03603452 Data Mining

Unsupervised Learning

Asst.Prof.Dr.Kulwadee Somboonviwat

Department of Computer Engineering

Faculty of Engineering at Sriracha

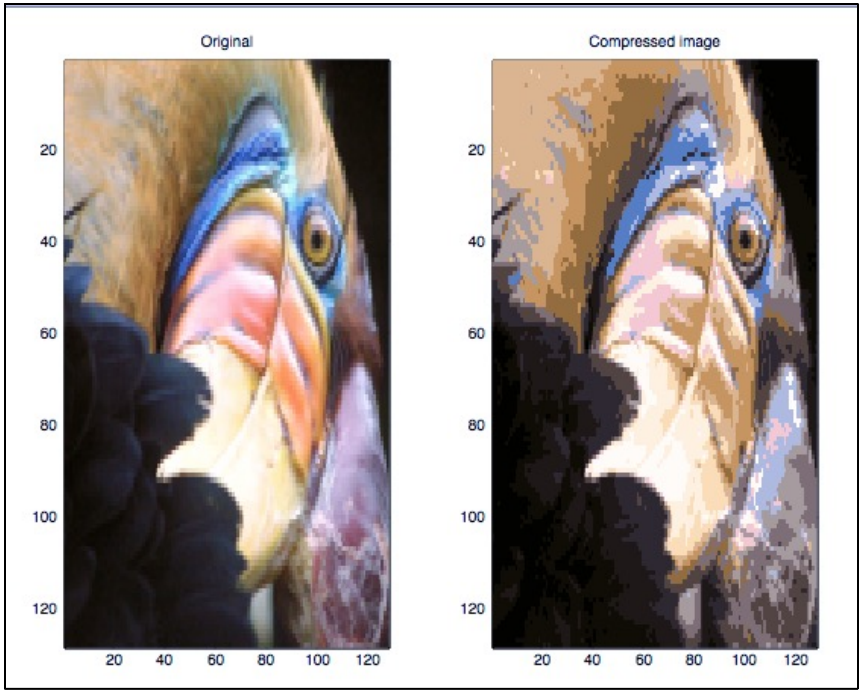
Kasetsart University Sriracha Campus

Semester 2/2565

Clustering

- Given a set of data points, group similar data points together
- Unsupervised learning: no predefined classes
- Why Clustering?
  - Basic assumption when doing data analytics (similar data points have similar behavior)
  - Applications: biology, searching, marketing, city planning, compression, outlier detection, etc.
- In this lecture, I will discuss two algorithms for clustering: k-Means and GMM

Data Compression with K-Means



K-means groups similar colors together into k=64 clusters.

Original image: each pixel was taking 24 (8X3) bits to store its corresponding color vector.

Compressed image: each pixel only takes 6 bits

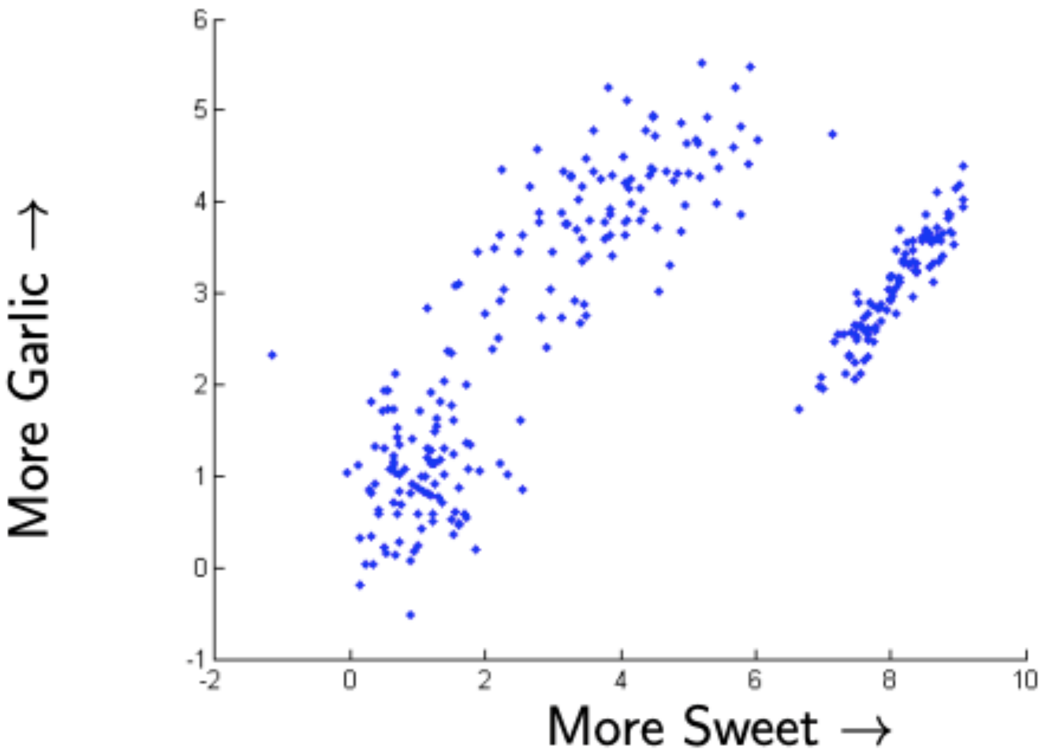
[Source: https://medium.com/@agarwalvibhor84/image-compression-using-k-means-clustering-8c0ec055103f](https://medium.com/@agarwalvibhor84/image-compression-using-k-means-clustering-8c0ec055103f)

Using K-Means to find a new tomato sacues

- A major tomato sauce company wants to tailer their sauces to suit the customers
- The company runs a market survey asking the test subject to rate different sauces
- The company gets the following data

* This example was taken from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

Tomato Sauce – Data



Each data point represents the preferred sauce characteristics of a specific customer

* This example was taken from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

Tomato Sauce – Questions

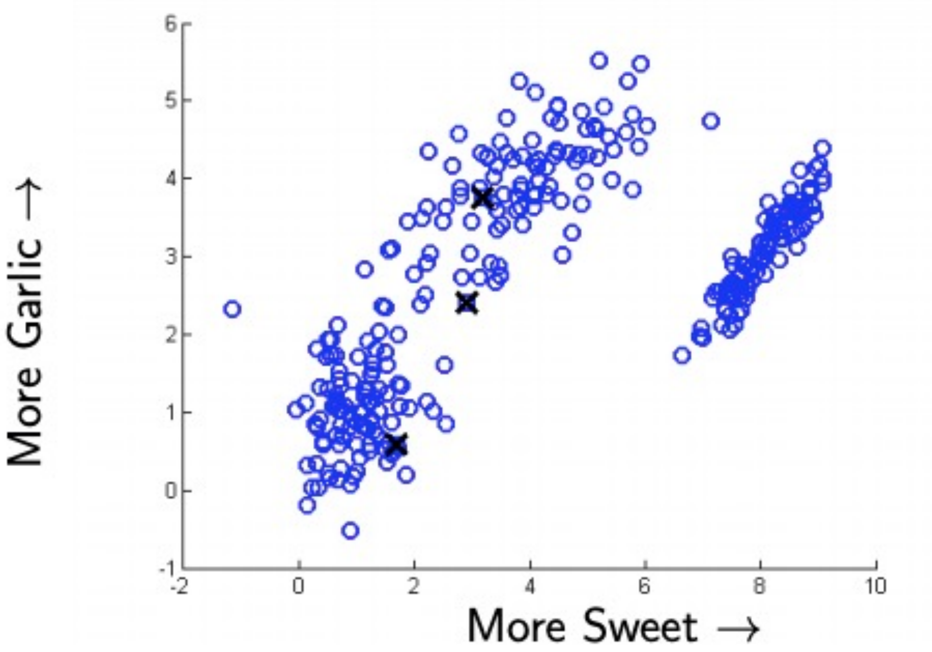
- How many different sauces should the company make?
- How sweet/garlicy should these sauces be?

Idea: Segment the customers into 3 groups, then find the best sauce for each group

* This example was taken from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

K-Means – Tomato Sauce

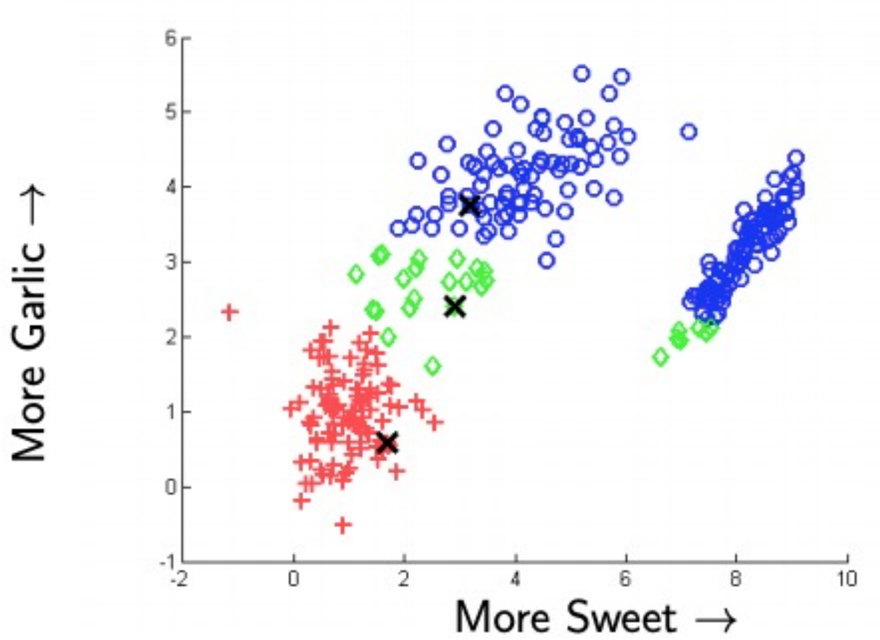
Given 3 sauces whose garlicy-ness and sweetness are marked by X



* This example was taken from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

K-Means – Tomato Sauce

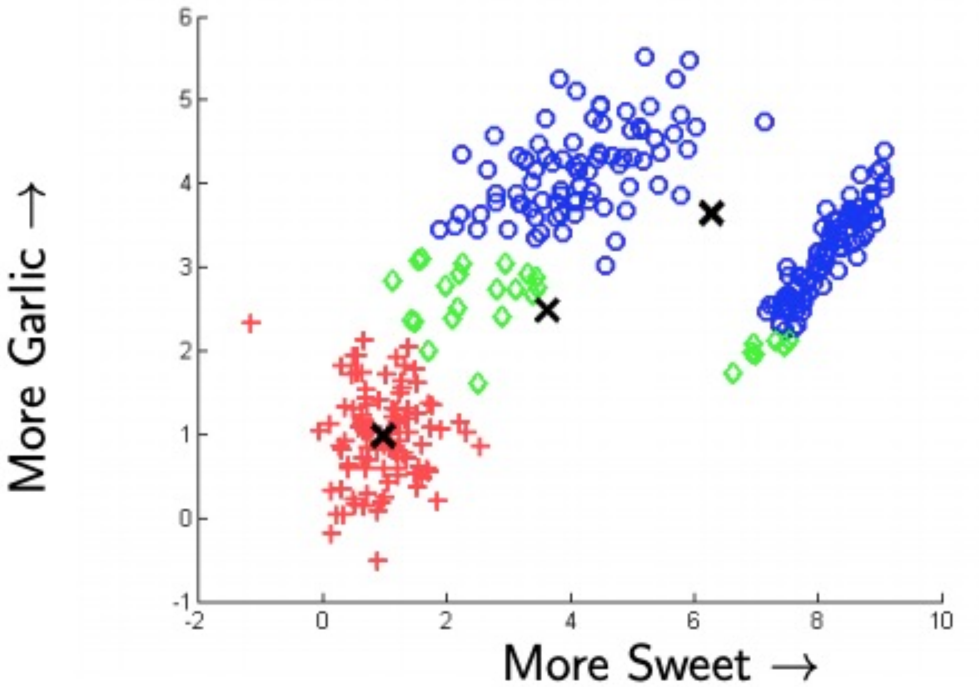
Group each customer by the sauce that most closely match his/her taste.



* This example was taken from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

K-Means – Tomato Sauce

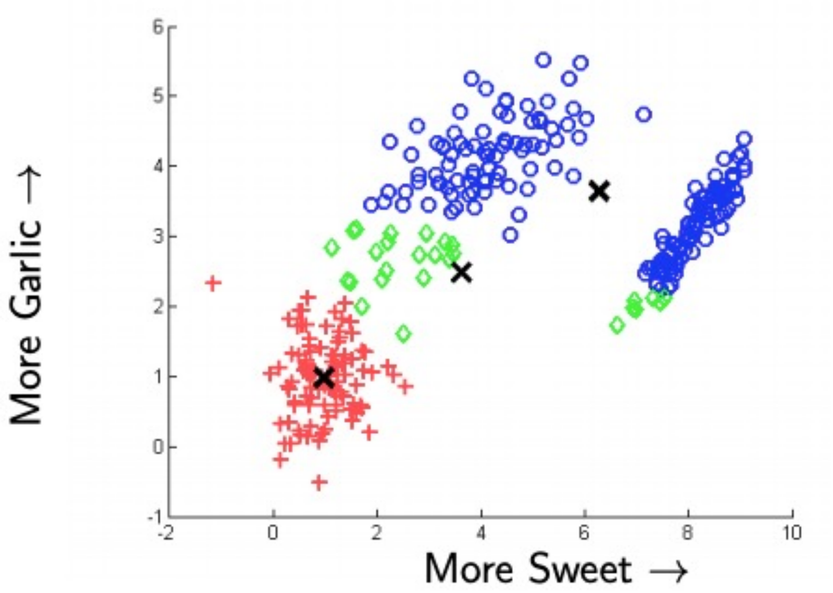
Choose a new sauce (location of X) that would make each group happier.



* This example was taken from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

K-Means – Tomato Sauce

Repeat the process until there is no need to change the characteristics of the new sauces



* This example was taken from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

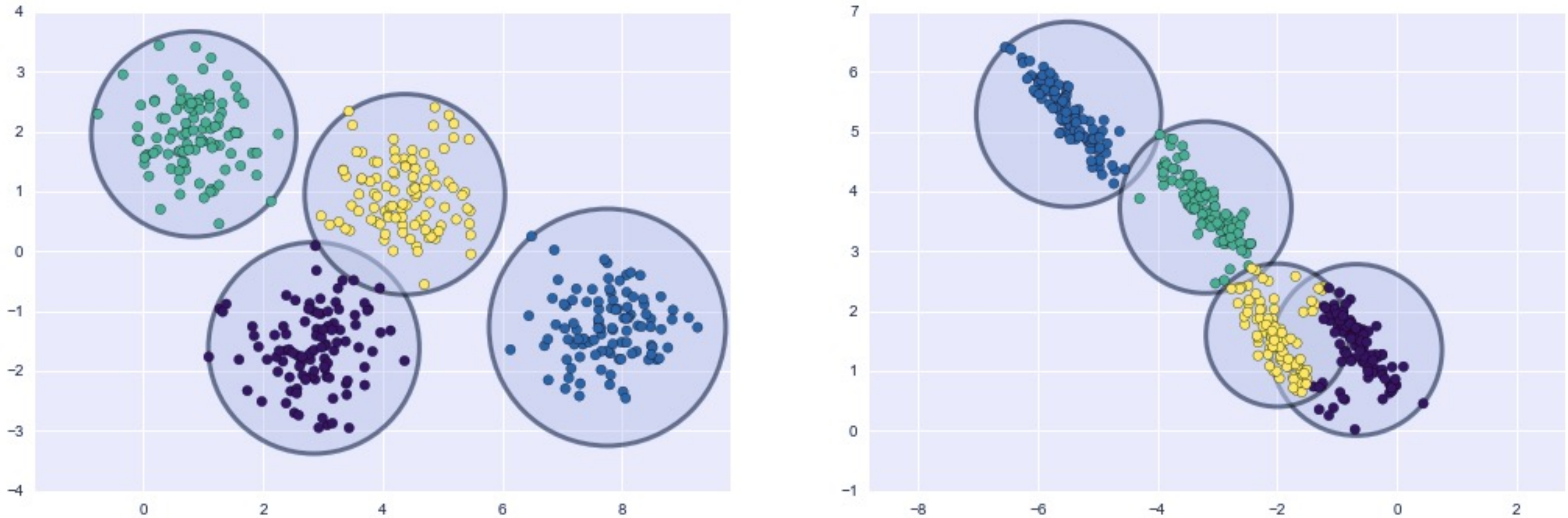
K-Means Algorithms

- Initialization: Choose k random points to act as a cluster centers
- Iterate until convergence:
  - Step 1: Assign points to closest center (forming k groups)
  - Step 2: Reset the centers to be the mean of the points respective groups

K-Means Challenges

- How to choose k?
- Which distance measure to use?
- K-means only find a **local** optimum
- K-means assumes data to be in a <u>spherical shape</u>

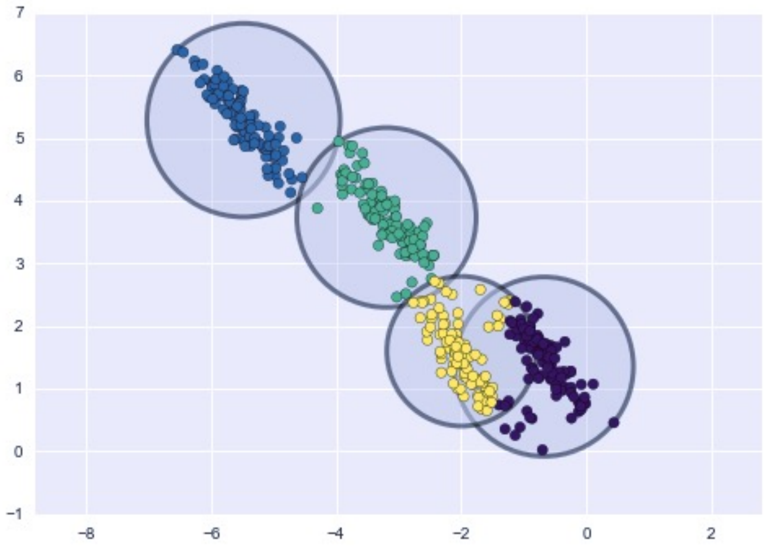
K-Means works well when data is circular



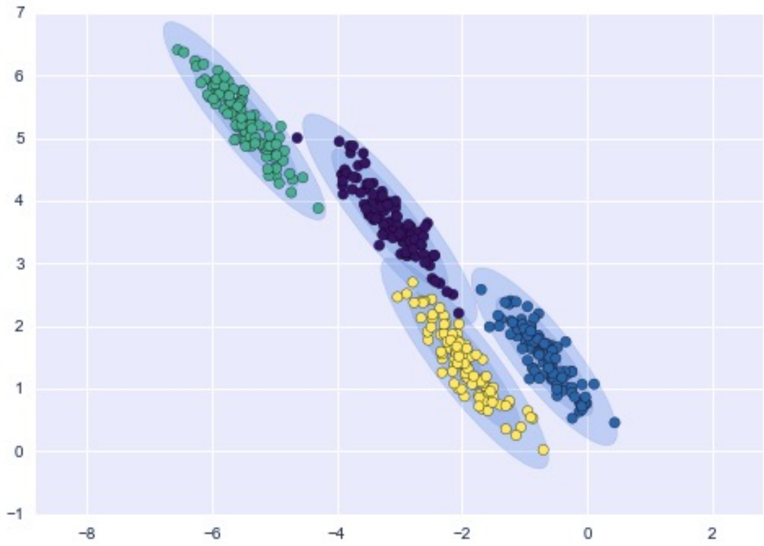
Source: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

Gaussian Mixture Model

K-Means

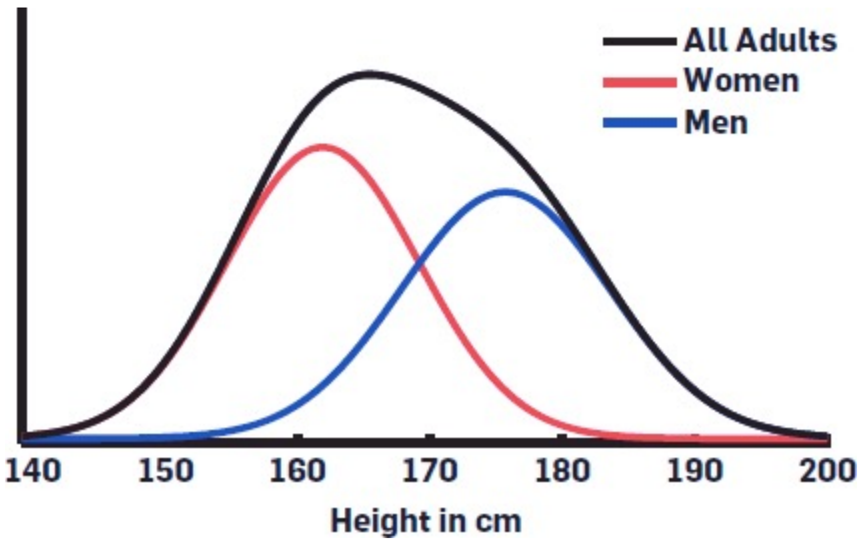


Gaussian Mixture Model



Source: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

Gaussian Mixture



Given only the height data and not the gender assignments for each data point, the distribution of all heights would follow the sum of two scaled (different variance) and shifted (different mean) normal distributions.

https://cacm.acm.org/magazines/2012/2/145412-disentangling-gaussians/fulltext

Gaussian Mixture Model

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}(x \mid \mu_i, \sigma_i)$$

$$\mathcal{N}(x \mid \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

$$\sum_{i=1}^{K} \phi_i = 1$$

A Gaussian mixture model is parameterized by two types of values, the mixture **component weights** and the component **means** and **variances**.

For a Gaussian mixture model with K components,the k-th component has a mean of μ_k and variance of σ_k

https://brilliant.org/wiki/gaussian-mixture-model/

Learning the Gaussian Mixture Model

EM (Expectation Maximization) for Gaussian Mixture Models

Step 1: (**Expectation** step or **E** step)
Calculate the expectationof the component assignments C_k for each data point $x_i \in X$, given the model parameters ϕ_k, μ_k, and σ_k.

Step 2: (**Maximization** step or **M** step)
Update the values of the model parameters ϕ_k, μ_k, and σ_k., given the current component assignments C_k

https://brilliant.org/wiki/gaussian-mixture-model/

Learning the Gaussian Mixture Model

Intialization Step

- Randomly assign samples without replacement from the dataset $X = \{x_1, ..., x_N\}$ to the component mean estimates $\hat{\mu}_1, ..., \hat{\mu}_K$. E.g. for $K = 3$ and $N = 100$, set $\hat{\mu}_1 = x_{45}, \hat{\mu}_2 = x_{32}, \hat{\mu}_3 = x_{10}$.
- Set all component variance estimates to the sample variance $\hat{\sigma}_1^2, ..., \hat{\sigma}_K^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$, where $\bar{x}$ is the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.
- Set all component distribution prior estimates to the uniform distribution $\hat{\phi}_1, ..., \hat{\phi}_K = \frac{1}{K}$.

https://brilliant.org/wiki/gaussian-mixture-model/

Learning the Gaussian Mixture Model

Expectation Step

Calculate $\forall i, k$

$$\hat{\gamma}_{ik} = \frac{\hat{\phi}_k \mathcal{N}(x_i \mid \hat{\mu}_k, \hat{\sigma}_k)}{\sum_{j=1}^{K} \hat{\phi}_j \mathcal{N}(x_i \mid \hat{\mu}_j, \hat{\sigma}_j)},$$

where $\hat{\gamma}_{ik}$ is the probability that x_i is generated by component C_k. Thus, $\hat{\gamma}_{ik} = p(C_k | x_i, \hat{\phi}, \hat{\mu}, \hat{\sigma})$.

https://brilliant.org/wiki/gaussian-mixture-model/

Learning the Gaussian Mixture Model

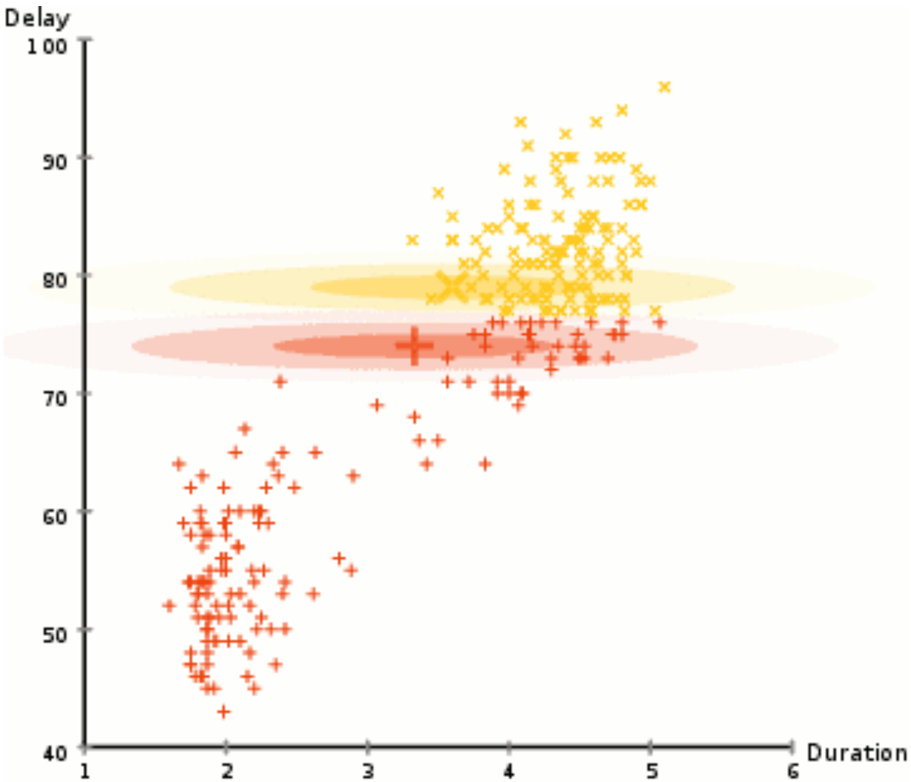
Maximization Step

Using the $\hat{\gamma}_{ik}$ calculated in the expectation step, calculate the following in that order $\forall k$:

- $\hat{\phi}_k = \sum_{i=1}^{N} \frac{\hat{\gamma}_{ik}}{N}$
- $\hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} x_i}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}$
- $\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} (x_i - \hat{\mu}_k)^2}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}.$

https://brilliant.org/wiki/gaussian-mixture-model/

Learning the Gaussian Mixture Model



https://brilliant.org/wiki/gaussian-mixture-model/

The EM algorithm updating the parameters of a two-component bivariate Gaussian mixture model