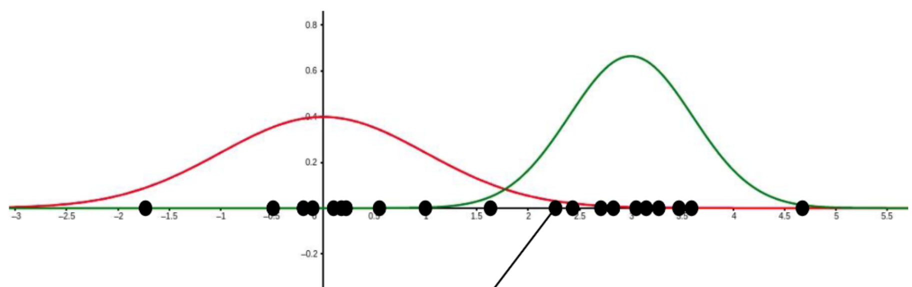
# Lecture 7: Soft Clustering

## 1) Soft Clustering

a) Previous: hard assignment(1 point -> 1 cluster)
   i) Sometimes data isn't accurately represented -> reasonable to have overlapping clusters
b) Assign points to every cluster with certain probability

## 2) Example

a) Things to consider:
   i) There is a prior probability of being one species
      1) Can have imbalance dataset ir there could be more than one species than the other
   ii) Weights within particular group/species follow a particular distribution
b) Generate data where $P(C_1) = P(C_2) = \frac{1}{2}$ and within $C_1$ and $C_2$ the weight distributions are $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$



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

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

- Any of these points could be generated from either curve
  - Can compute probability each point is generated from either curve
- Create soft assignment based on probabilities

c) Mixture Model

i) X comes from mixture model with k mixture components if 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 $C_j$

ii) Gaussian Mixture Model

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

## 3) Maximum Likelihood Estimation (intuition)

a) Find parameters that maximize probability of having seen data we got

b) Scenario: given dataset of coin tosses and asked to estimate parameter that distribution

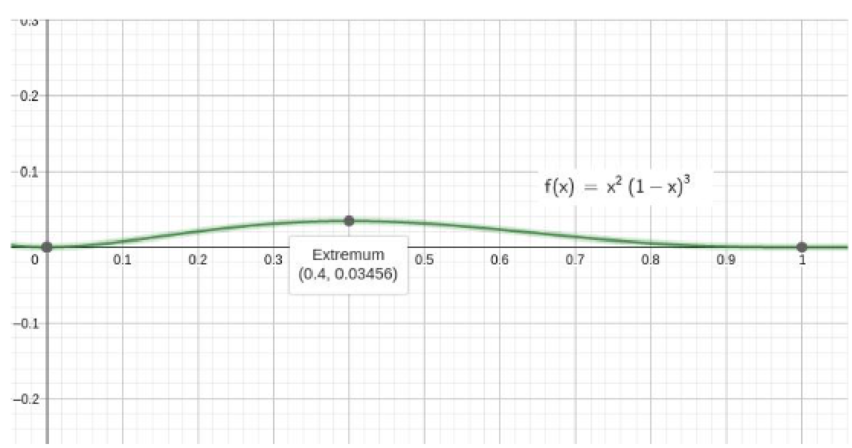
i) Assume Bernoulli(p) iid coin tosses

ii) Values: H T T H T

iii) Goal: find p that maximized probability

P(having seen the data we saw) = P(H)P(T)P(T)P(H)P(T)

= $p^2(1\text{-}p)^3$




- Sample proportion ⅖ maximizes this probability

## 4) GMM Clustering

a) Find GMM that maximizes probability of seeing data we have

i) Probability of seeing data we saw, assuming each data point was sampled independently, is the product of probabilities of observing each data point

**$P(C_i)$** & **$\mu_i$** & **$\sigma_i$** for all **k** components.

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

b) Goal:

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

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

- Joint probability distribution of our data, assuming data is independent

c) Log transform does not change critical points

i) Define:

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

ii) For $\mu = [\mu_1, \ldots, \mu_k]^T$ and $\Sigma = [\Sigma_1, \ldots, \Sigma_k]^T$, we can solve

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

To get:

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

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

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

d) Still need $P(C_j|X_i)$ (Probability $X_l$ was drawn from $C_j$)

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

need $P(C_j)$ to get $P(C_j \mid X_i)$ and $P(C_j \mid X_i)$ to get $P(C_j)$

## 5) Expectation Maximization Algorithm

a) Start with random θ
b) Computer P($C_j$|$X_i$) for all $X_i$ by using θ
c) Compute/update θ from previous P($C_j$|$X_i$)
d) Repeat 2 & 3 until convergence