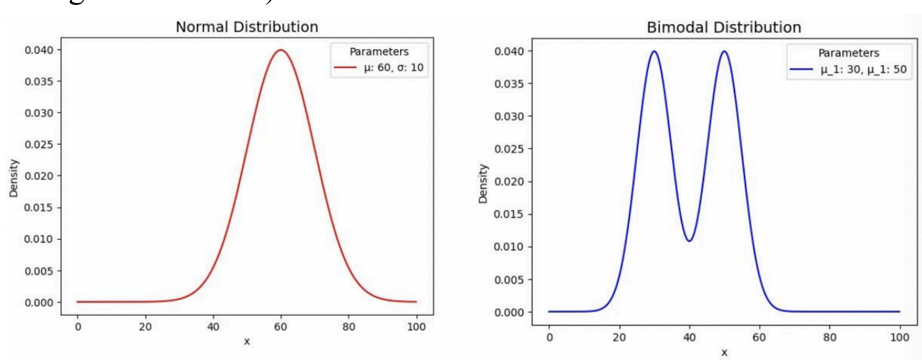
CAS CS 506
Lec 08

Soft Clustering

1. Problem Statement
    a. Given a dataset of weights sampled from N different animals
    b. Can determine which weight belongs to which animal?
2. Output
    a. Makes more sense to provide, for each data point (weight) the probability that it came from each species
    $P(S_j \mid X_i)$
    Where S_j is species j and X_i is the ith weight in the dataset
3. Things to Consider
    a. There is a prior probability of being one species (i.e. could have an imbalance dataset or there could just be more of one species than the other)
        i. Ex: Some dinosaurs are more common than others: for example there are many Stegosauruses than Raptors in the park. This means a given data point, knowing nothing about it would just have a higher chance of being a Stegosaurus than a Raptor
    b. Weights vary differently depending on the species (i.e. each species could have a different weight distribution)
        i. 

4. How to Compute
    a. $$P(S_j|X_i) = \frac{P(X_i|S_j)P(S_j)}{P(X_i)}$$
    b. $P(S_j)$ is the prior probability of seeing species S_j (that probability would be higher for Stegosauruses than the Raptors for example)
    c. $P(X_i \mid S_j)$ is the PDF of species S_j weights evaluated at weight X_i (seeing a Sauropod that weighs 100 tons is way more likely than seeing a Raptor that weights 100 tons)

5. What about $P(X_i)$?
    a. $$P(X_i) = \sum_j P(S_j)P(X_i|S_j)$$
6. Mixture Model
    a. X comes from a mixture model with k mixture components if the probability distribution of X is

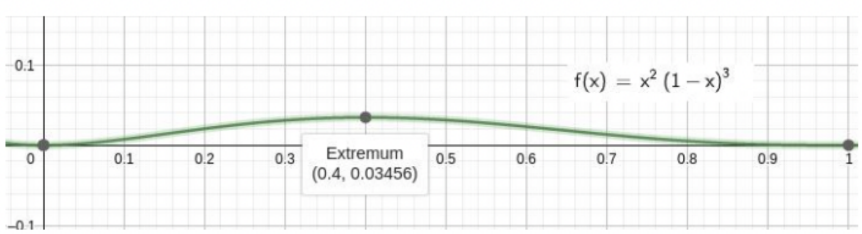
$$P(X) = \sum_j \underbrace{P(S_j)}_{\substack{\text{Mixture proportion} \\ \text{Represents the probability} \\ \text{of belonging to } S_j}} \underbrace{P(X|S_j)}_{\substack{\text{Probability of seeing x} \\ \text{when sampling from } S_j}}$$

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

$$P(X|S_j) \sim N(\mu, \sigma)$$

8. Maximum Likelihood Estimation (Intuition)
    a. Suppose there is a given dataset of coin tosses and need to estimate the parameters that characterize that distribution - how would that be done?
        i. MLE: find the parameters that maximized the probability of having seen the data given
    b. Example: Assume Bernoulli(p) iid coin tosses

Val
H
T
T
H
T

        i. Find p that maximized that probability
        ii. P(having seen the data we saw) = p^2 * (1-p)^3
        iii. 

        iv. The sample proportion ⅖ is what maximizes this probability

9. GMM Clustering
    a. Goal: Find the GMM that maximizes the probability of seeing the data gathered
    b. Recall: $$P(X_i) = \sum_j P(S_j)P(X_i|S_j)$$
    c. Finding the GMM means finding the parameters that uniquely characterize it.
    d. Parameters
        i. $P(S_j)$ & $\boldsymbol{\mu}_j$ & $\boldsymbol{\sigma}_j$ for all k components
        ii. Let's call $\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_k, P(S_1), \ldots, P(S_k)\}$
    e. 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
    f. Goal:
$$\prod_i P(X_i) = \prod_i \sum_j P(S_j)P(X_i|S_j)$$
    g. How do we find the critical points of this function?
        i. Take the log-transform since it does not change the critical points
        ii. $$\log\left(\prod_i \sum_j P(S_j)P(X_i|S_j)\right) = \sum_i \log\left(\sum_j P(S_j)P(X_i|S_j)\right)$$
$$\hat{\mu_j} = \frac{\sum_i P(S_j|X_i)X_i}{\sum_i P(S_j|X_i)}$$
$$\hat{\Sigma}_j = \frac{\sum_i P(S_j|X_i)(X_i - \hat{\mu_j})^T(X_i - \hat{\mu_j})}{\sum_i P(S_j|X_i)}$$
        iii. $$\hat{P}(S_j) = \frac{1}{N}\sum_i P(S_j|X_i)$$
10. Expectation Maximization Algorithm
    a. Start with random $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $P(S_j)$
    b. Compute $P(S_j \mid X_I)$ for all X_i by using $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $P(S_j)$
    c. Compute / Update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $P(S_j)$ from $P(S_j \mid X_I)$
    d. Repeat 2 & 3 until convergence