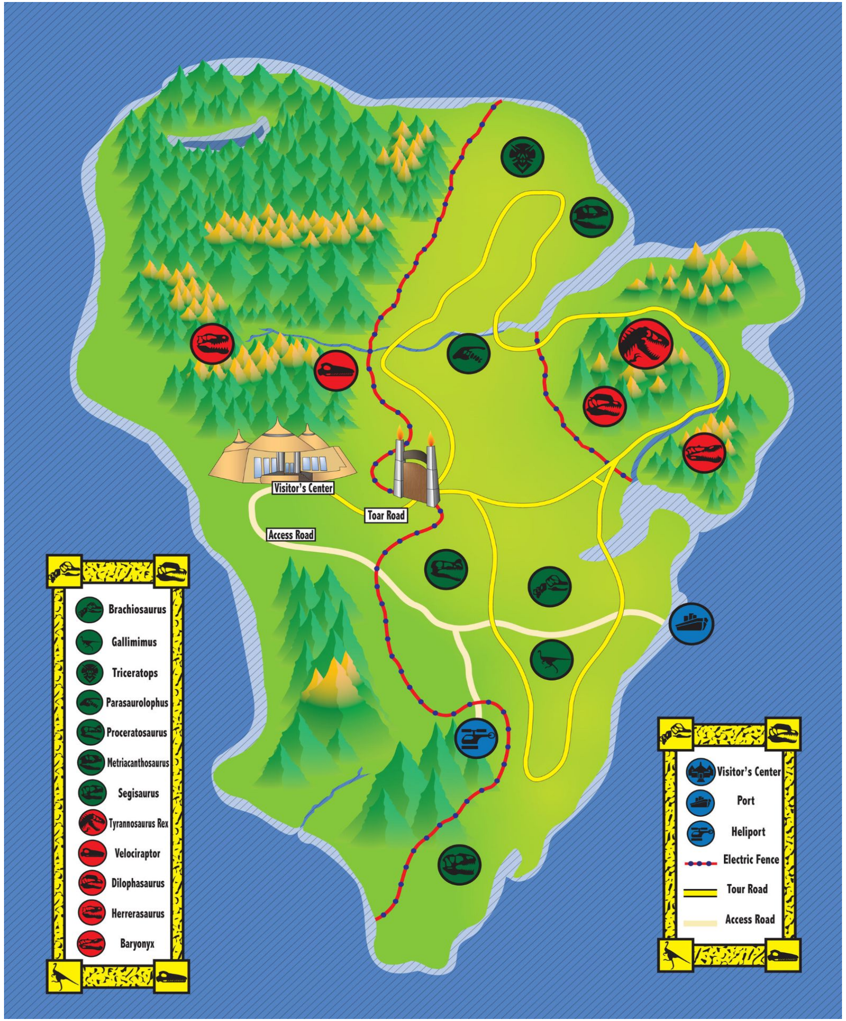
Soft Clustering

Boston University CS 506 - Lance Galletti


Visitor's Center
Toar Road
Access Road
Brachiosaurus
Gallimimus
Triceratops
Parasaurolophus
Proceratosaurus
Metriacanthosaurus
Segisaurus
Tyrannosaurus Rex
Velociraptor
Dilophasaurus
Herrerasaurus
Baryonyx
Visitor's Center
Port
Heliport
Electric Fence
Tour Road
Access Road

Problem Statement

Given a dataset of weights sampled from N different animals.

Can we determine which weight belongs to which animal?

Output

Makes more sense to provide, for each data point (weight) the probability that it came from each species.

$$P(S_j | X_i)$$

Where S_j is species j and X_i is the i^{th} weight in the dataset.

Things To Consider

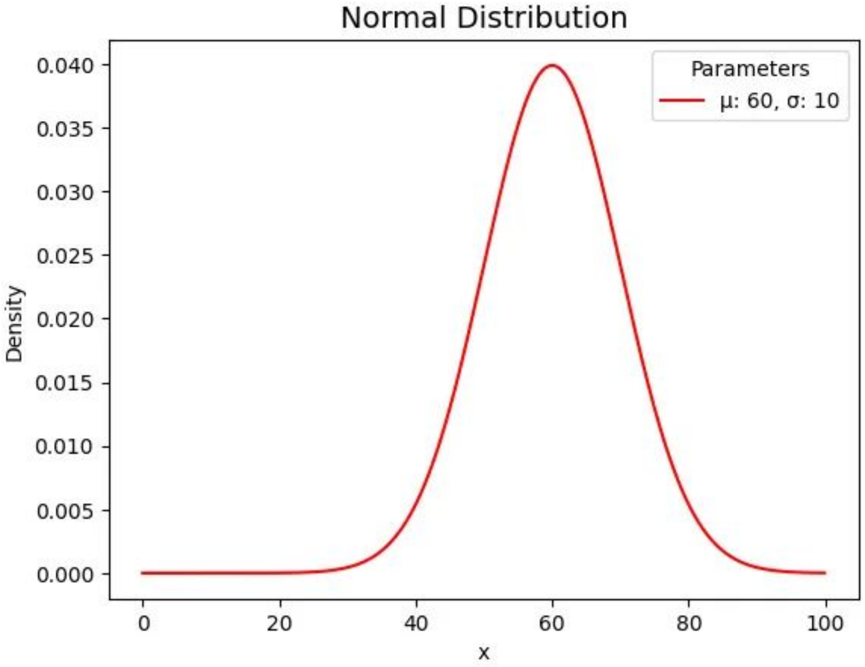
1. There is a prior probability of being one species (i.e. we could have an imbalanced dataset or there could just be more of one species than the other)

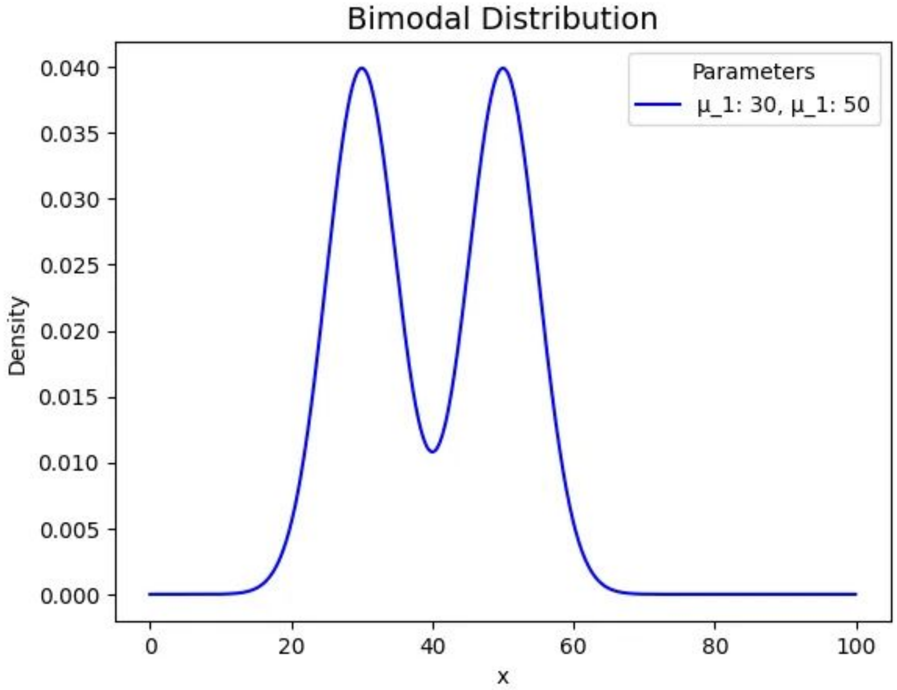
Some dinosaurs are more common than others: for example there are many more Stegosauruses than Raptors in the park. This means a given data point, knowing nothing about it would just have a higher chance of being a Stegosaurus than a Raptor.

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 vary differently depending on the species (i.e. each species could have a different weight distribution)

Things to Consider





How to compute $P(S_j \mid X_i)$?

$$P(S_j|X_i) = \frac{P(X_i|S_j)P(S_j)}{P(X_i)}$$

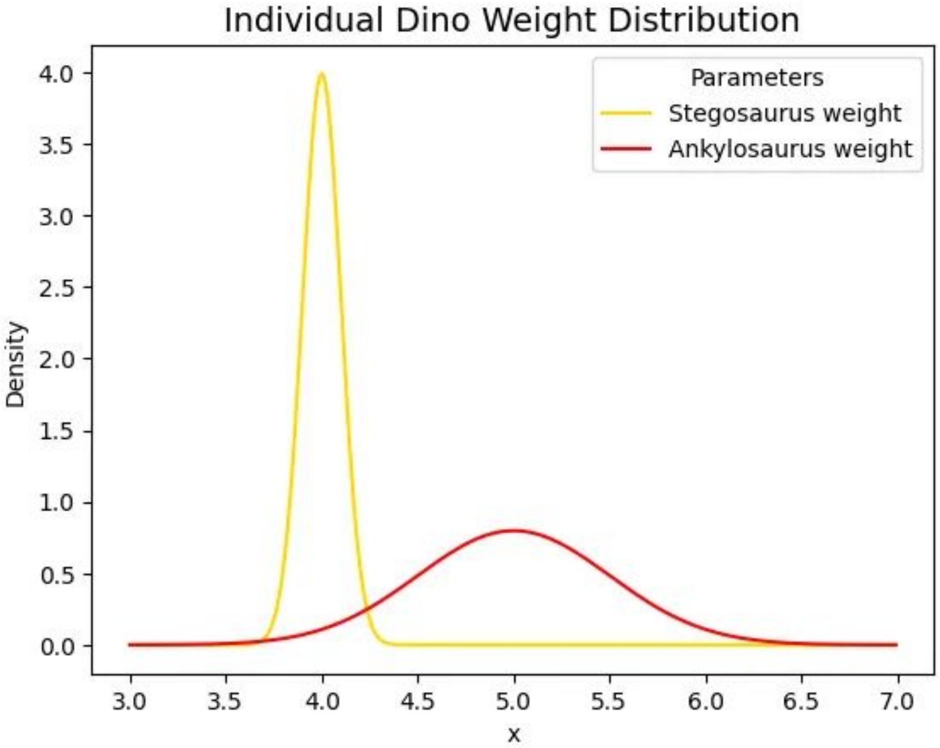
How to compute $P(S_j \mid X_i)$?

$$P(S_j|X_i) = \frac{P(X_i|S_j)P(S_j)}{P(X_i)}$$

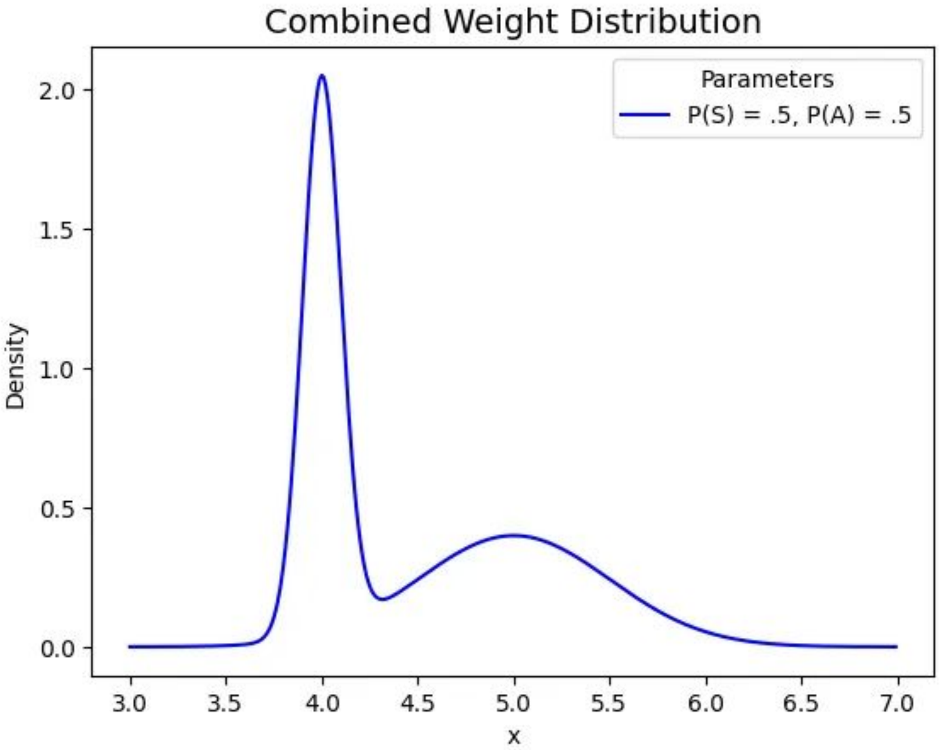
$P(S_j)$ is the prior probability of seeing species S_j (that probability would be higher for the Stegosauruses than the Raptors for example)

$P(X_i \mid S_j)$ 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 weighs 100 tons)

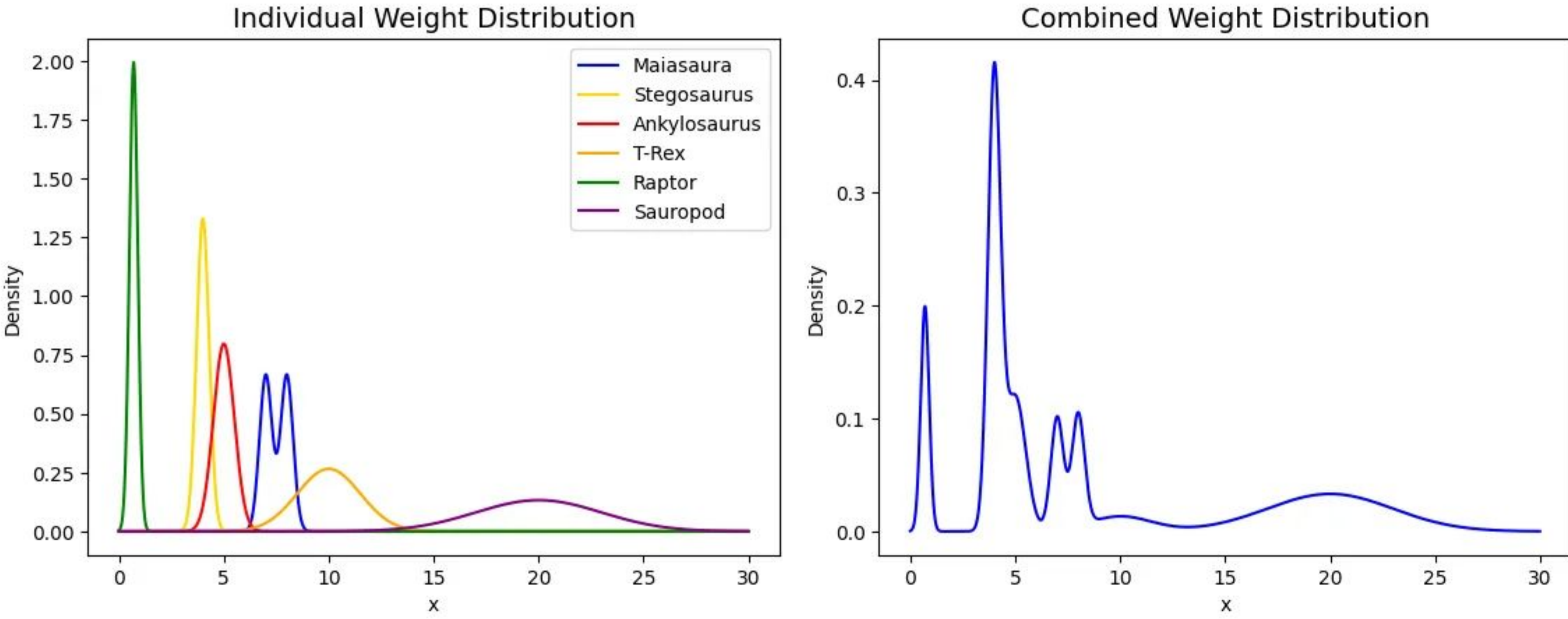
What about $P(X_i)$?



What about $P(X_i)$?



What about P(X_i) ?



What about $P(X_i)$?

$$P(X_i) = \sum_j P(S_j)P(X_i|S_j)$$

Mixture Model

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

$$P(X) = \sum_j P(S_j)P(X|S_j)$$

Mixture proportion
Represents the probability
of belonging to S_j

Probability of seeing x
when sampling from S_j

Gaussian Mixture Model

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

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

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

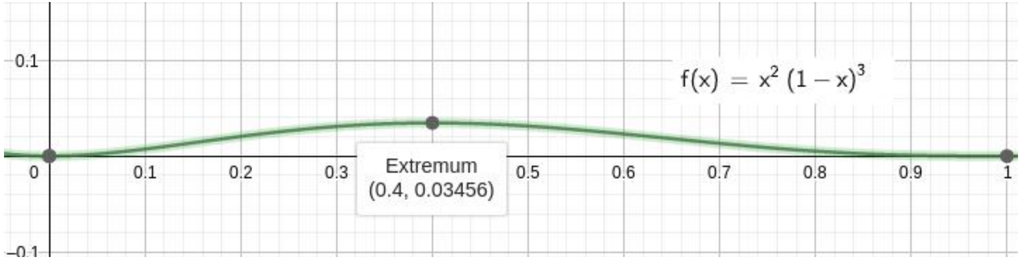
Val
H
T
T
H
T

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

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:

$$P(X_i) = \sum_j P(S_j)P(X_i|S_j)$$

GMM Clustering

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

$$P(X_i) = \sum_j P(S_j)P(X_i|S_j)$$

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.

$$P(X_i) = \sum_j P(S_j)P(X_i|S_j)$$

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

$P(S_j)$ & **μ_j** & **σ_j** for all **k** components.

Lets call ***Θ*** = **{**μ_1, …, μ_k , σ_1, …, σ_k, $P(S_1)$, …, $P(S_k)$**}**

GMM Clustering

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

$$P(X_i) = \sum_j P(S_j)P(X_i|S_j)$$

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:

$$\prod_{i} P(X_i) = \prod_{i} \sum_{j} P(S_j) P(X_i | S_j)$$

GMM Clustering

How do we find the critical points of this function?

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

Define:

$$\log\left(\prod_i \sum_j P(S_j)P(X_i|S_j)\right) = \sum_i \log\left(\sum_j P(S_j)P(X_i|S_j)\right)$$

GMM Clustering

To get

$$\hat{\mu_j} = \frac{\sum_i P(S_j|X_i)X_i}{\sum_i P(S_j|X_i)}$$

$$\hat{\Sigma}_j = \frac{\sum_i P(S_j|X_i)(X_i - \hat{\mu_j})^T(X_i - \hat{\mu_j})}{\sum_i P(S_j|X_i)}$$

critical point.

$$\hat{P}(S_j) = \frac{1}{N}\sum_i P(S_j|X_i)$$

GMM Clustering

Do we have everything we need to solve this?

Expectation Maximization Algorithm

1. Start with random **μ, Σ, $P(S_j)$**
2. Compute **$P(S_j \mid X_l)$** for all **X_i** by using **μ, Σ, $P(S_j)$**
3. Compute / Update **μ, Σ, $P(S_j)$** from **$P(S_j \mid X_l)$**
4. Repeat 2 & 3 until convergence

Worksheet d) -> h)