Expectation Maximization

Pattabiramann Balasubrahmaniam – 110097754

Gladson George – 110091793

Ajay Kanagarethinam Rajakumar – 110082468

Instructor:

Dr. Yasser Alginahi

November 3, 2023

Will I do good?

Determine my success rate

Look at the problem another way

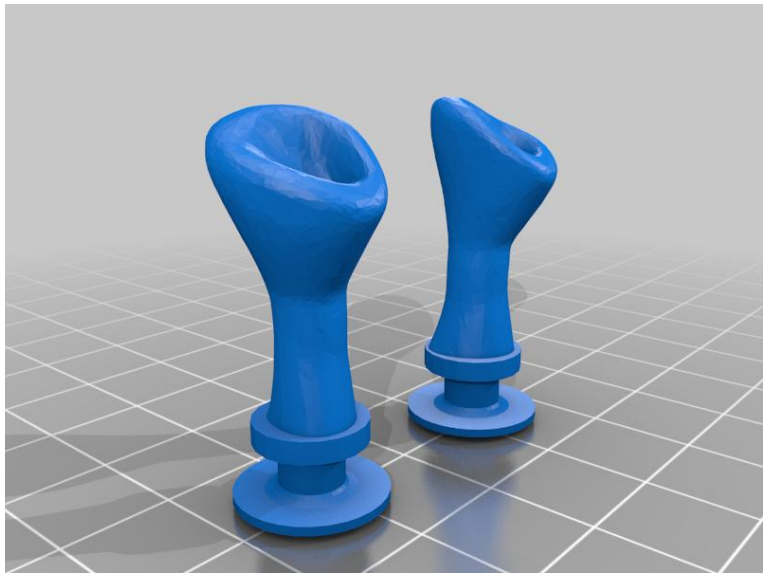
Reviews!

A Real Life Problem

-The 3D Print

The Hunt for 3-D Printers

- Send the model->Model is perfect->Happy customer



Source : https://tinyurl.com/mrmarxx5

Source : https://tinyurl.com/mrmarxx5

The Hunt for 3-D Printers

- First Shop : Shop A = 188/200 Reviews
- Second Shop : Shop B = 48/50 Reviews
- Third Shop : Shop C = 10/10 Reviews

- Shop A : 93% success
- Shop B : 96% success
- Shop C : 100% success

Let's Apply!

- Laplace adds : +2 to total, +1 to the success
- Shop A : original = 188/200 = 94%
- New : 189/202 = 93.5%
- Shop B : original = 48/50 = 96%
- New : 49/52 = 94.2%
- Shop C : original = 10/10 = 100%
- New : 11/12 = 91.6%

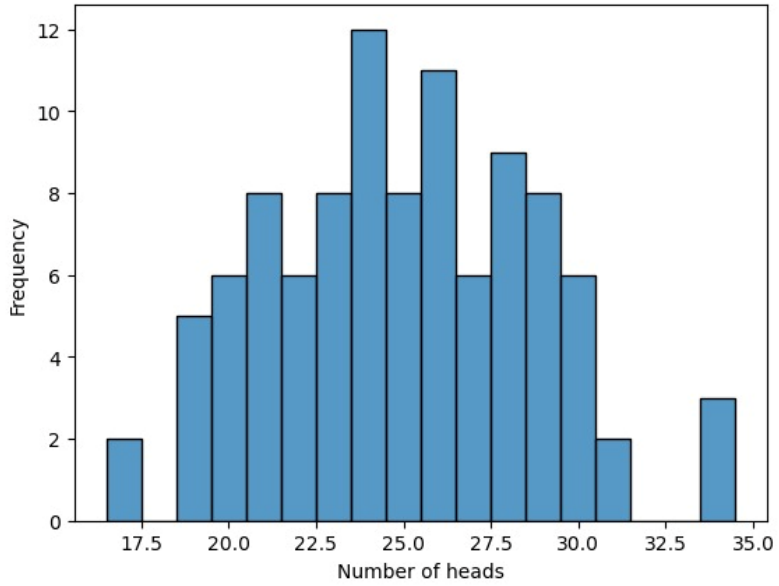
A Happy Customer!

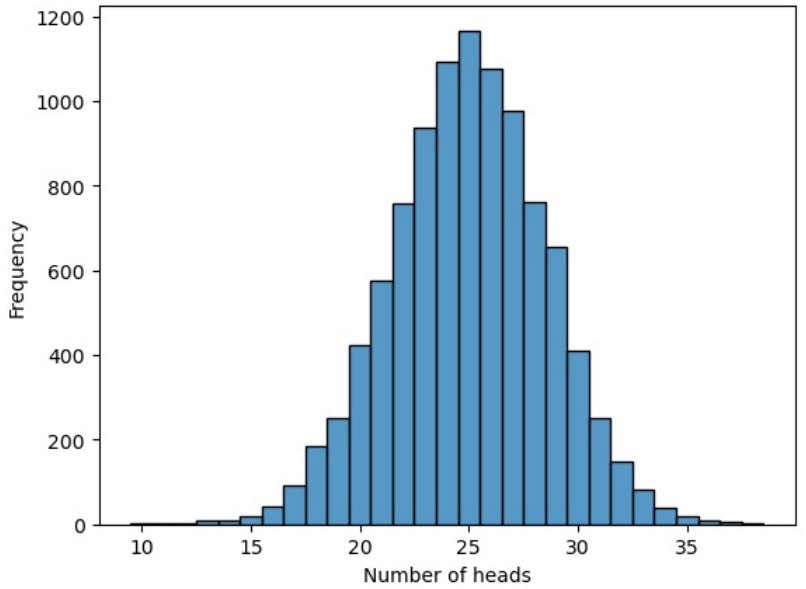
Laplace's Rule of Succession

- In a given sample space and success rate,
- Introduce **one success and one failure** observation in the sample space.
- Calculate the **new success rate**, which should be a close approximation of your own observation.
- If the change in the rates is significant, the sample space is not **mature enough** to be trusted.

A Simpler Model:

-The Coin Flip





University of Windsor

Central Limit Theorem

- "Let X1,X2,…,Xn be a sample from a population having mean μ and standard deviation σ. For n large, the sum X1+X2+⋯+Xn will approximately have a ***normal distribution*** with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.[1]"
- Normal distributions have an area of $\sqrt{\pi}$, probabilitiy adds to 1.
- The **density of the probability with the sample space**,

then, can be written with:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Source, Galton Board: https://tinyurl.com/yeyrnhjc

Probability Densities

- The central limit theorem gives us a model for probabilities.
- Let us make the variable continuous!
- That gives us a problem : no sample spaces.
- So, we define a function for **probability *density.***

Probability density function

- In the coin flip example, the possibility of each number of heads happening is visualized.
- Therefore, taking the area over each ‘rectangle’ gives us subsequent probability.
- In a continuous variable probability density function, we use integration to get the probability.
- Therefore, Area of probability density function = 1(total probability).

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

A Quick Review:

- The same experiment repeated and value of probability plotted against sample space is a normal distribution, mean μ and standard deviation σ.
- The μ is the *most likely* and the average value, and thus called the **expected value**.
- For continuous variables, the sample space is written as a function called a probability density function.

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Attempt to Arrive at a Rigorous Approach

Expectation Maximization

- It involves the process of maximizing the parameter of an assumed probability density function in an iterative process to find and fit a variable to a cluster of values arranged like a Gaussian mixture.
- The Gaussian mixture follows the equation:

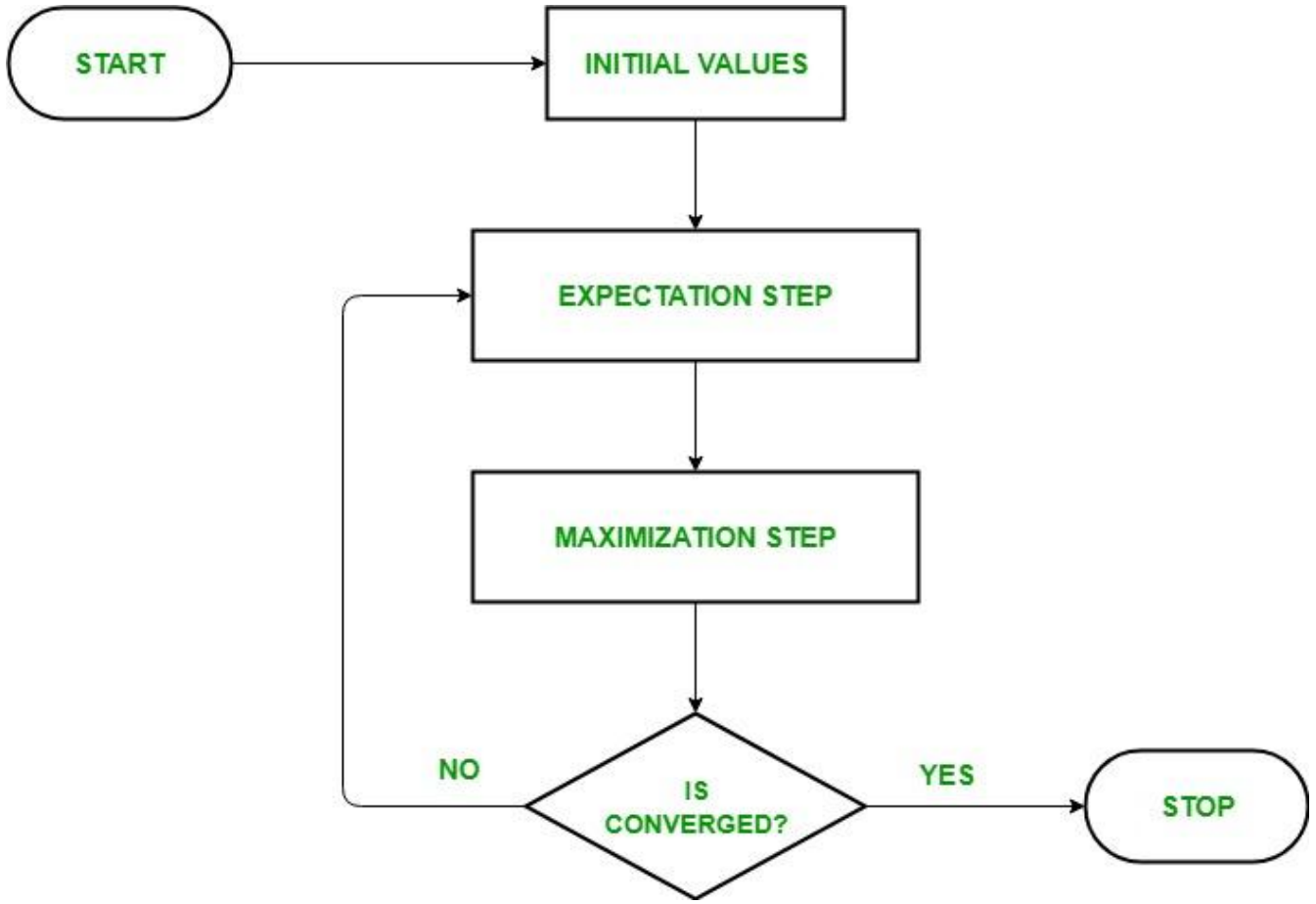
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Where μ and σ are the parameters of the function f(x).
- We define a likelihood function, where x is the parameter, i.e., L(μ , σ)
- Both the functions are equivalent as both of the arguments are random variables.

Expectation Maximization(contd…)

- If we take the log of the likelihood function, we get the function as a polynomial, letting us study the change much easier.
- The First Iteration: Assume parameters and create a likelihood function.
- Maximize the argument of the function and get the new parameters.
- Repeat till the parameters don’t change.

Review of the process



Source : https://tinyurl.com/bdedvhea

The General Process - Definitions

- The probability density function is given by:
- $P(x_i) = \sum_{c=1}^{n} P_c . Norm(x_i, \mu_c, \sigma_c)$

$Where,$
$x = observable\ random\ variable,$
$m = size\ of\ dataset,$
$n = number\ of\ clusters,$
$c = cluster\ number.$
$For\ each\ cluster,$
$\mu_c = mean,$
$\sigma_c = variance,$
$P_c = Probability\ of\ cluster$
$Norm = normal\ function.$

Step 1: Allocating Parameters

- Let the latent variable be z, such that:
- $P(z = c) = P_c$
- Then, we can write:
- $Norm(x_i, \mu_c, \sigma_c) = P(x_i | z = c)$
- To find z,
- $P(z = c|x_i) = \frac{P(z=c).P(x_i|z=c)}{P(x_i)}$
- $P(z = c|x_i) = \frac{P_c.Norm(x_i,\mu_c,\sigma_c)}{\sum_{c=1}^{n} P_c.Norm(x_i,\mu_c,\sigma_c)} = P_{ic}$

$Where,$
$x = observable\ random\ variable,$
$z = latent\ random\ variable$
$m = size\ of\ dataset,$
$n = number\ of\ clusters,$
$c = cluster\ number.$
$For\ each\ cluster,$
$\mu_c = mean,$
$\sigma_c = variance,$
$P_c = Probability\ of\ cluster$
$Norm = normal\ function.$

Step 2: Maximization of Parameters

- Cluster size can be defined as:
- $m_c = \sum_{i=1}^{m} P(z = c|x_i)$
- $P_c = \frac{m_c}{m}$
- Then,
- $\mu_c = \frac{1}{m_c} \sum_{i=1}^{m} (P(z = c|x_i).x_i)$
- And,
- $\sigma_c = \frac{1}{m_c} \sum_{i=1}^{m} (P(z = c|x_i).(x_i - \mu_c)^2)$

$Where,$
$x = observable\ random\ variable,$
$z = latent\ random\ variable$
$m = size\ of\ dataset,$
$n = number\ of\ clusters,$
$c = cluster\ number.$
$For\ each\ cluster,$
$\mu_c = mean,$
$\sigma_c = variance,$
$P_c = Probability\ of\ cluster$
$Norm = normal\ function.$

Advantages of EM

- **Unsupervised Learning**
- **Handles Incomplete Data**
- **Guaranteed Convergence**
- **Commendable Flexibility**
- **Soft Assignments – Probabilistic Information**

Disadvantages of EM

- **Initialization Sensitivity**
- **Slow Convergence makes it more expensive**
- **Not robust to noisy data**
- **Singularities and non-invertible covariances**
- **Local Optima convergence**
- **Computationally Expensive**
- **Overfitting**

Application of EM

- **Anomaly Detection**:
  - EM can be employed in anomaly detection to model the distribution of normal data points. Any data point that deviates significantly from this model may be considered an anomaly.
- **Gaussian Mixture Model (GMM) in Clustering**:
  - EM is commonly used for clustering data into multiple groups. A classic example is clustering customer purchasing behavior in marketing. You can use EM to identify different segments of customers based on their buying habits.
- **Image Compression**:
  - EM can be applied to image compression. For instance, in the context of video coding, you can use EM to model the distribution of pixel intensities in an image, allowing for efficient compression based on the estimated model parameters.

Appendix

- If the user needs more information, can refer to these youtube videos:
  - Expectation Maximization Algorithm | Intuition & General Derivation by Machine Learning & Simulation [2]
  - Expectation Maximization for the Gaussian Mixture Model | Full Derivation by Machine Learning & Simulation[3]
  - EM Algorithm for Latent Variable Models by Inside Bloomberg[4]
  - Lecture 14 - Expectation-Maximization Algorithms | Stanford CS229: Machine Learning (Autumn 2018) Stanford Online[5]
- Sklearn Gaussian Mixture Models[6]
- J. A. Bilmes and Others, 'A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models', 1998.[7]

References

- [1] “Central limit theorem,” Central Limit Theorem - an overview | ScienceDirect Topics, https://www.sciencedirect.com/topics/mathematics/central-limit-theorem (accessed Nov. 3, 2023).
- [2] Machine Learning & Simulation. Expectation Maximization Algorithm | Intuition & General Derivation. (Apr. 8, 2021). Accessed: Nov. 3, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=MujtZ4A23t8&ab_channel=MachineLearning%26Simulation
- [3] Machine Learning & Simulation. Expectation Maximization for the Gaussian Mixture Model | Full Derivation. (Apr. 9, 2021). Accessed: Nov. 3, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=ZI98eg3CRls&ab_channel=MachineLearning%26Simulation
- [4] Inside Bloomberg. 27. EM Algorithm for Latent Variable Models. (July 11, 2018). Accessed: Nov. 3, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=lMShR1vjbUo&ab_channel=InsideBloomberg
- [5]Stanford Online. Lecture 14 - Expectation-Maximization Algorithms | Stanford CS229: Machine Learning (Autumn 2018). (Apr. 17, 2020). Accessed: Nov. 3, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=rVfZHWTwXSA
- [6] “Gaussian mixture models”, [Online]. Available: https://scikit-learn.org/stable/modules/mixture.html. [Accessed: 3-Nov.-2023].
- [7] J. A. Bilmes and Others, ‘A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models’, 1998. [Accessed: 3-Nov.-2023]