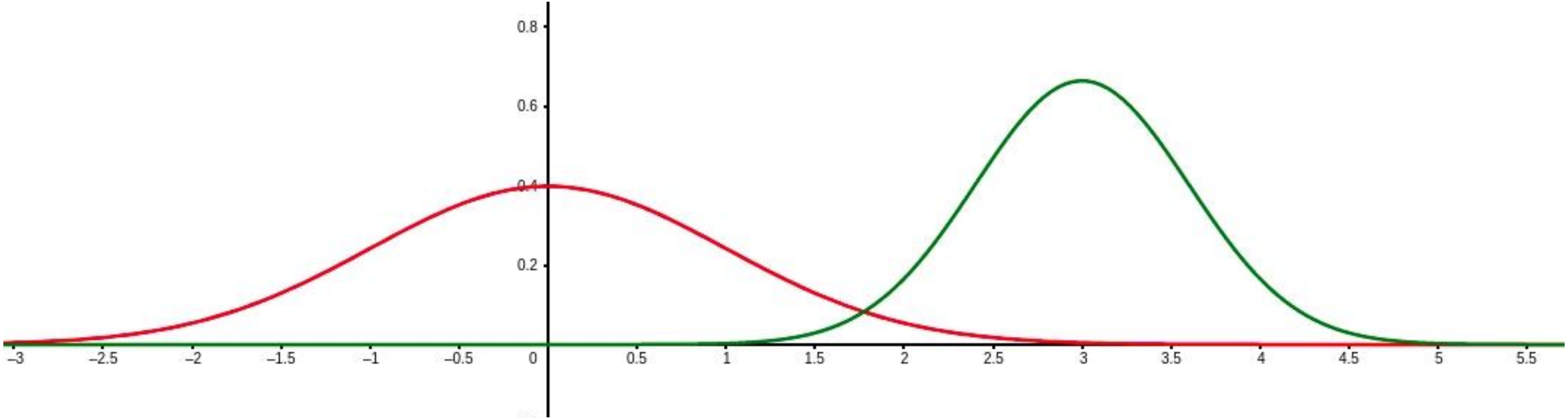
Soft Clustering

Boston University CS 506 - Lance Galletti

Soft Clustering - Example

Generate data where $P(C_1) = P(C_2) = ½$ and within C_1 and C_2 the distributions are $\mathbf{N(\mu_1, \sigma_1)}$ and $\mathbf{N(\mu_2, \sigma_2)}$



Ex: we are given the weights of animals. Unknown to us these are weights from two different species. Can we determine the species (group / assignment) from the height?

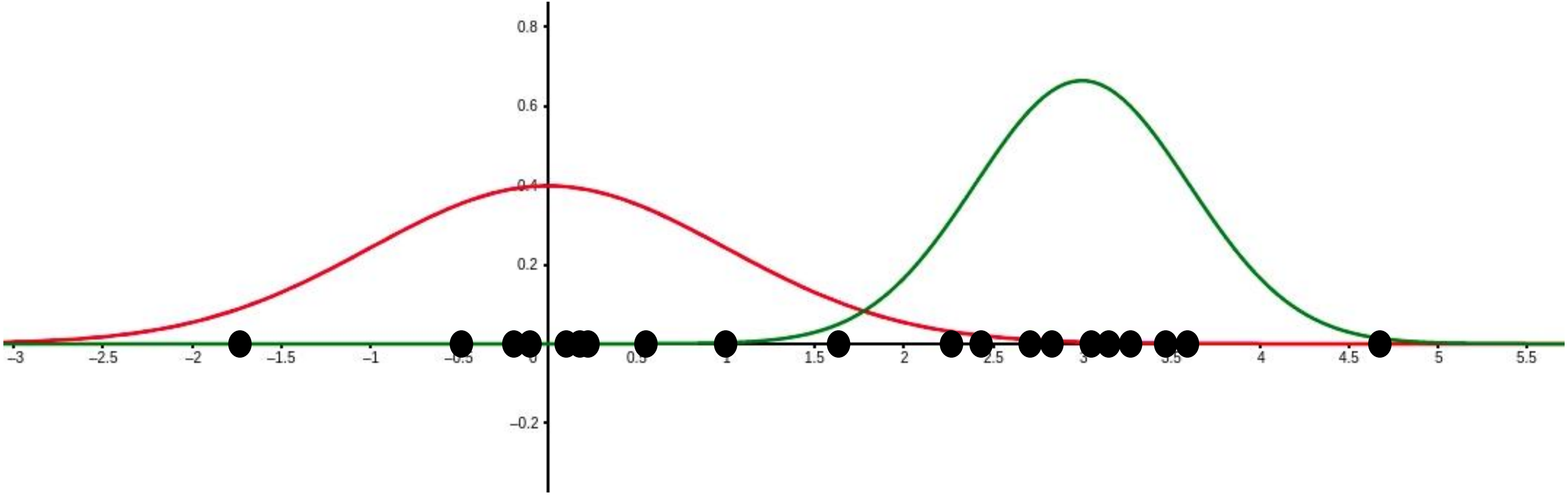
Soft Clustering - Example

Things to consider:

1. There is a prior probability of being one species (i.e. we could have an imbalanced dataset or there could just be more of one species than the other)
2. Weights within a particular group / species follow a particular distribution

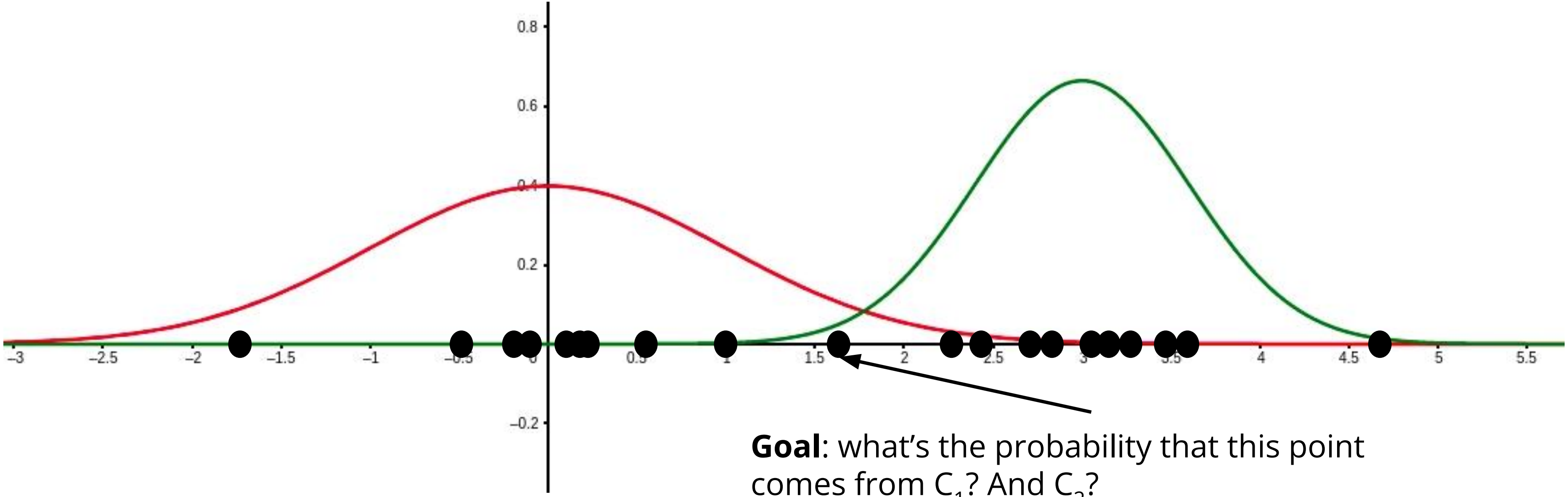
Soft Clustering - Example

Generate data where $P(C_1) = P(C_2) = ½$ and within C_1 and C_2 the weight distributions are $\mathbf{N(\mu_1, \sigma_1)}$ and $\mathbf{N(\mu_2, \sigma_2)}$



Soft Clustering - Example

Generate data where $P(C_1) = P(C_2) = ½$ and within C_1 and C_2 the weight distributions are $\mathbf{N(\mu_1, \sigma_1)}$ and $\mathbf{N(\mu_2, \sigma_2)}$

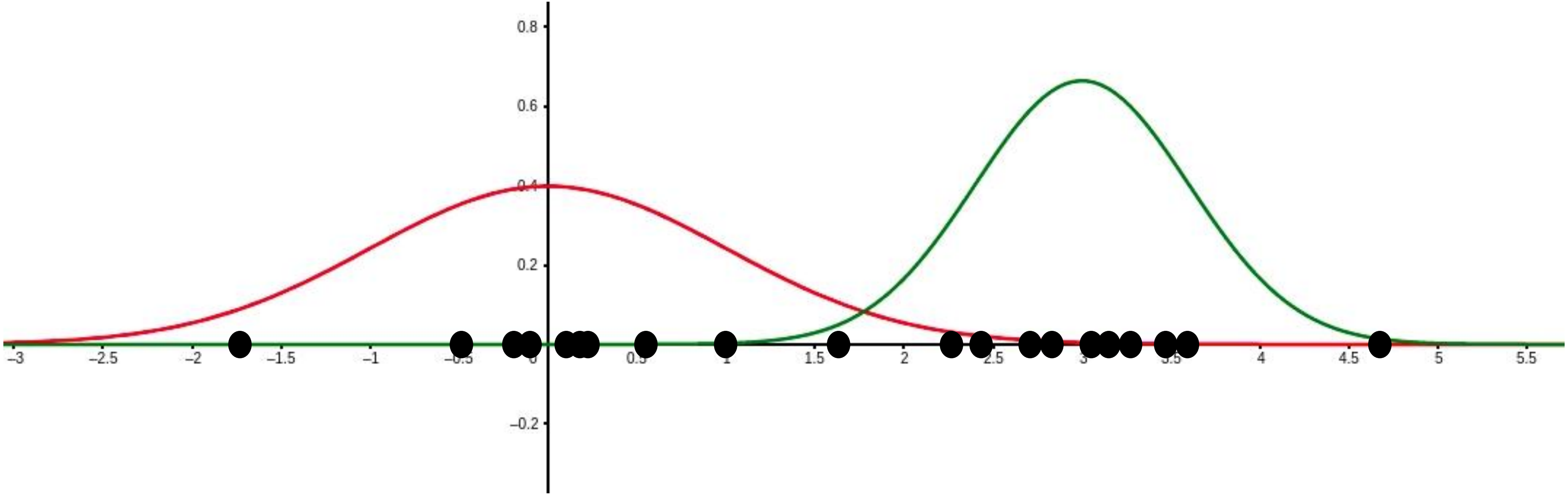


Goal: what's the probability that this point comes from C_1? And C_2?

$\mathbf{P(C_1 \mid X_i)}$

Soft Clustering - Example

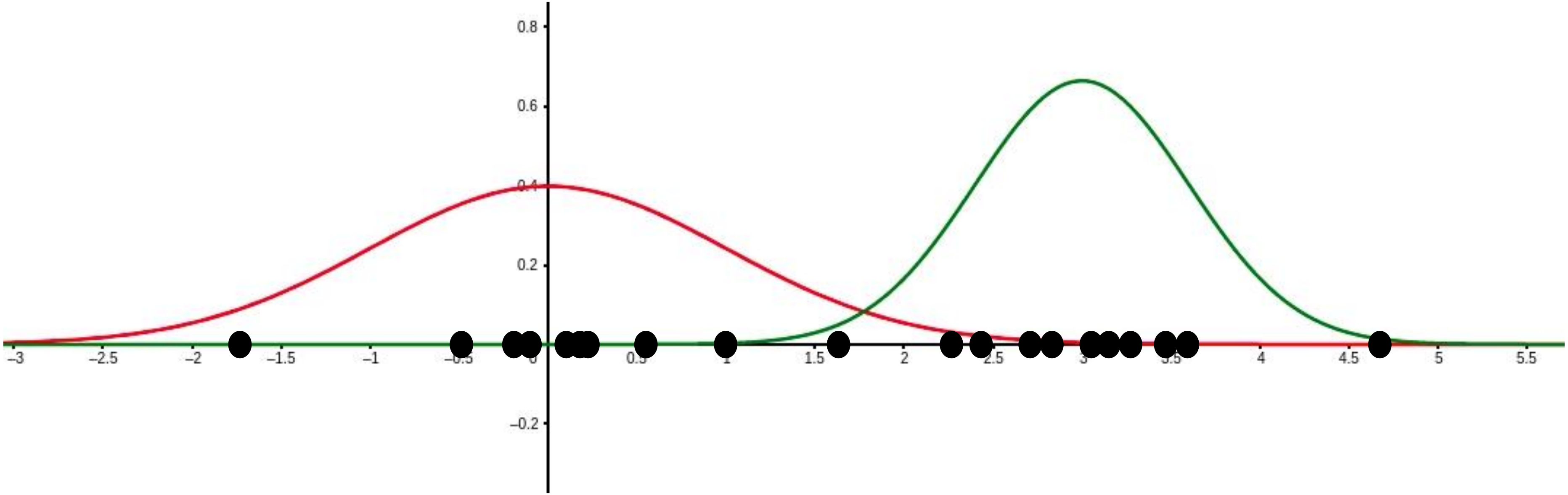
Generate data where $P(C_1) = P(C_2) = ½$ and within C_1 and C_2 the weight distributions are $\mathbf{N(\mu_1, \sigma_1)}$ and $\mathbf{N(\mu_2, \sigma_2)}$



Any of these points could technically have been generated from either curve.

Soft Clustering - Example

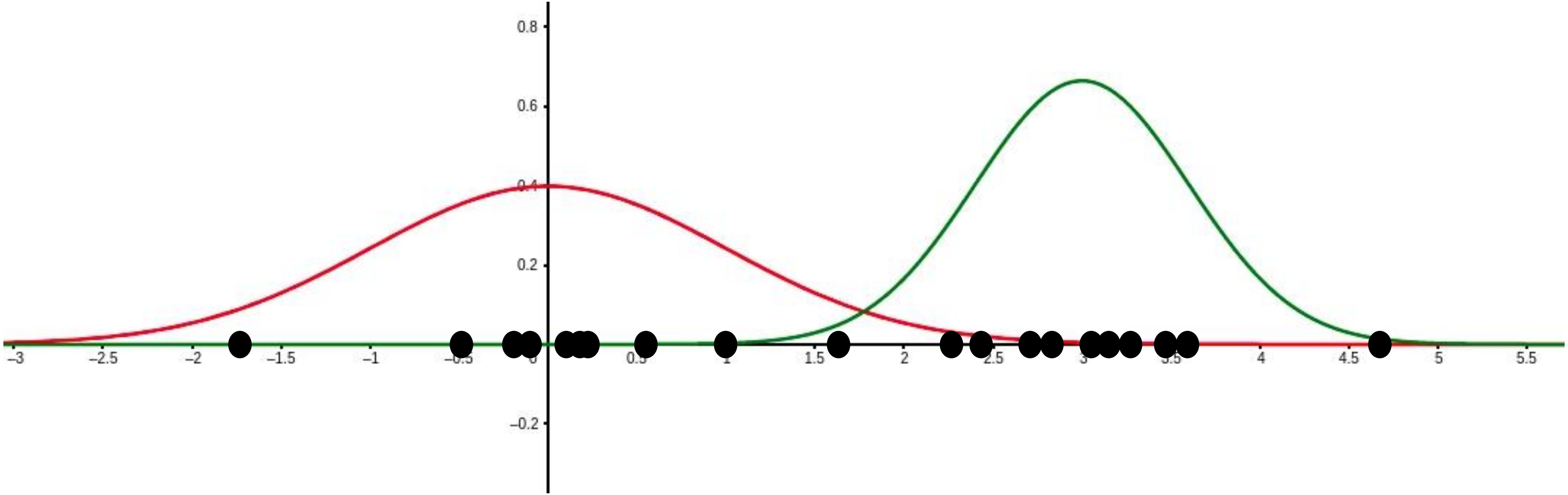
Generate data where $P(C_1) = P(C_2) = ½$ and within C_1 and C_2 the weight distributions are $\mathbf{N(\mu_1, \sigma_1)}$ and $\mathbf{N(\mu_2, \sigma_2)}$



For each point we can compute the probability of it being generated from either curve

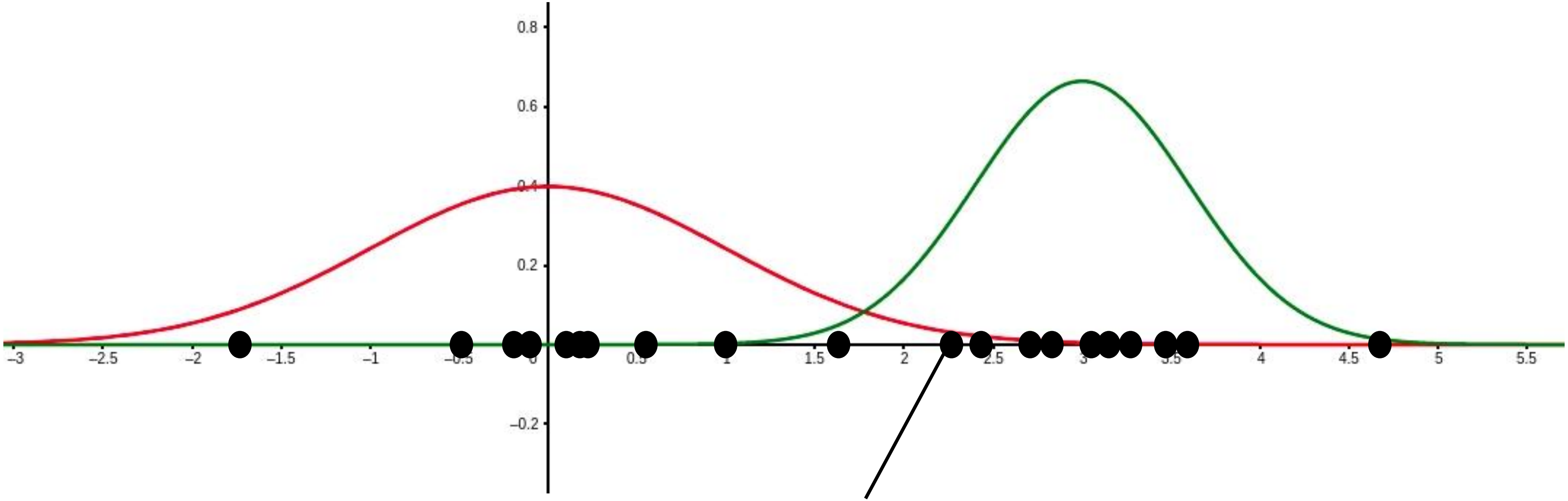
Soft Clustering - Example

Generate data where $P(C_1) = P(C_2) = ½$ and within C_1 and C_2 the weight distributions are $\mathbf{N(\mu_1, \sigma_1)}$ and $\mathbf{N(\mu_2, \sigma_2)}$



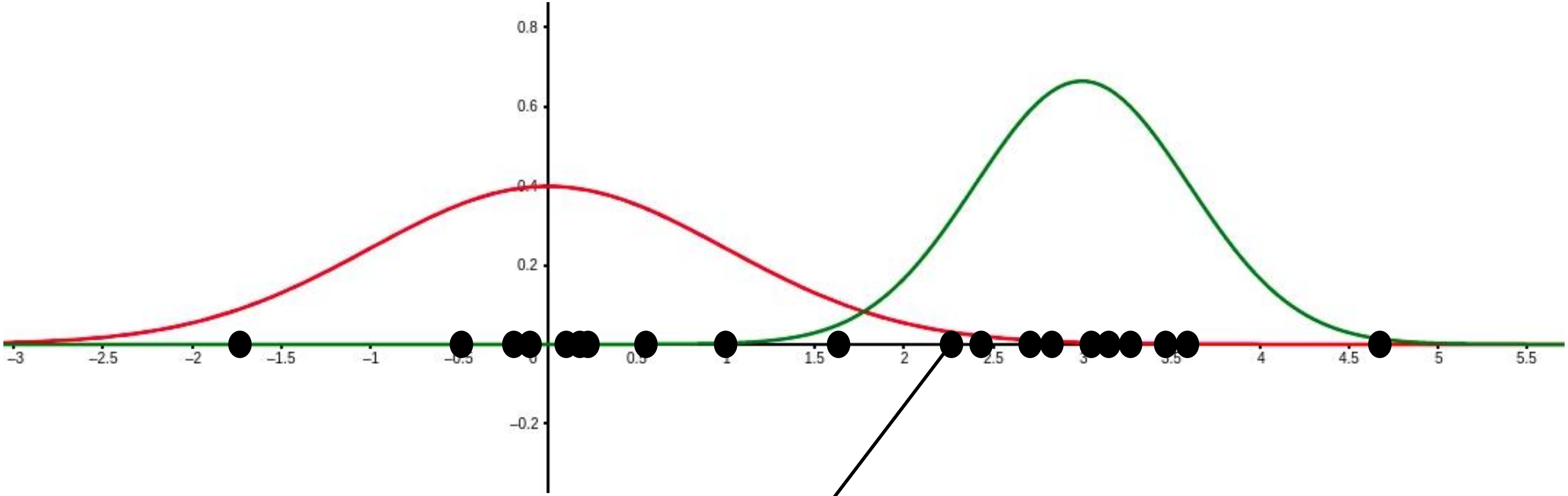
We can create soft assignments based on these probabilities.

Example



What is the probability density here?

Example



$$P(X = x) = P(C_1)P(X = x|C_1) + P(C_2)P(X = x|C_2)$$

Mixture Model

X comes from a mixture model with k mixture components if the probability distribution of X is:

$$P(X = x) = \sum_{j=1}^{k} P(C_j) P(X = x | C_j)$$

Mixture proportion
Represents the probability of belonging to C_j

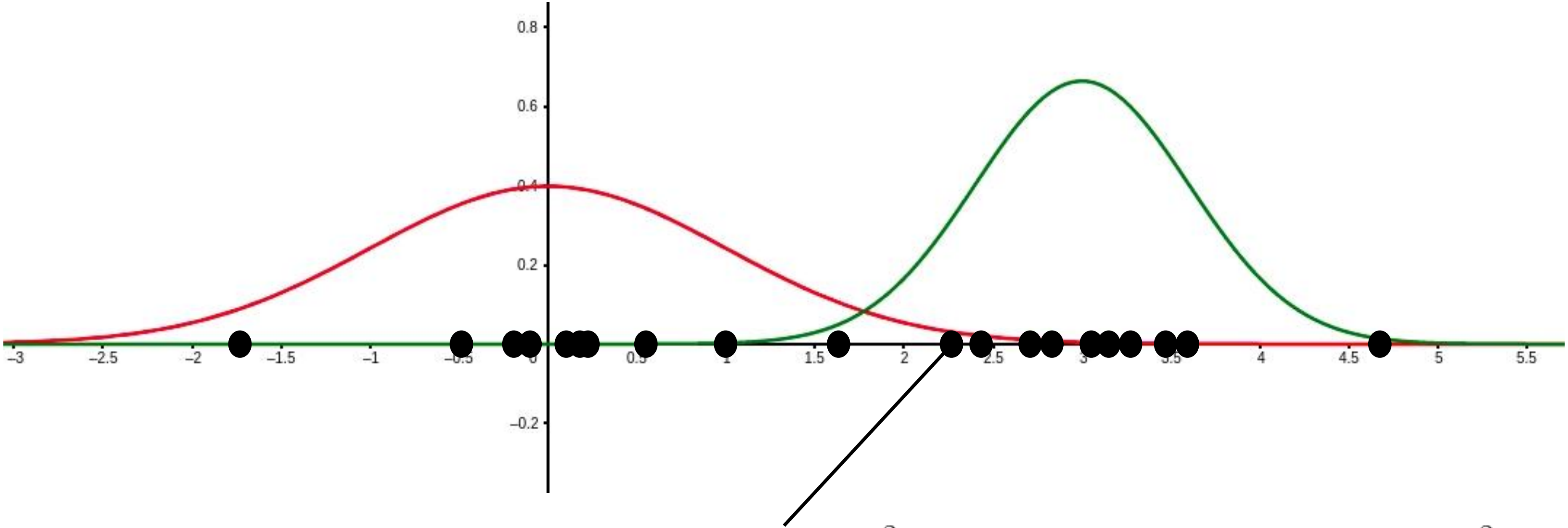
Probability of seeing x when sampling from C_j

Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a mixture model where

$$P(X = x|C_i) \sim N(\mu, \sigma)$$

Example



$$P(X=x) = P(C_1)\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + P(C_2)\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$

Worksheet a) -> c)

Maximum Likelihood Estimation (intuition)

Suppose you are given a dataset of coin tosses and are asked to estimate the parameters that characterize that distribution - how would you do that?

MLE: find the parameters that maximized the probability of having seen the data we got

Maximum Likelihood Estimation (intuition)

Example: Assume Bernoulli(p) iid coin tosses

Val
H
T
T
H
T

Goal: find p that maximized that probability

Maximum Likelihood Estimation (intuition)

Example: Assume Bernoulli(p) iid coin tosses

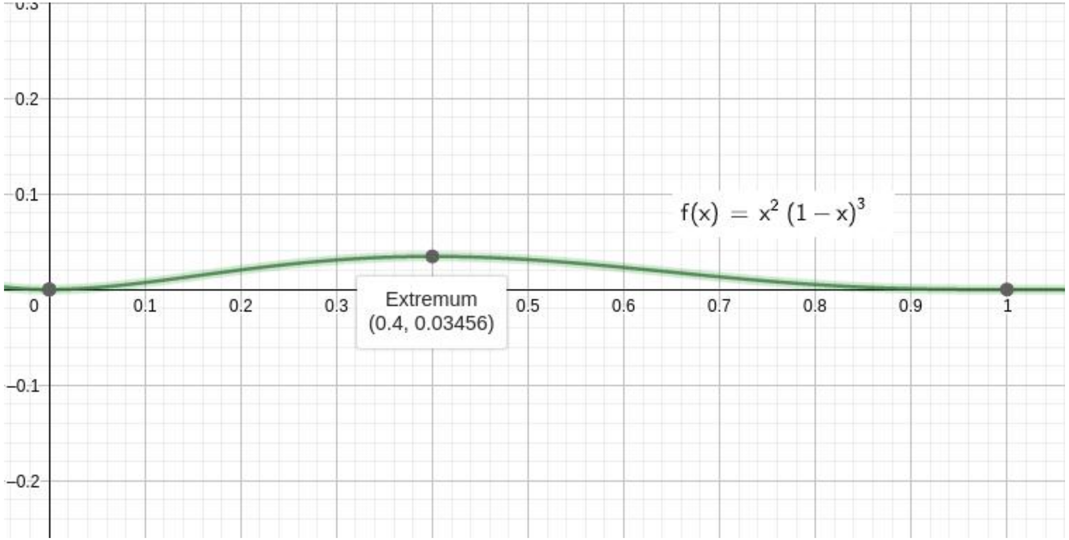
Val
H
T
T
H
T

$$\begin{aligned} P(\text{having seen the data we saw}) &= P(H)P(T)P(T)P(H)P(T) \\ &= p^2(1-p)^3 \end{aligned}$$

Goal: find p that maximized that probability

Maximum Likelihood Estimation (intuition)

Val
H
T
T
H
T



The sample proportion ⅖ is what maximizes this probability

GMM Clustering

Goal: Find the GMM that maximizes the probability of seeing the data we have.

Recall: $$P(X = x) = \sum_{j=1}^{k} P(C_j)P(X = x|C_j)$$

GMM Clustering

Goal: Find the GMM that maximizes the probability of seeing the data we have.

$$P(X = x) = \sum_{j=1}^{k} P(C_j)P(X = x|C_j)$$

Finding the GMM means finding the parameters that uniquely characterize it. What are these parameters?

GMM Clustering

Goal: Find the GMM that maximizes the probability of seeing the data we have.

$$P(X = x) = \sum_{j=1}^{k} P(C_j)P(X = x|C_j)$$

Finding the GMM means finding the parameters that uniquely characterize it. What are these parameters?

$\mathbf{P(C_i)}$ & $\boldsymbol{\mu_i}$ & $\boldsymbol{\sigma_i}$ for all **k** components.

Lets call $\boldsymbol{\Theta}$ = $\{\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_k}, \boldsymbol{\sigma_1}, \ldots, \boldsymbol{\sigma_k}, P(C_1), \ldots, P(C_k)\}$

GMM Clustering

Goal: Find the GMM that maximizes the probability of seeing the data we have.

$$P(X = x) = \sum_{j=1}^{k} P(C_j) P(X = x | C_j)$$

The probability of seeing the data we saw is (**assuming each data point was sampled independently**) the product of the probabilities of observing each data point.

GMM Clustering

Goal:

$$\theta^* = \arg\max_{\theta} \prod_{i=1}^{n} \sum_{j=1}^{k} P(C_j) P(X_i \mid C_j)$$

Where $\boldsymbol{\Theta}$ = **{**$\boldsymbol{\mu}_1$, …, $\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_1$, …, $\boldsymbol{\sigma}_k$, $P(C_1)$, …, $P(C_k)$**}**

Joint probability distribution of our data

Assuming our data are independent

GMM Clustering

How do we find the critical points of this function?

Notice: taking the log-transform does not change the critical points

Define:

$$
\begin{aligned}
l(\theta) &= \log(L(\theta)) \\
&= \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} P(C_j)P(X_i \mid C_j)\right)
\end{aligned}
$$

GMM Clustering

For $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k]^T$ and $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k]^T$

We can solve

$$\frac{d}{d\Sigma} l(\theta) = 0 \qquad \frac{d}{d\mu} l(\theta) = 0$$

GMM Clustering

To get

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} P(C_j|X_i)X_i}{\sum_{i=1}^{n} P(C_j|X_i)}$$

$$\hat{\Sigma}_j = \frac{\sum_{i=1}^{n} P(C_j|X_i)(X_i - \hat{\mu}_j)^T(X_i - \hat{\mu}_j)}{\sum_{i=1}^{n} P(C_j|X_i)}$$

$$\hat{P}(C_j) = \frac{1}{n}\sum_{i=1}^{n} P(C_j|X_i)$$

GMM Clustering

Do we have everything we need to solve this?

Still need **$P(C_j \mid X_i)$** (i.e. the probability that X_i was drawn from C_j)

GMM Clustering

$$P(C_j|X_i) = \frac{P(X_i|C_j)}{P(X_i)} P(C_j)$$
$$= \frac{P(X_i|C_j)P(C_j)}{\sum_{j=1}^{k} P(C_j)P(X_i|C_j)}$$

Looks like a loop! Seems we need $P(C_j)$ to get $P(C_j \mid X_i)$ and $P(C_j \mid X_i)$ to get $P(C_j)$

Expectation Maximization Algorithm

1. Start with random $\boldsymbol{\theta}$
2. Compute $\mathbf{P(C_j \mid X_l)}$ for all $\mathbf{X_i}$ by using $\boldsymbol{\theta}$
3. Compute / Update $\boldsymbol{\theta}$ from $\mathbf{P(C_j \mid X_l)}$
4. Repeat 2 & 3 until convergence

Worksheet d) -> h)