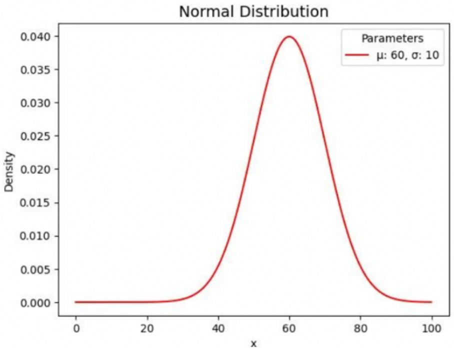
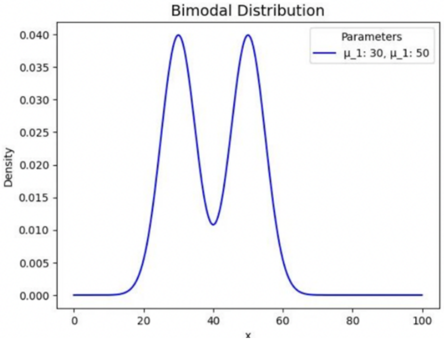
Lecture 7 Soft Clustering

- Given a dataset of weights sampled from N different animals, can we determine which weight belongs to which animal?
- Makes the most sense to provide a probability value that it came from each species from each data point
- $P(S_j|X_i)$
    - Where Sj is species j
    - Where Xi is the ith weight in the data set
    - “the probability that that belongs to species Sj given weight Xi
- Issue: their could be an imbalanced data set w a lot more or a lot less of one species
- Issue: weights vary differently depending on species (each species could have a different weight distribution)





-
- Computig the probability $P(S_j|X_i)$
    - $P(S_j|X_i) = P(X_i|S_j)P(S_j) \div P(X_i)$
    - Where P(Sj) is the prior probability of seeing the species how much more frequently the species occurs in nature could effect this
    - $P(X_i|S_j)$ is the PDF (prob. Density function) of species Sj weights evaluated at weight Xi (answers is it more likely for a rat or an elephant to weight ½ pound)
    - $P(X_i) = \sum_j P(S_j)P(X_i|S_j)$
- Mixture Model
    - X comes from a mixture model w k mixture components if the prob disturbution of X is $P(X) = \sum_j P(S_j)P(X|S_j)$
        - P(Sj) is the mitxutre proportion representing the prob of belonging to Sj
        - P(X | sj) is the probability of seeing x when samping from Sj
    - P(X | Sj) is approximitly the normal distrubtion N (μ, σ)
- Maximum Likelihood Estimation
    - Coin tossing
        - Imagine you are given a dataset of coin tosses and are asked to estimate the parameters that characterize that distubtion
        - MLE : find the parameters that maximized the probability of having seen the data we got
        - Goal : find optimal way to fit distribution to data
        - Plot the likelihood of observing the data at particular points and extract the location that maximizes the likelihood of observing the weights measured

    - and assign it to the mean -> gives you the MLE for the mean of the distribution
    - For SD Plot the likelihood of observing the data at particular points for the sd and extract the location that maximizes the likelihood of observing the weights measured and assign it to the sd
    - Liklyhood != probability
    - Likelyhood strictly refers to the process of finding the optimal value for the mean or SD for a distribution given a set of observed measurements
  - GMM (Gaussian Mixture Model)
    - Goal : to find the GMM that maximized the probability of seeing the data we have seen so far
    - 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.
    - $P(X_i) = \sum_j P(S_j)P(X_i|S_j)$
    - $\prod_i P(X_i) = \prod_i \sum_i P(S_j)P(X_i|S_j$