calculated as:

WCSS(cₖ) = Σ(dist(xᵢ, μₖ)²), where xᵢ belongs to cluster cₖ.

- Total WCSS:

- The total WCSS for all clusters is calculated as:

Total WCSS = Σ(WCSS(cₖ)).

Limitations of clustering are given as follows:

1 It requires the user to specify the number of clusters in advance.

2 It may not work well with complex or overlapping data patterns.

The concept which is unclear to me is Naives Bayes.

Reference – “Hands on Machine learning with scikit - learn and Tenserflow” by Aurelien Geron