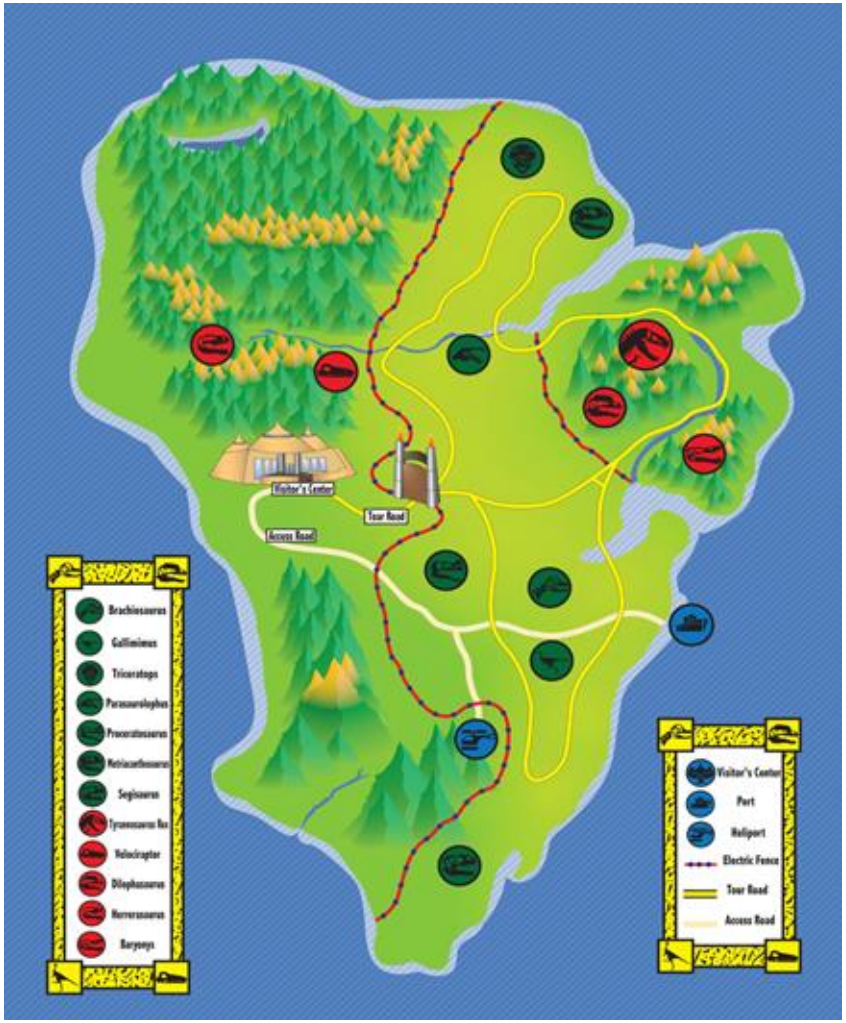
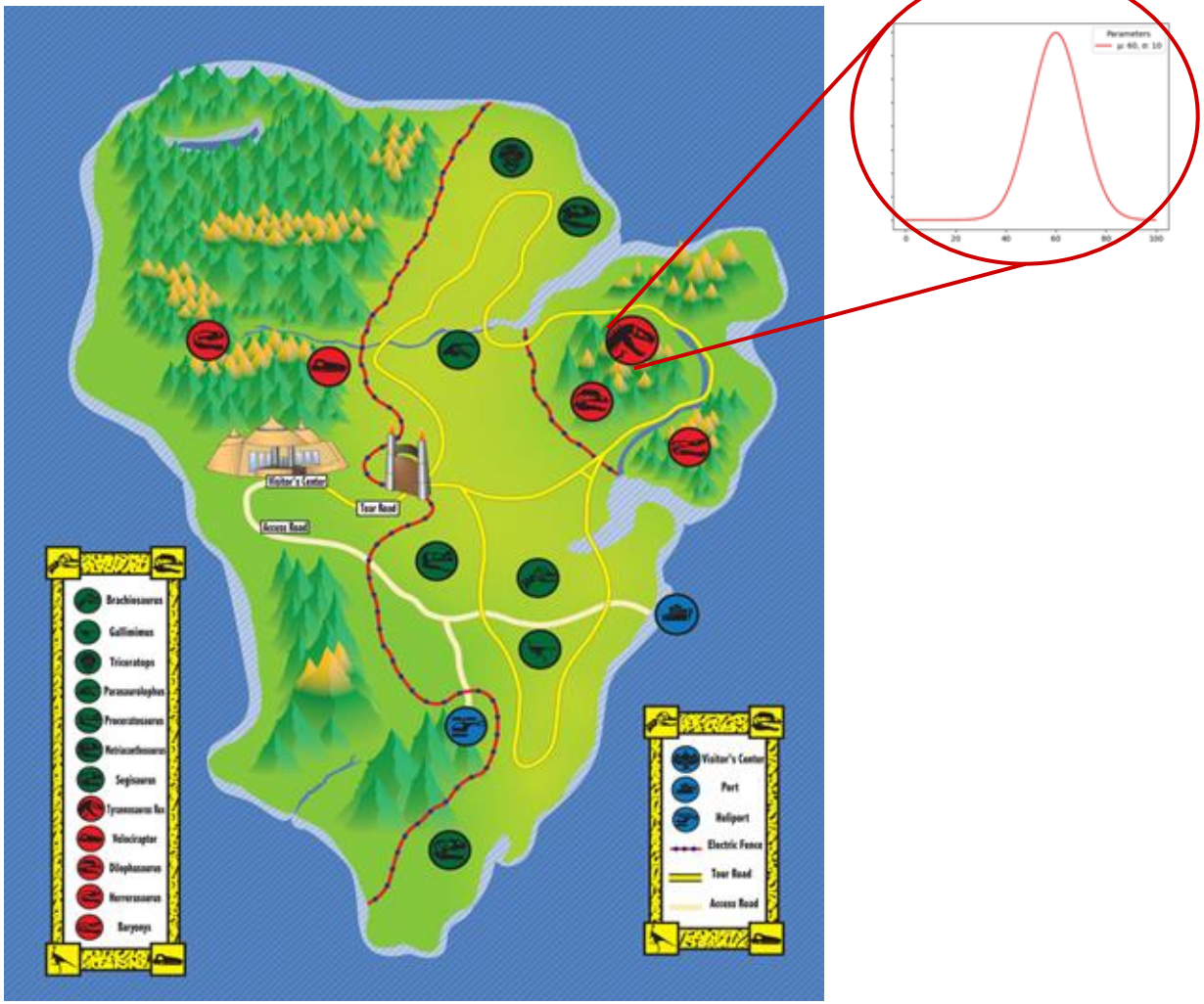
Soft Clustering

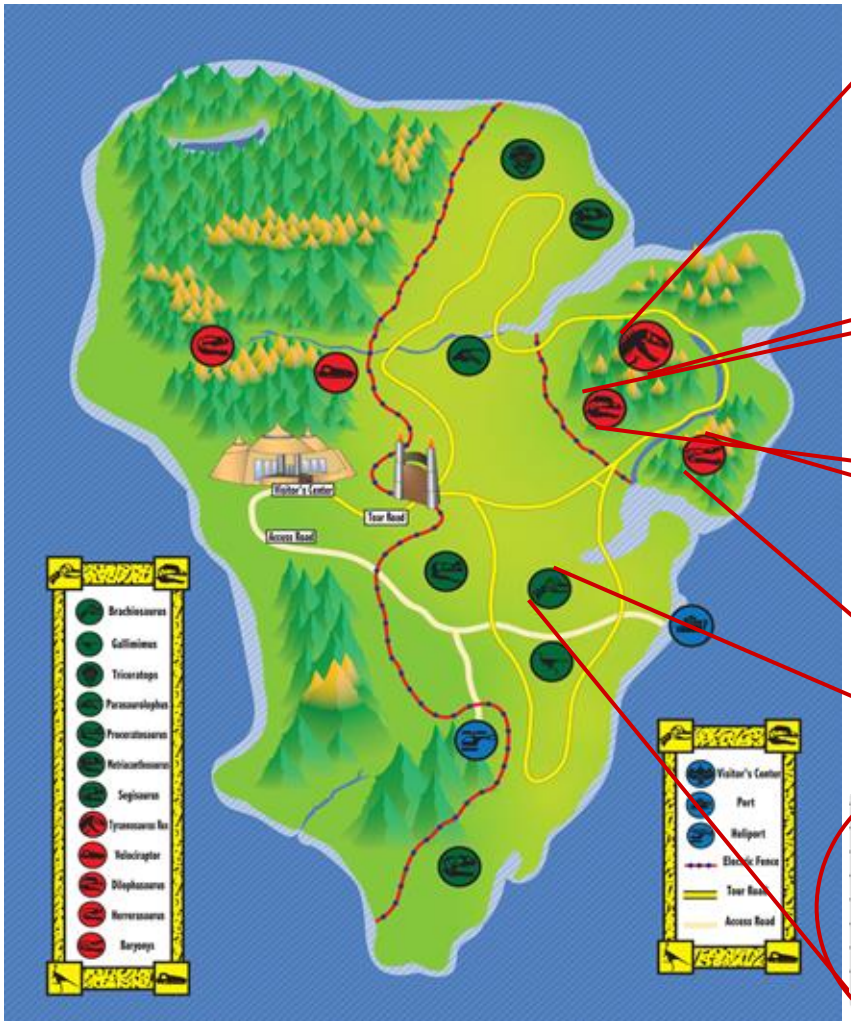
Boston University CS 506 - Lance Galletti

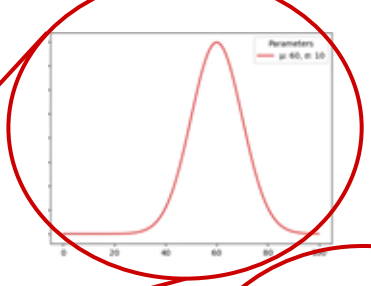

WELCOME

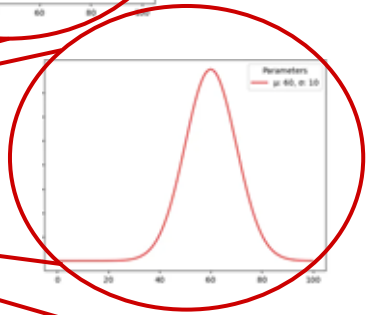

Visitor's Center
Tour Road
Access Road
Brachiosaurus
Gallimimus
Triceratops
Parasaurolophus
Proceratosaurus
Metriacanthosaurus
Segisaurus
Tyrannosaurus Rex
Velociraptor
Dilophosaurus
Herrerasaurus
Baryonyx
Visitor's Center
Port
Heliport
Electric Fence
Tour Road
Access Road

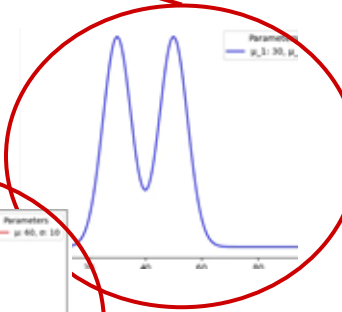

WELCOME

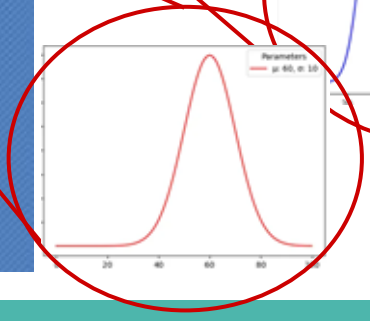

Visitor's Center
Tour Road
Access Road
Brachiosaurus
Gallimimus
Triceratops
Parasaurolophus
Proceratosaurus
Metriacanthosaurus
Segisaurus
Tyrannosaurus Rex
Velociraptor
Dilophosaurus
Herrerasaurus
Baryonyx
Visitor's Center
Port
Heliport
Electric Fence
Tour Road
Access Road
Parameters
μ: 60, σ: 10
0
20
40
60
80
100

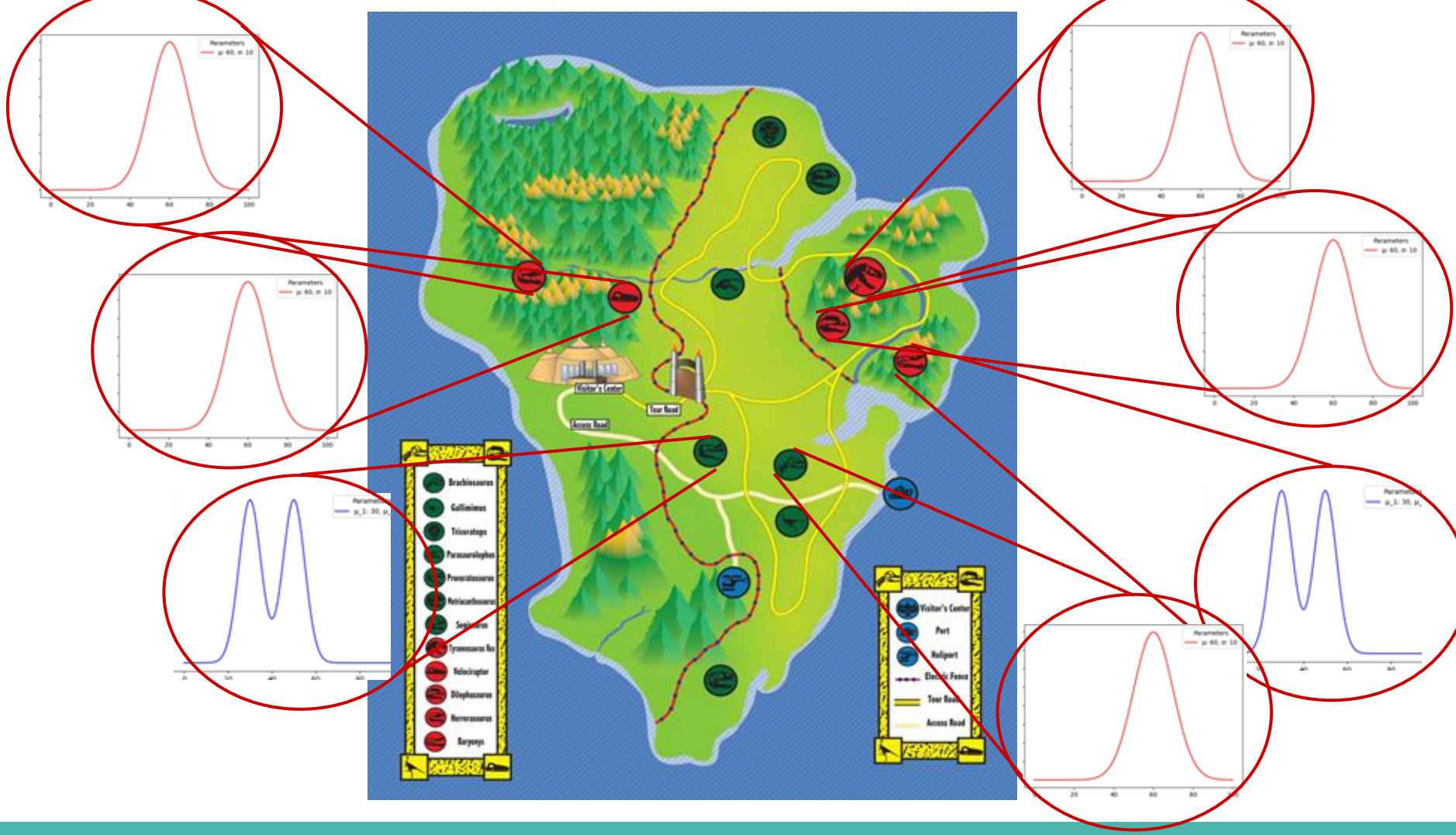

Visitor's Center
Tour Road
Access Road
Brachiosaurus
Gallimimus
Triceratops
Parasaurolophus
Proceratosaurus
Segisaurus
Velociraptor
Dilophosaurus
Herrerasaurus
Baryonyx
Visitor's Center
Port
Heliport
Electric Fence
Tour Road
Access Road


Parameters
μ: 60, σ: 10


Parameters
μ: 60, σ: 10




Parameters
μ: 60, σ: 10


Parameters
μ: 60, σ: 10
Visitor's Center
Tour Road
Access Road
Brachiosaurus
Gallimimus
Triceratops
Parasaurolophus
Proceratosaurus
Tyrannosaurus Rex
Velociraptor
Dilophosaurus
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

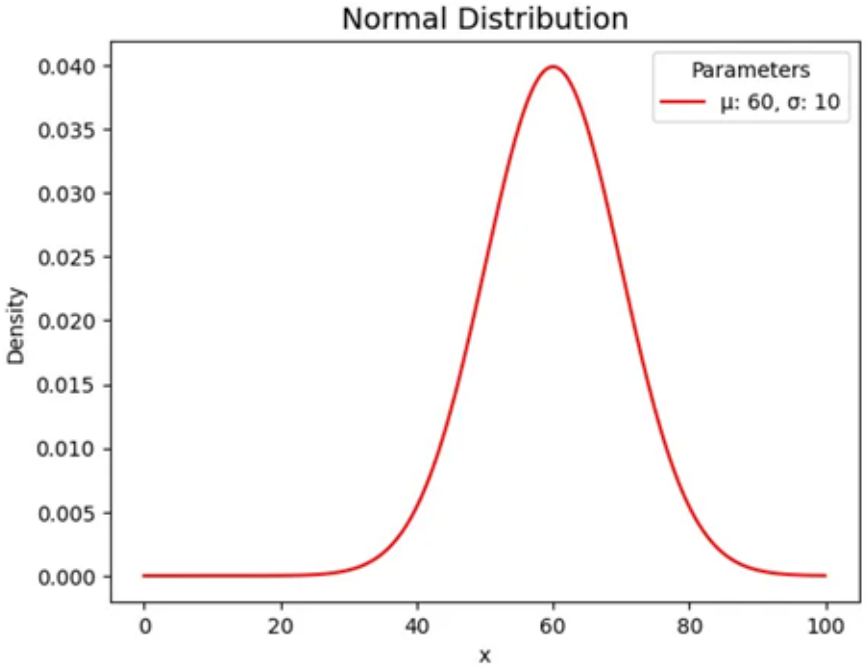
1. There is a prior probability of being one species (i.e. we could have an imbalanced dataset or there could just be more of one species than the other)

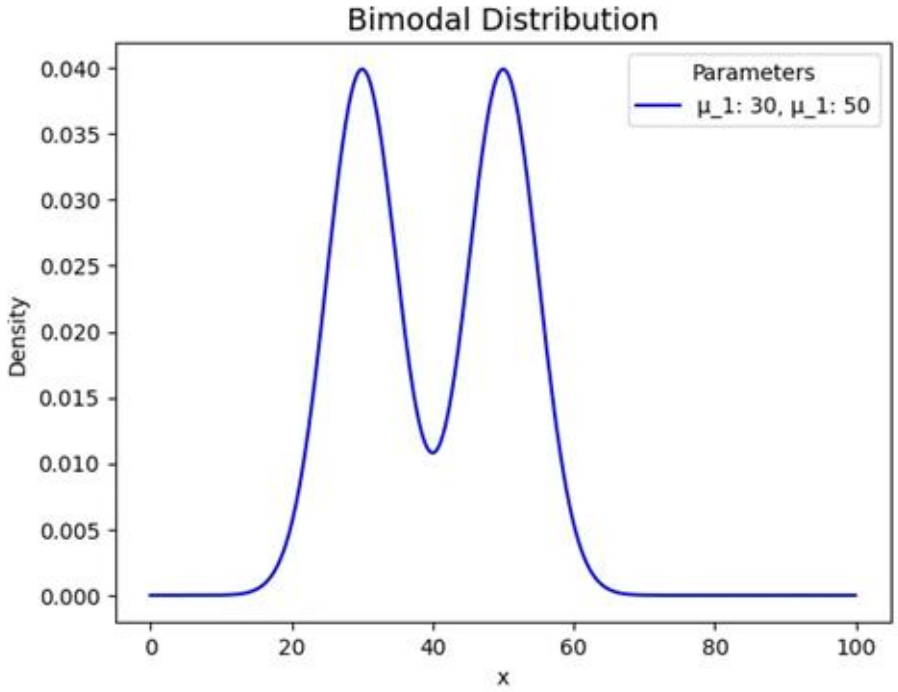
Some dinosaurs are more common than others: for example there are many more Stegosauruses than Raptors in the park. This means a given data point, knowing nothing about it would just have a higher chance of being a Stegosaurus than a Raptor.

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 $P(S_j \mid X_i)$?

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

Conditional Probability

Recall

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$

Conditional Probability

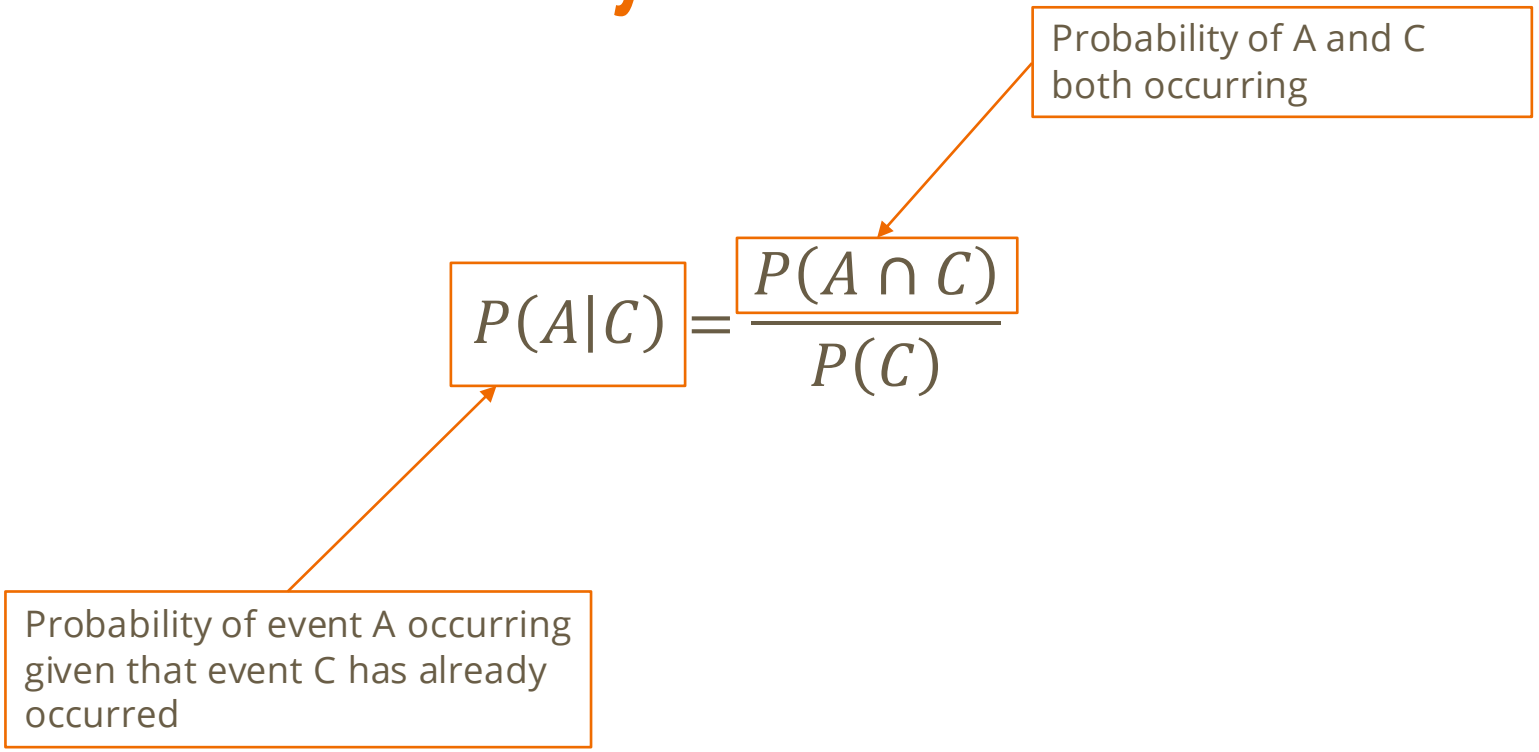
Recall

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$

Probability of event A occurring given that event C has already occurred

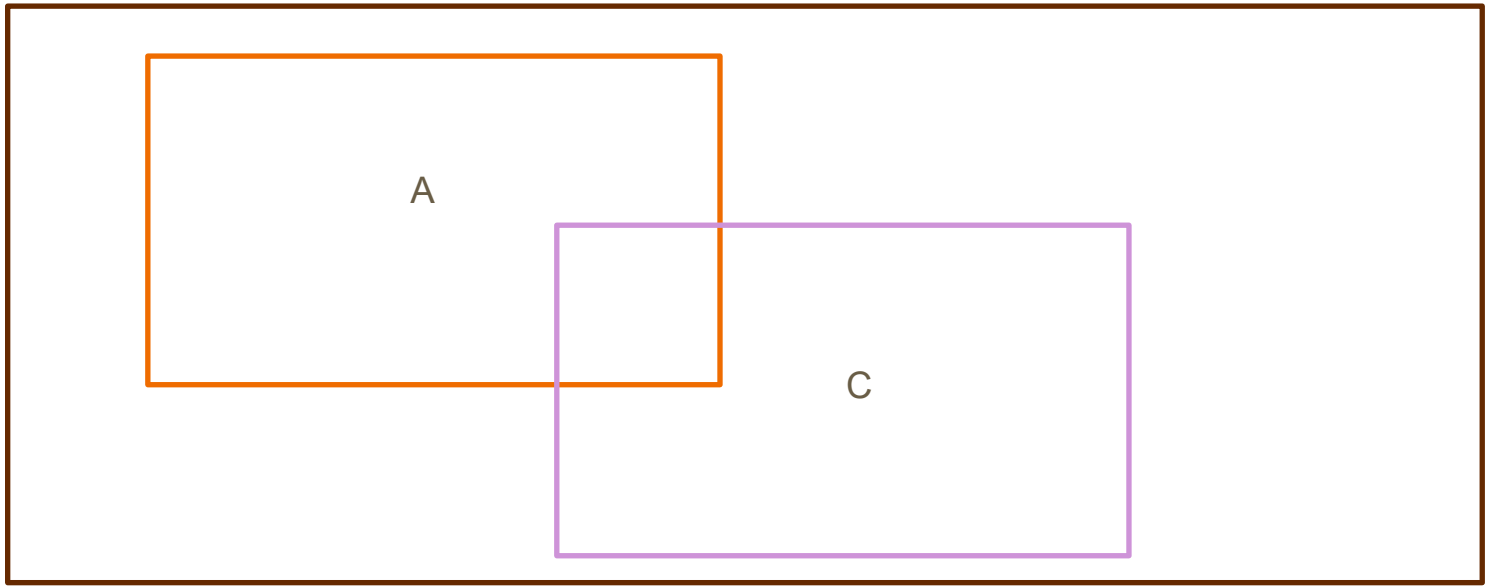
Conditional Probability

Recall



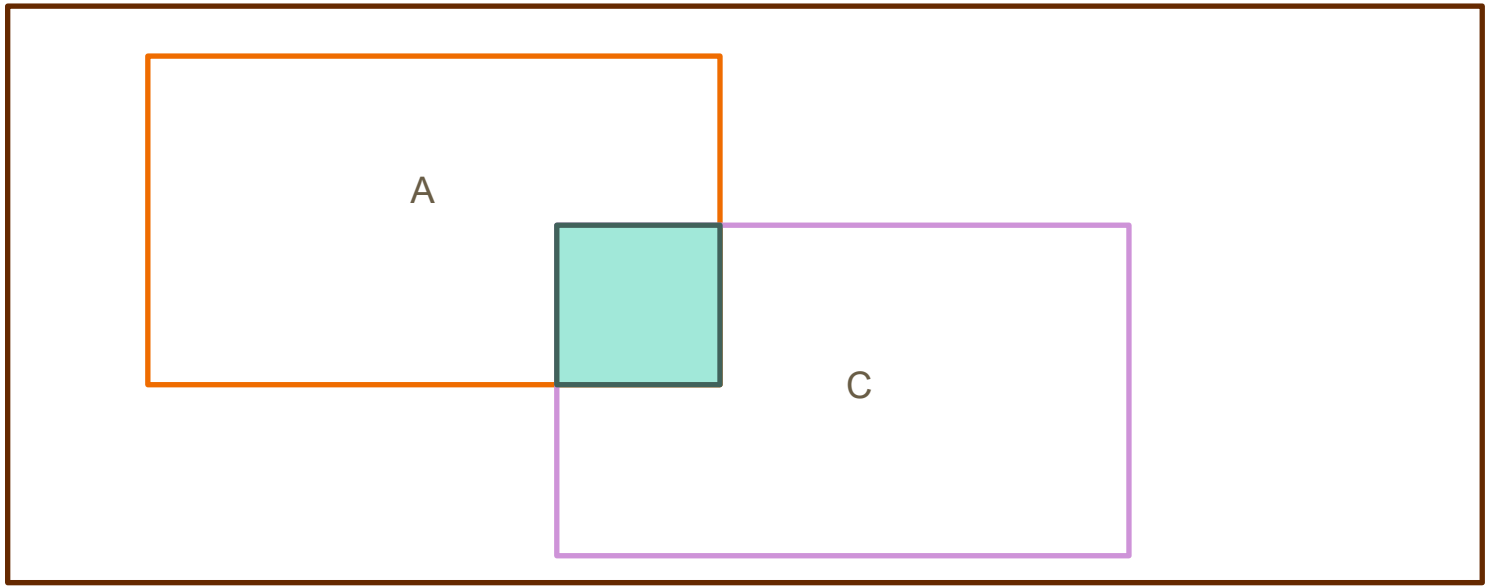
Conditional Probability

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$



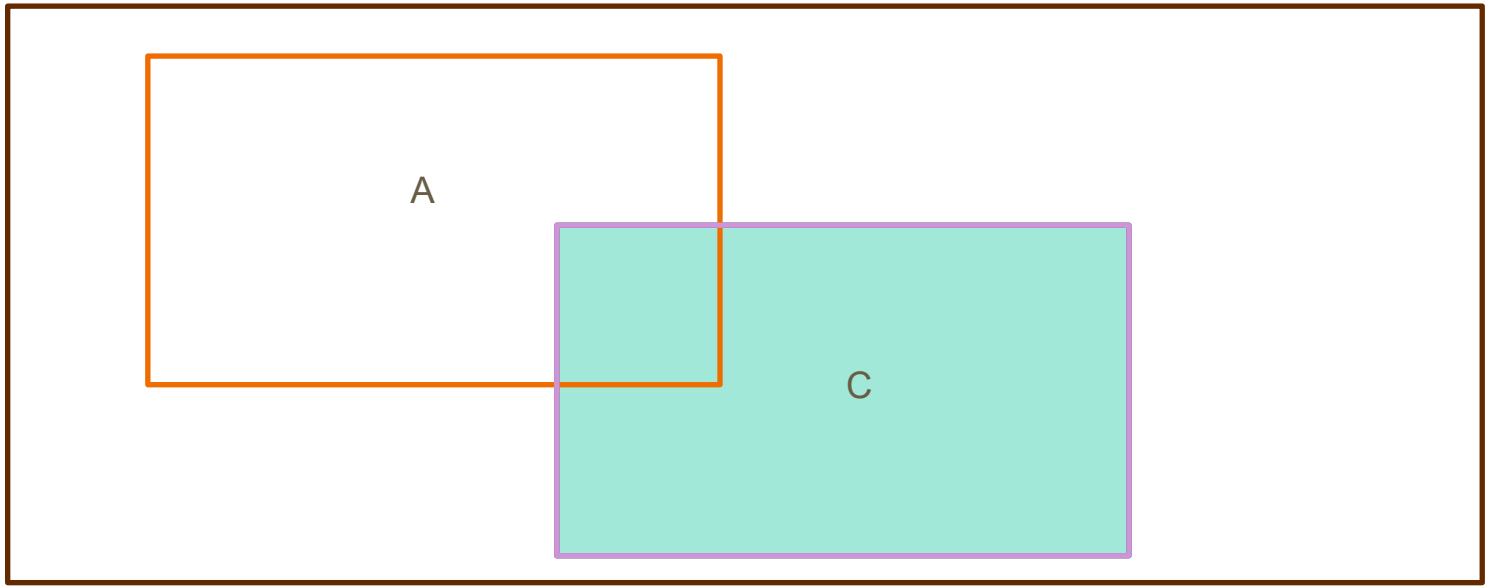
Conditional Probability

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$



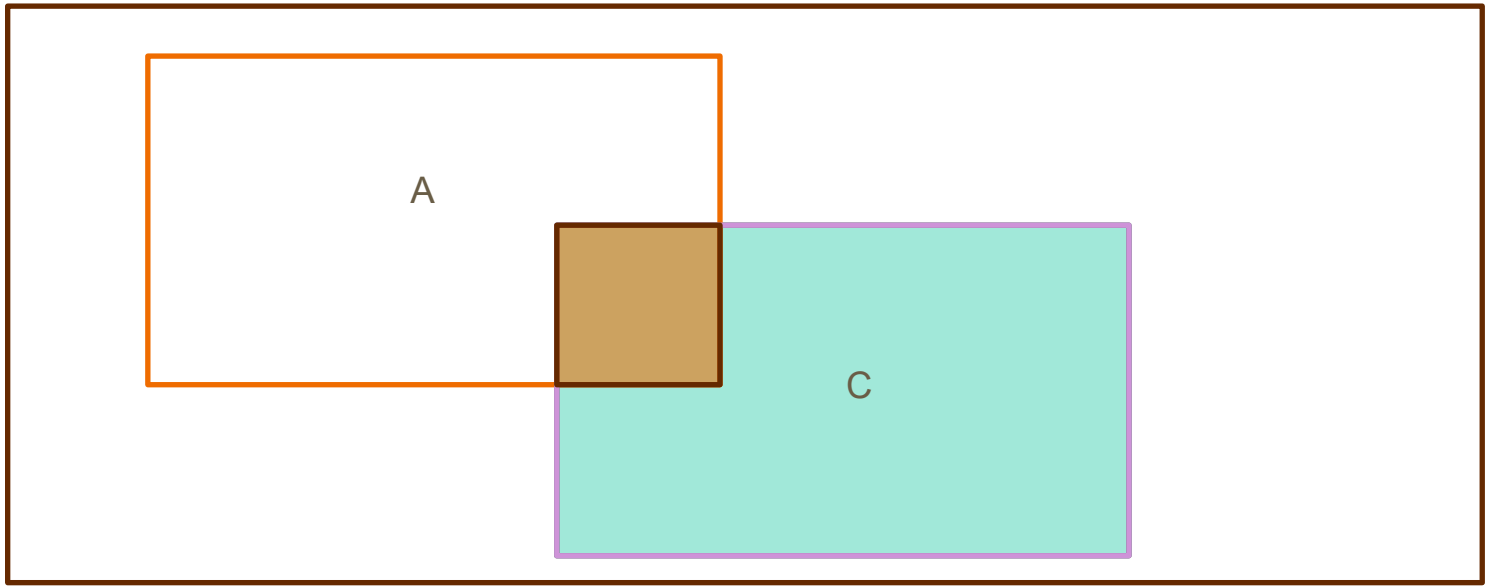
Conditional Probability

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$



Conditional Probability

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$



Conditional Probability

$$P(A \cap C) = ?$$

Conditional Probability

$$P(A \cap C) = P(A)P(C)$$

If A and C are independent

Conditional Probability

$$P(A \cap C) = P(C|A)P(A)$$

If we cannot assume independence

Bayes Theorem

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)}$$

Bayes Theorem

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)}$$

$$P(C|A) = \frac{P(A \cap C)}{P(A)}$$

Bayes Theorem

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)}$$

$$P(C|A) = \frac{P(A \cap C)}{P(A)}$$

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$

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

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

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

$\boldsymbol{P(Xi \mid S_j)}$ is the **PDF** of species j's weights evaluated at weight i

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

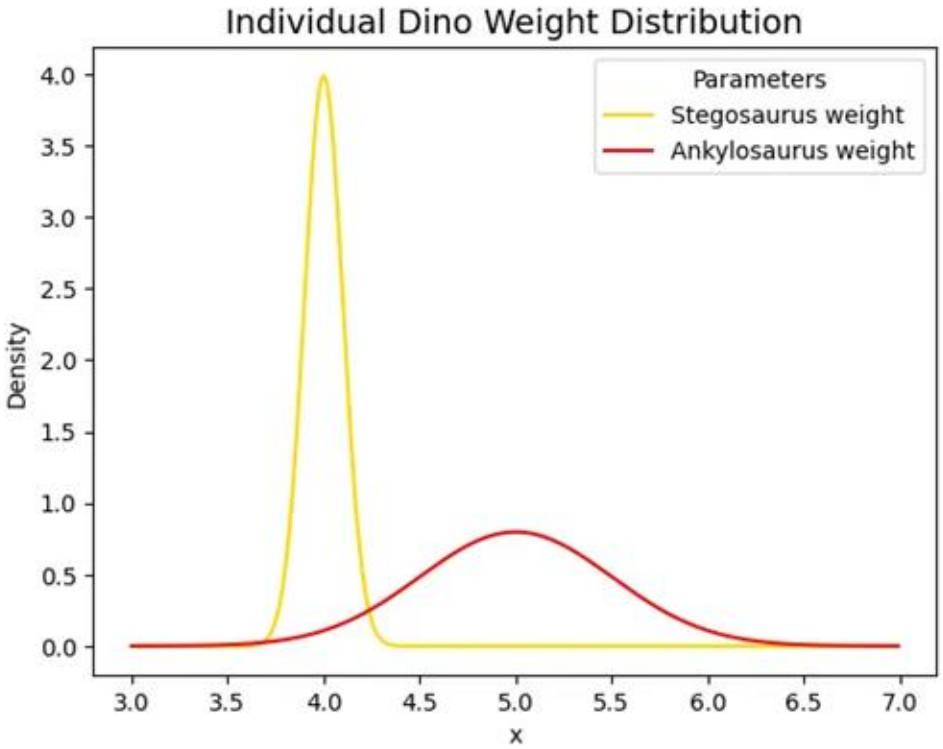
$$P(S_j|X_i) = \frac{P(X_i|S_j)P(S_j)}{P(X_i)}$$

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

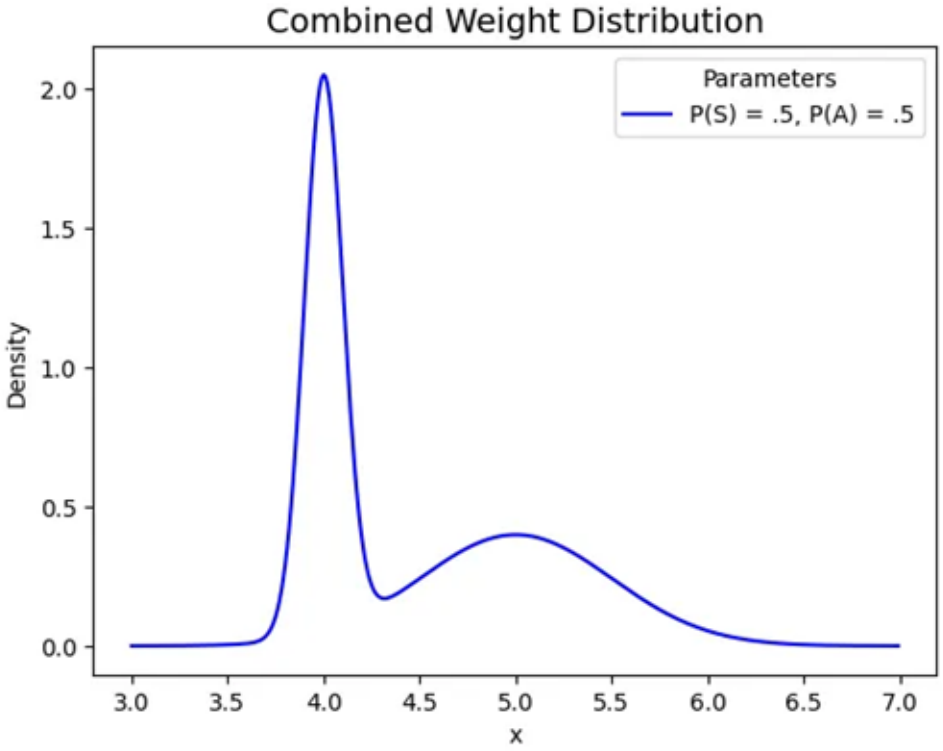
$\boldsymbol{P(X_i|S_j)}$ is the **PDF** of species j's weights evaluated at weight i

- E.g., seeing a Sauropod that weighs 100 tons is way more likely than seeing a Raptor that weighs 100 tons

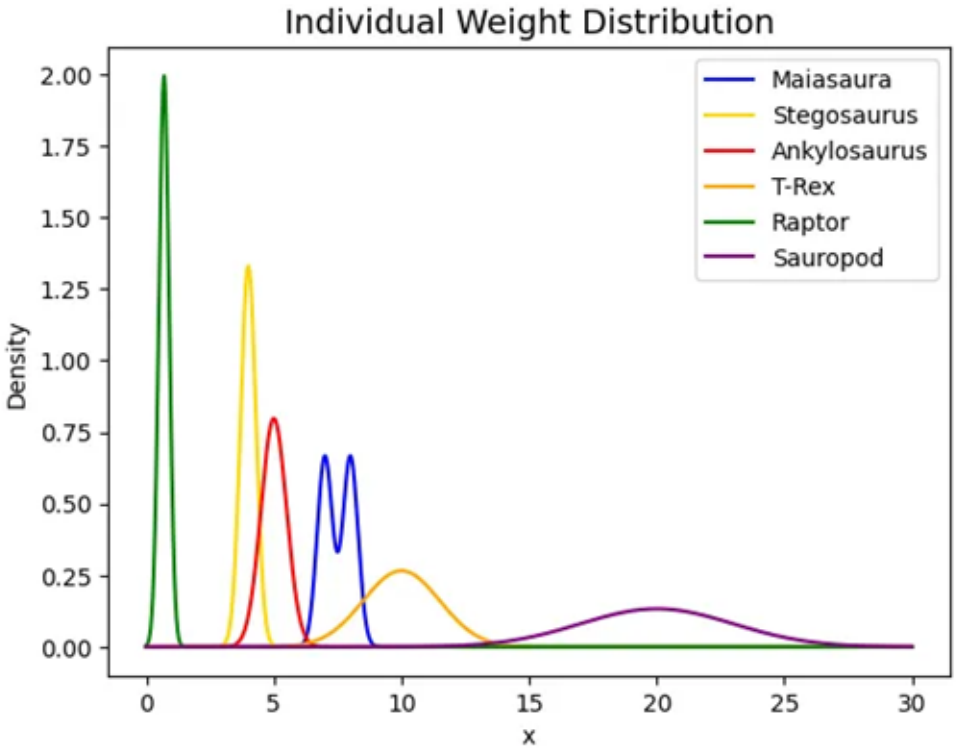
What about $P(X_i)$?

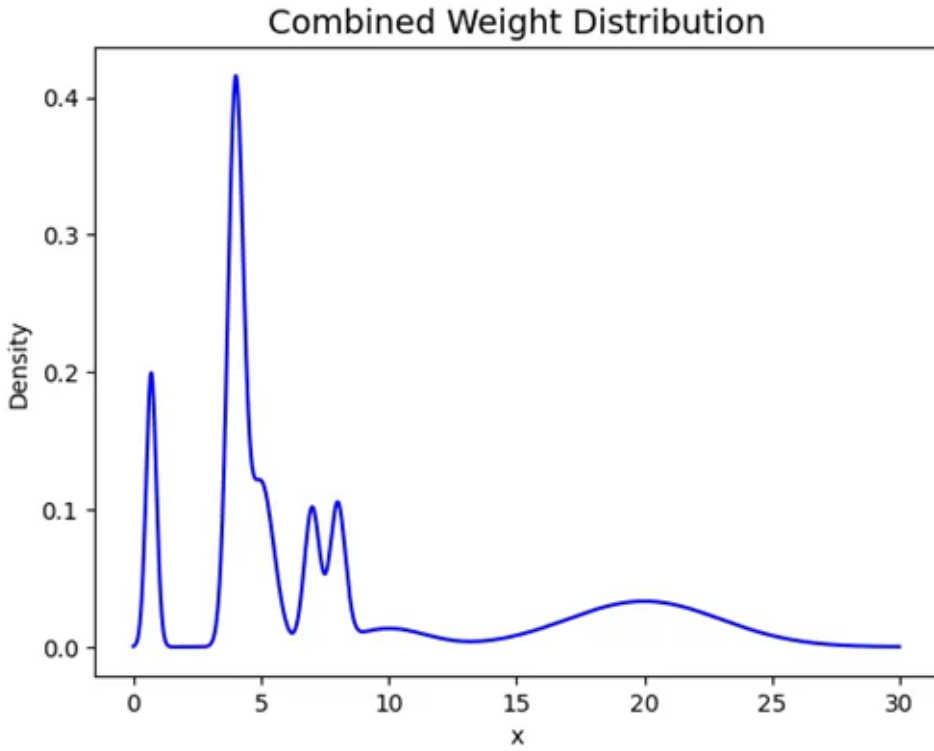


What about $P(X_i)$?



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

$$P(X|S_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

Gaussian Mixture Model

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

$$P(X|S_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

If you know an animal comes from species j, then its weight X follows a normal distribution with mean μ_j and variance σ_j^2

Gaussian Mixture Model

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

$$P(X|S_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

If you know an animal comes from species j, then its weight X follows a normal distribution with mean μ_j and variance σ_j^2

Q: what are the parameters of a GMM?

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?

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

Maximum Likelihood Estimation (intuition)

Example: Assume Bernoulli(p) iid coin tosses.
Success outcome → Heads

Val
H
T
T
H
T

Goal: find p that maximizes the probability of 2 heads and 3 tails

Maximum Likelihood Estimation (intuition)

Example: Assume Bernoulli(p) iid coin tosses

Val
H
T
T
H
T

$$P(\text{having seen the outcomes we saw}) = P(H)P(T)P(T)P(H)P(T)$$

Goal: find p that maximizes the probability of 2 heads and 3 tails

Maximum Likelihood Estimation (intuition)

Example: Assume Bernoulli(p) iid coin tosses

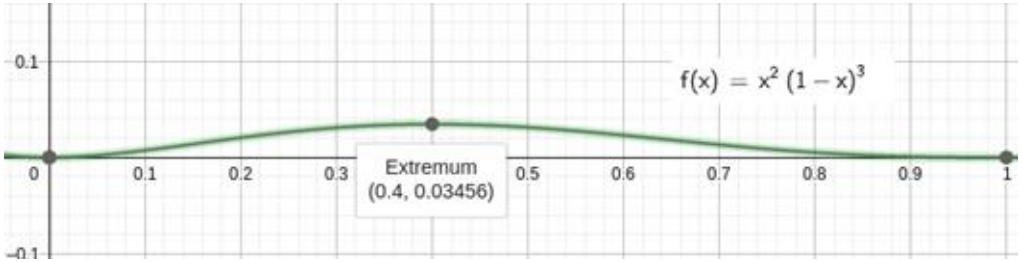
Val
H
T
T
H
T

$$P(\text{having seen the outcomes we saw}) = P(H)P(T)P(T)P(H)P(T)$$
$$= p^2(1-p)^3$$

Goal: find p that maximizes the probability of 2 heads and 3 tails

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:

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

Such that $P(X|S_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$

GMM Clustering

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

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

Such that $P(X|S_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$

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.

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

Such that $P(X|S_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$

Finding the GMM means finding the parameters that uniquely characterize it.

What are these parameters? → $P(S_j), \mu_j, \sigma_j^2 \; \forall k \text{ components}$

Let $\theta = \{\mu_1, \dots, \mu_k, \sigma_1, \dots, \sigma_k, P(S_1), \dots, P(S_k)\}$

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:

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

GMM Clustering

Goal:

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

How do we find the critical points of this function? It's hard to take the derivative of a product!

GMM Clustering

Goal:

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

How do we find the critical points of this function? It's hard to take the derivative of a product! → Log transforms don't change the critical points

GMM Clustering

Goal:

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

How do we find the critical points of this function? It's hard to take the derivative of a product! → Log transforms don't 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

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

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

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

GMM Clustering

Do we have everything we need to solve this?

Expectation Maximization Algorithm

1. Start with random $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$

Expectation Maximization Algorithm

1. Start with random $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$
2. Compute $\boldsymbol{P}(\boldsymbol{S_j}|\boldsymbol{X_i})$for all $\boldsymbol{X_i}$ by using $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$

Expectation Maximization Algorithm

1. Start with random $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$
2. Compute $\boldsymbol{P}(\boldsymbol{S_j}|\boldsymbol{X_i})$for all $\boldsymbol{X_i}$ by using $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$
    - We are calculating the probability that X_i belongs to each species, for all X_i

Expectation Maximization Algorithm

1. Start with random $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$
2. Compute $\boldsymbol{P}(\boldsymbol{S_j}|\boldsymbol{X_i})$for all $\boldsymbol{X_i}$ by using $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$
3. Compute / Update $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$ from $\boldsymbol{P}(\boldsymbol{S_j}|\boldsymbol{X_i})$

Expectation Maximization Algorithm

1. Start with random $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$
2. Compute $\boldsymbol{P}(\boldsymbol{S_j}|\boldsymbol{X_i})$for all $\boldsymbol{X_i}$ by using $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$
3. Compute / Update $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{P}(\boldsymbol{S_j})$ from $\boldsymbol{P}(\boldsymbol{S_j}|\boldsymbol{X_i})$
4. Repeat 2 & 3 until convergence

Random initialization?

Good initialization makes EM more reliable

Random initialization?

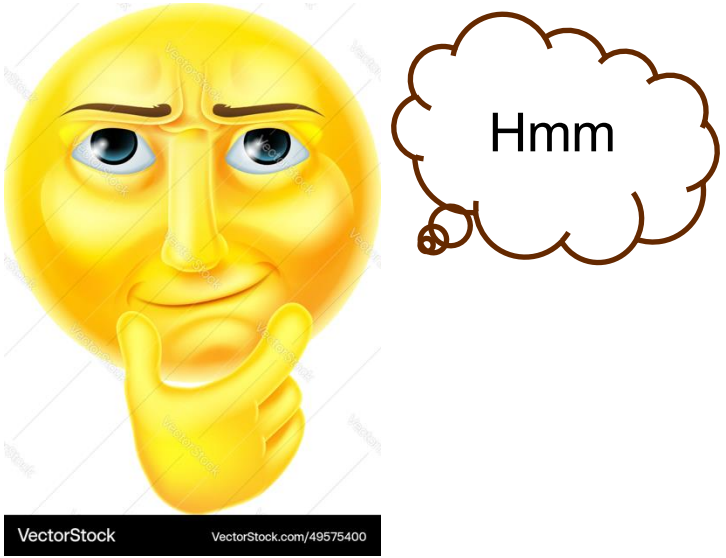
Good initialization makes EM more reliable

What is an algorithm we know which we can use to initialize k clusters and get initial values for $\mu, \Sigma, P(S_j)$?

Random initialization?

Good initialization makes EM more reliable

What is an algorithm we know which we can use to initialize k clusters and get initial values for $\mu, \Sigma, P(S_j)$?



K-means for parameter initialization

1. Run K-means(++) with k clusters

K-means for parameter initialization

1. Run K-means(++) with k clusters
    - Output: $C_1, \dots, C_k$

K-means for parameter initialization

1. Run K-means(++) with k clusters
    - Output: $C_1, \dots, C_k$
2. Initialize parameters
    - μ_j = centroid of cluster j
    - $\Sigma_j = \frac{1}{|C_j|}\sum_{x \in C_j}(x - \mu_j)(x_i - \mu_j)^T$
    - $P(S_j) = \frac{1}{n}|C_j|$ where n is the total number of data points in the dataset

Worksheet d) ->