Notation

1D Gaussian PDF: $f(x) = \mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Standard normal distribution

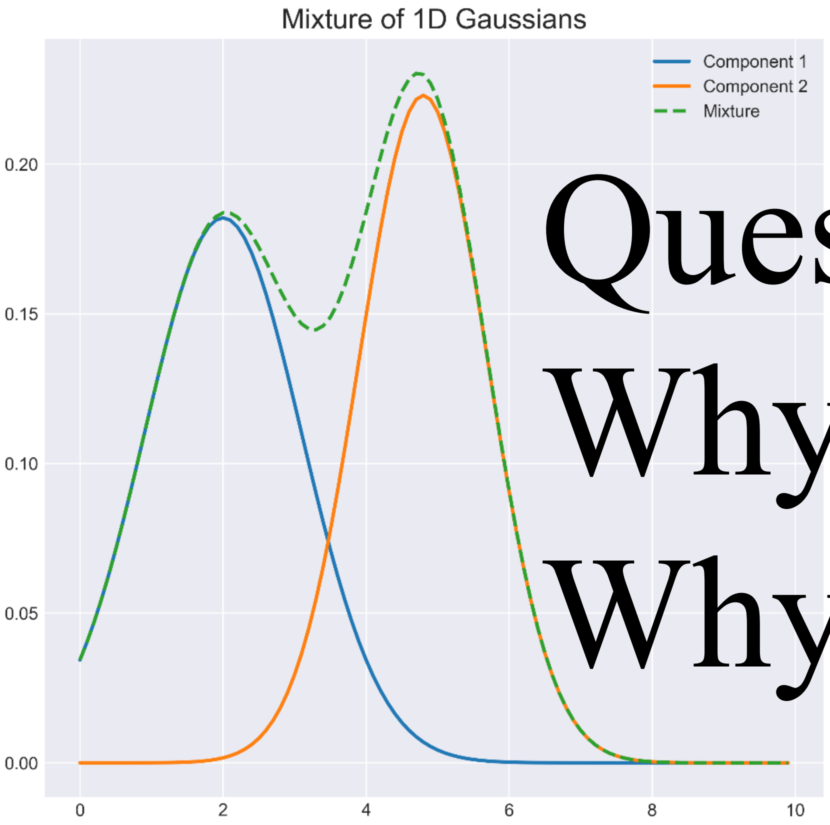
$$\mathcal{N}(x|0,1) = \mathcal{N}(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) dx = 1$$

1D Gaussian Mixture Model

- Mixture of two 1D Gaussians

$$f(x) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(x|\mu_2, \sigma_2^2)$$



Question:

Why do we need π_1 and π_2 ?

Why $\pi_1 + \pi_2 = 1$?

https://angusturner.github.io/generative_models/2017/11/03/pytorch-gaussian-mixture-model.html

A Property of PDF: $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \left[\boldsymbol{\pi}_1 \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2\right) + \boldsymbol{\pi}_2 \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2\right)\right] dx$$

$$= \boldsymbol{\pi}_1 \int_{-\infty}^{\infty} \left[\mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2\right)\right] dx + \boldsymbol{\pi}_2 \int_{-\infty}^{\infty} \left[\mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2\right)\right] dx$$

$$= \boldsymbol{\pi}_1 + \boldsymbol{\pi}_2$$

What is $\gamma_{(n,k)}$ when there are two Gaussians in a GMM ?

$$\gamma_{(n,1)} = \frac{\boldsymbol{\pi_1} \mathcal{N}\left(\boldsymbol{x_n} | \boldsymbol{\mu_1}, \boldsymbol{\sigma_1^2}\right)}{\boldsymbol{\pi_1} \mathcal{N}\left(\boldsymbol{x_n} | \boldsymbol{\mu_1}, \boldsymbol{\sigma_1^2}\right) + \boldsymbol{\pi_2} \mathcal{N}\left(\boldsymbol{x_n} | \boldsymbol{\mu_2}, \boldsymbol{\sigma_2^2}\right)}$$

$$\gamma_{(n,2)} = \frac{\boldsymbol{\pi_2} \mathcal{N}\left(\boldsymbol{x_n} | \boldsymbol{\mu_2}, \boldsymbol{\sigma_2^2}\right)}{\boldsymbol{\pi_1} \mathcal{N}\left(\boldsymbol{x_n} | \boldsymbol{\mu_1}, \boldsymbol{\sigma_1^2}\right) + \boldsymbol{\pi_2} \mathcal{N}\left(\boldsymbol{x_n} | \boldsymbol{\mu_2}, \boldsymbol{\sigma_2^2}\right)}$$

$$\gamma_{(n,1)} + \gamma_{(n,2)} = 1$$

What are the E-step and M-step for this1D GMM ?

- E-step:

Given $\boldsymbol{\pi_1}, \boldsymbol{\pi_2}\ \boldsymbol{\mu_1}, \boldsymbol{\sigma_1^2}, \boldsymbol{\mu_2}, \boldsymbol{\sigma_2^2}$, calculate $\boldsymbol{\gamma_{(n,1)}}$ and $\boldsymbol{\gamma_{(n,2)}}$ for each data point x_n

- M-step:

Given $\boldsymbol{\gamma_{(n,1)}}$ and $\boldsymbol{\gamma_{(n,2)}}$, calculate $\boldsymbol{\pi_1}, \boldsymbol{\pi_2}\ \boldsymbol{\mu_1}, \boldsymbol{\sigma_1^2}, \boldsymbol{\mu_2}, \boldsymbol{\sigma_2^2}$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{(n,k)} x_n$$

$$\sigma_k^2 = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{(n,k)} (x_n - \mu_k)^2$$

$$\pi_k = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N} \gamma_{(n,k)}$

How can we do clustering using $\gamma_{(n,k)}$?

- If there are two clusters, for the data point x_n,

  if $\gamma_{(n,1)} > \gamma_{(n,2)}$, then x_n is assigned to cluster-1

  if $\gamma_{(n,1)} < \gamma_{(n,2)}$, then x_n is assigned to cluster-2

- $\gamma_{(n,k)}$ is the probability of x_n belonging to cluster-k
- For each data point x_n, there is a probability distribution over the K clusters

  $[\gamma_{(n,1)}, \gamma_{(n,2)}, \gamma_{(n,3)}, \cdots, \gamma_{(n,k)}, \cdots, \gamma_{(n,K)}]$

  $\gamma_{(n,1)} + \gamma_{(n,2)} + \gamma_{(n,3)} + \cdots + \gamma_{(n,k)} + \cdots + \gamma_{(n,K)} = 1$

  $\gamma_{(n,k)}$ is also called membership of x_n in cluster-k