

# Finite Mixture Model: Multinomial

RECAP

# SESSION I:

- General concepts of probability distribution
- Discrete Probability Distribution:
  - Bernoulli, Binomial, Categorical, Multinomial
- Continuous Probability Distribution:
  - Gaussian, Beta, Dirichlet
- Bayes Theorem:
  - Likelihood, Prior, Posterior

# SESSION II

- Conjugate Distribution:
  - Beta-Binomial Distribution
  - Dirichlet-Multinomial Distribution
  - Normal with NormalGamma
- Sampling
  - Markov Chain
  - Markov Chain and Transition Matrix
  - Markov Chain Monte Carlo
    - Gibbs Sampling

# OUTLINE:

- Representing knowledge through graphical models
- Mixture Model:
  - Introduction
  - Multinomial Mixture Model
    - Known parameters
    - Unknown parameters
      - Posterior
      - Fully Collapsed

# Representing knowledge through graphical models

- Nodes correspond to random variables.
- Edges represent statistical dependencies between the variables.

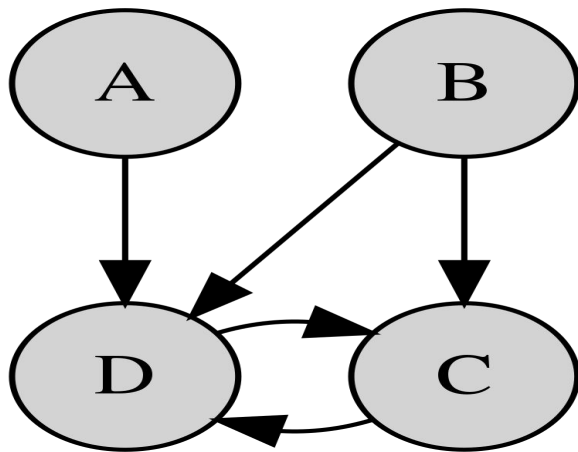


Figure 1: An example of a graphical model. Each arrow indicates a dependency. In this example: D depends on A, B, and C; and C depends on B and D; whereas A and B are each independent.  
[Image Source: Wikipedia]

# Mixture models

## Clustering:

- Involves grouping of similar objects into a set known as cluster.
- Applications: creating newsfeeds, customer segmentation, social network analysis and so on.



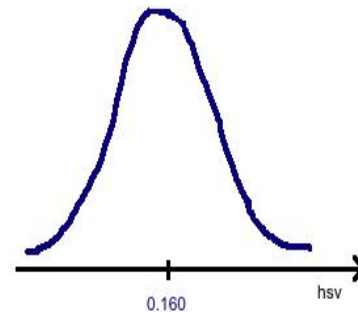
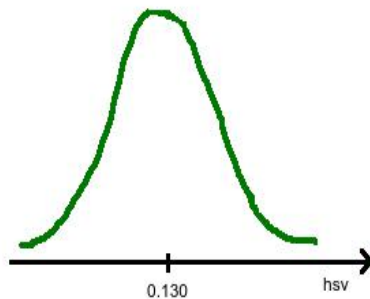
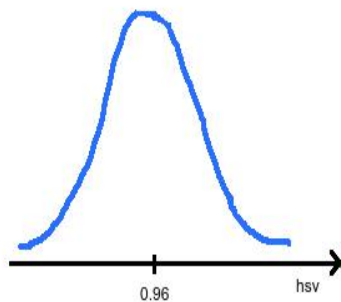
What is the picture representing (sky, river or forest)?



Picture from: Unsplash

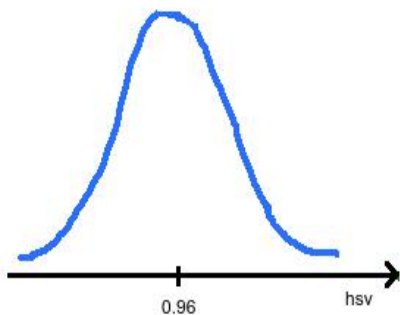


# HSV color distribution

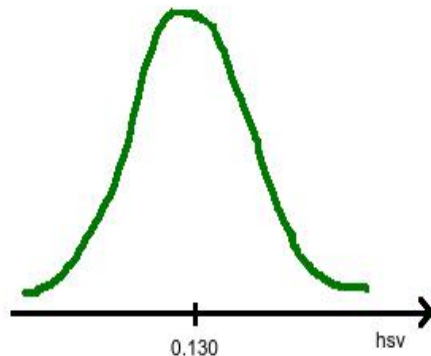


# Weight assignment

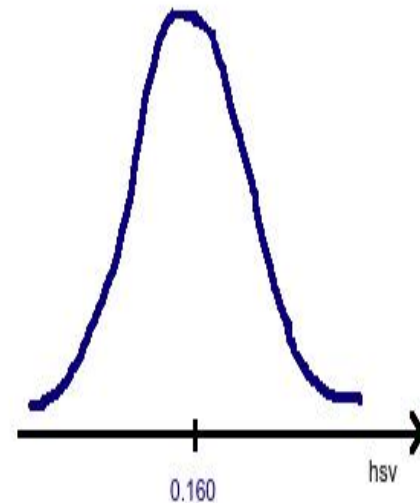
$$\Pi_k = [0.2, 0.3, 0.5]$$



$P(x | \mu_1)$



$P(x | \mu_2)$



$P(x | \mu_3)$

- Each value of  $\Pi_k$  lies between 0 and 1.
- The sum of all the values of  $\Pi_k$  must be equal to 1.

# Combining Multiple Mixture Models

Combining all the distributions, we get:

$$p(x|\theta) = \pi_1 P(x|\mu_1) + \pi_2 P(x|\mu_2) + \pi_3 P(x|\mu_3)$$

Here,  $\theta = \{\pi_k, \mu_k\}$  are the parameters.

Therefore,  $p(x|\theta) = \sum_{k=1}^K \pi_k P(x|\mu_k)$

# Mixture Distribution

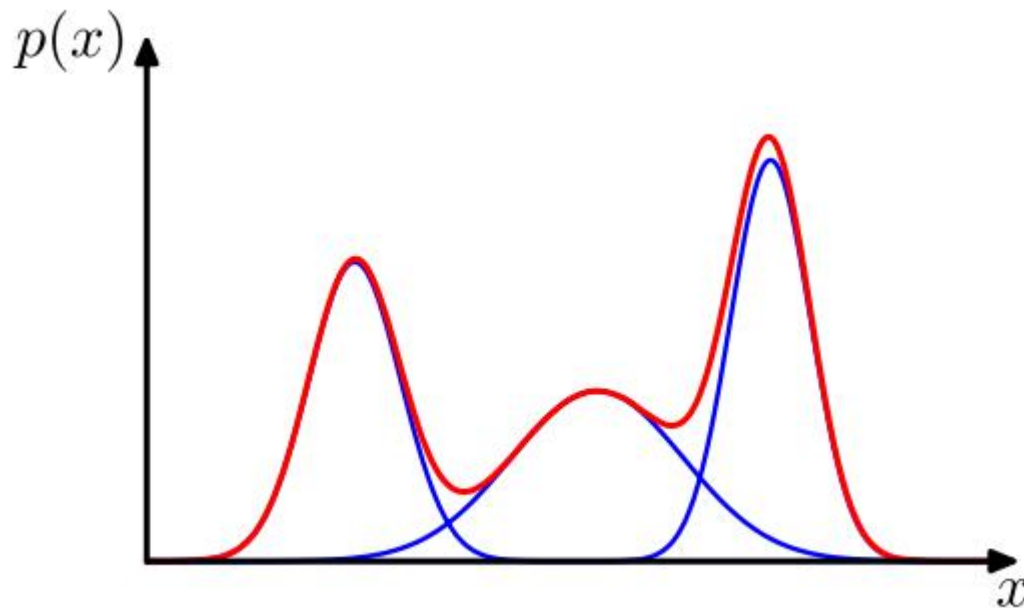


Figure 3: Illustration of mixture distribution in one dimension showing three distributions (each scaled by a coefficient) in blue and their sum in red

# Bimodal density function

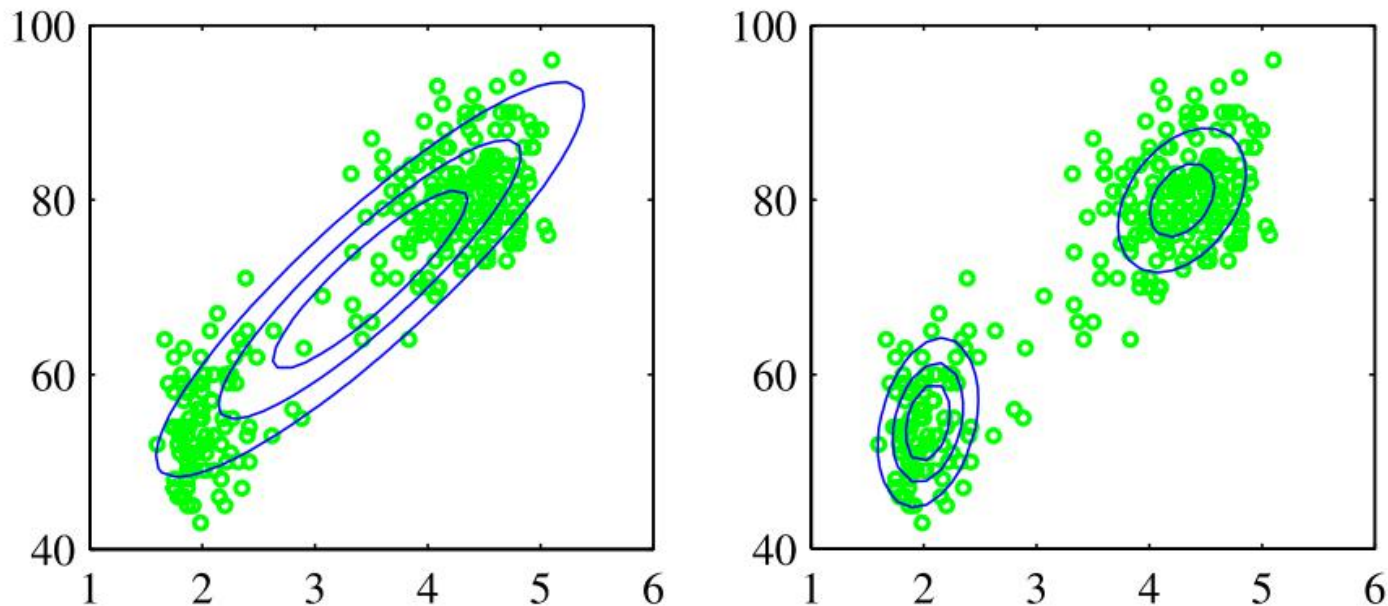


Figure 2: Left - single distribution representing data, Right- two distribution to model different set of data

# Mixture models

- Probabilistic model
- Allows soft clustering
- Can capture even oddly shaped clusters
- Allows bimodal density functions

# Latent Variable Models

Given,  $X = \{x_1, x_2, \dots, x_n\}$ , we assume that

$$Z = \{z_1, z_2, \dots, z_n\}$$

in which the corresponding latent variable indicates the mixture component

- $Z$ 's are like switches to indicate which component was used.

$$p(x_i|\theta) = \sum_k p(x_i|\theta_k) p(z_i = k)$$

$$p(x_i|\theta) = \sum_k p(x_i|\theta_k) \pi_k$$

Here,  $\pi_k$  is the mixture component.



# Posterior Dirichlet Categorical Conditional Distribution

- Find the probability of belonging to observed class  $k$ , i.e,  $p(\mathbf{z}_i=\mathbf{k})$ .

**Example: Classes** = [Math, English]

**Biasedness ( $\theta$ ):** Math: English = 0.25: 0.75, i.e,  $\theta = [0.25, 0.75]$

**Random data points:**  $X_1, X_2, X_3, X_4$

**To find:**  $P(Z=k|\theta)$

$$P(Z_{X1} = \text{Math} \mid [0.25, 0.75]) = 1$$

$$P(Z_{X2} = \text{Math} \mid [0.25, 0.75]) = 0$$

$$P(Z_{X3} = \text{Math} \mid [0.25, 0.75]) = 0$$

$$P(Z_{X4} = \text{Math} \mid [0.25, 0.75]) = 0$$



# Posterior Dirichlet Categorical Conditional Distribution

Using Dirichlet Categorical, we can compute the probability of observing class  $k$  having already observed counts  $(\mathbf{c}_1, \dots, \mathbf{c}_K)$ .

We will use an indicator variable  $\mathbf{z} = \mathbf{k}$  to indicate that the observed class is  $\mathbf{k}$ :

$$p(\mathbf{z} = \mathbf{k} | \mathbf{c}, \alpha) = \frac{p(\mathbf{z} = \mathbf{k} | \alpha)}{p(\mathbf{c} | \alpha)} = \frac{p(\mathbf{z} = \mathbf{k} | \theta) p(\theta | \alpha)}{\int p(\mathbf{c} | \theta) p(\theta | \alpha) d\theta}$$

Using Dirichlet prior,

$$p(\mathbf{z} = \mathbf{k} | \mathbf{c}, \alpha) = \frac{\frac{\tau(A)}{\prod_i \tau(\alpha_i)} \frac{[\tau(c_k + \alpha_k + 1)] \prod_{i \neq k} \tau(c_i + \alpha_i)}{\tau(C + A)}}{\frac{\tau(A)}{\prod_i \tau(\alpha_i)} \frac{\prod_i \tau(c_i + \alpha_i)}{\tau(C + A - 1)}}$$

# Posterior Dirichlet Categorical Conditional Distribution

Here,  $C-1$  is the total number of items before adding new  $\mathbf{z}$ . Simplifying the above equation, we get,

$$p(z = k|c, \alpha) = \frac{\tau(c_k + \alpha_k + 1)}{\tau(c_k + \alpha_k)} \frac{\tau(C + A - 1)}{\tau(C + A)}$$

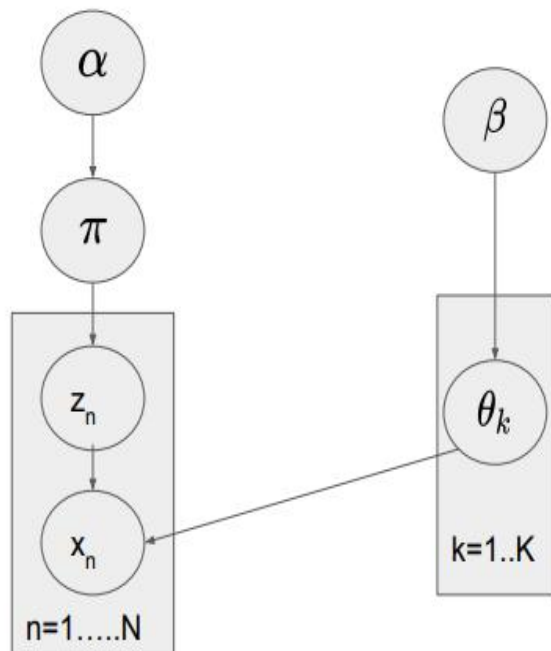
Using  $\tau(N+1) = N\tau(N)$ ,

$$= \frac{c_k + \alpha_k}{C + A - 1} \quad \text{[equation 1]}$$

This says that the probability of a new data point being assigned the class  $\mathbf{k}$  is proportional to  $\mathbf{c_k} + \boldsymbol{\alpha_k}$ .

- Thus, the Dirichlet exhibits the **rich-gets-richer** property.

# Modeling Mixture Component: Finite Mixture Model



To model mixture components  $\pi$ , we use:

$$\begin{aligned} p(X, Z, \pi | \theta, \alpha) &= p(X, Z | \theta, \pi) p(\pi | \alpha) \\ &= p(X | Z, \theta) p(Z | \pi) p(\pi | \alpha) \end{aligned}$$

Here,

- $p(\pi | \alpha)$  gives distribution over mixture weights

$$p(\pi | \alpha) \sim \text{Dir} \left( \frac{A}{K}, \dots, \frac{A}{K} \right)$$

- $p(Z | \pi) \sim \text{Categorical}(\pi)$
- $p(\theta_k | \beta) \sim \text{Beta}(\beta)$

We can integrate out mixture proportions ' $\pi$ ' using Dirichlet categorical distribution:

$$\begin{aligned} p(X, Z|\theta, \alpha) &= \int p(X, Z, \pi|\theta, \alpha) d\pi \\ &= p(X|Z, \theta) \int p(Z|\pi) p(\pi|\alpha) d\pi \\ &= p(X|Z, \theta) p(Z|\alpha) \\ &= p(Z|\alpha) \prod_i p(x_i|Z, \theta) \\ &= p(Z|\alpha) \prod_i p(x_i|\theta_{z_i}) \end{aligned}$$

# Gibbs Sampling for Finite Mixtures

As conditional distribution converges to joint distribution in the limit,

$$\begin{aligned}\text{we can write, } p(x_i, z_i | X_{-i}, Z_{-i}, \theta, \alpha) &= \frac{p(X, Z | \theta, \alpha)}{p(X_{-i}, Z_{-i} | \theta, \alpha)} \\ &= \frac{p(X | Z, \theta)}{p(X_{-i} | Z_{-i}, \theta)} \frac{p(Z | \alpha)}{p(Z_{-i} | \alpha)} \\ &= p(x_i | z_i, \theta) \frac{p(Z | \alpha)}{p(Z_{-i} | \alpha)}\end{aligned}$$

Thus,  $p(x_i, z_i | X_{-i}, Z_{-i}, \theta, \alpha) = p(x_i | z_i, \theta) p(z_i | \alpha)$

# Multinomial Mixture Model

- Mixture components are multinomial distribution

# Problem Statement: Multinomial

Suppose 2 dice are rolled with the following parameters.

```
params = {0: { $\pi$ 1: 0.2,  $\theta$ 1:0.1},  
          1: { $\pi$ 2: 0.8,  $\theta$ 2:0.9}}
```

The data generated has two mixture components for each of the given parameters.

# Cases

1. We know the value of ' $\theta$ '.
2. We do not know the value of ' $\theta$ '.



## Case 1 (' $\theta$ ' known)

1. Randomly assign values to cluster, i.e,  $z_i = k$  with uniform probability.

2. For each  $i$ : i) Remove data point  $(x_i, z_i)$

ii) Count the number of data points,  $c_k$ , in class  $k$ , given by:

$$c_k = |\{z_i = k \mid z_i \in Z\}|$$

iii) Compute multinomial distribution for each cluster  $k$ :

$$p_k(x_i | \theta_k) = \frac{C!}{\prod_i c_i!} \prod_i \theta_k^{c_i}$$

Here,  $c_i$  is the count of each element within the cluster.

iv) Use Dirichlet categorical conditional distribution to compute the probability of observing class  $k$  having already observed counts as in equation (1):

$$p(z = k|c, \alpha) = \frac{c_k + \alpha_k}{C + A - 1}$$

v) Add data point  $(x_i, z_i)$  back by sampling  $z_i = k$ , using,

$$z_i = k \sim p(z = k|c, \alpha)p_k(x_i|\theta_k)$$

where,  $p_k(x_i|\theta_k)$  and  $p(z = k|c, \alpha)$  are obtained from equation (iv) and (v) respectively.

Mixture proportions can be estimated at the end.

## Case 2 (' $\theta$ ' unknown) (Finding parameters with posterior distribution)

1. Randomly assign values to cluster, i.e,  $z_i = k$  with uniform probability.

2. For each  $i$ : i) Remove data point  $(x_i, z_i)$ .

ii) Count the number of data points,  $c_k$ , in class  $k$ , given by:

$$c_k = |\{z_i = k \mid z_i \in Z\}|$$

iii) Estimate multinomial parameters  $\theta_k$  for each cluster  $k$  using:

$$p(\theta|c, \beta) = \frac{\tau(C+B)}{\prod_i \tau(c_i + \beta_i)} \prod_i \theta_i^{c_i + \beta_i - 1}$$

iv) Compute multinomial distribution for each cluster  $k$ :

$$p_k(x_i|\theta_k) = \frac{C!}{\prod_i c_i!} \prod_i \theta_i^{c_i}$$

Here,  $c_i$  is the count of each element within the cluster.

v) Use Dirichlet categorical conditional distribution to compute the probability of observing class  $k$  having already observed counts as in equation (1),

$$p(z = k|c, \alpha) = \frac{c_k + \alpha_k}{C + A - 1}$$

vi) Add data point  $(x_i, z_i)$  back by sampling  $z_i = k$ , using,  
$$z_i = k \sim p(z = k|c, \alpha)p_k(x_i|\theta_k)$$

where,  $p_k(x_i|\theta_k)$  and  $p(z = k|c, \alpha)$  are obtained from equation (iv) and (v) respectively.

Mixture proportions can be estimated at the end.

# Sample code

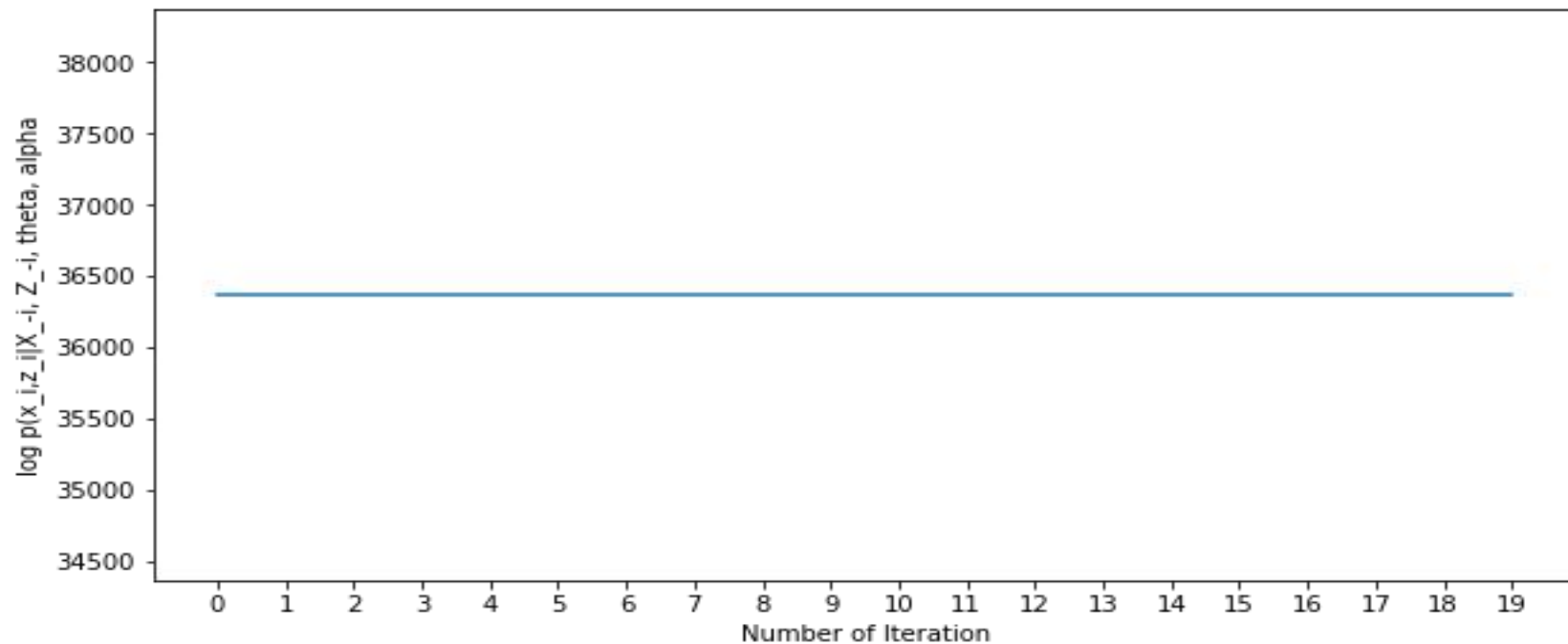
- Finite Mixtures with multinomial prior: (2 mixture components)

<https://colab.research.google.com/drive/1zrtdCgLfVS6gUtbC4pDgplLGp-4gyEAs>

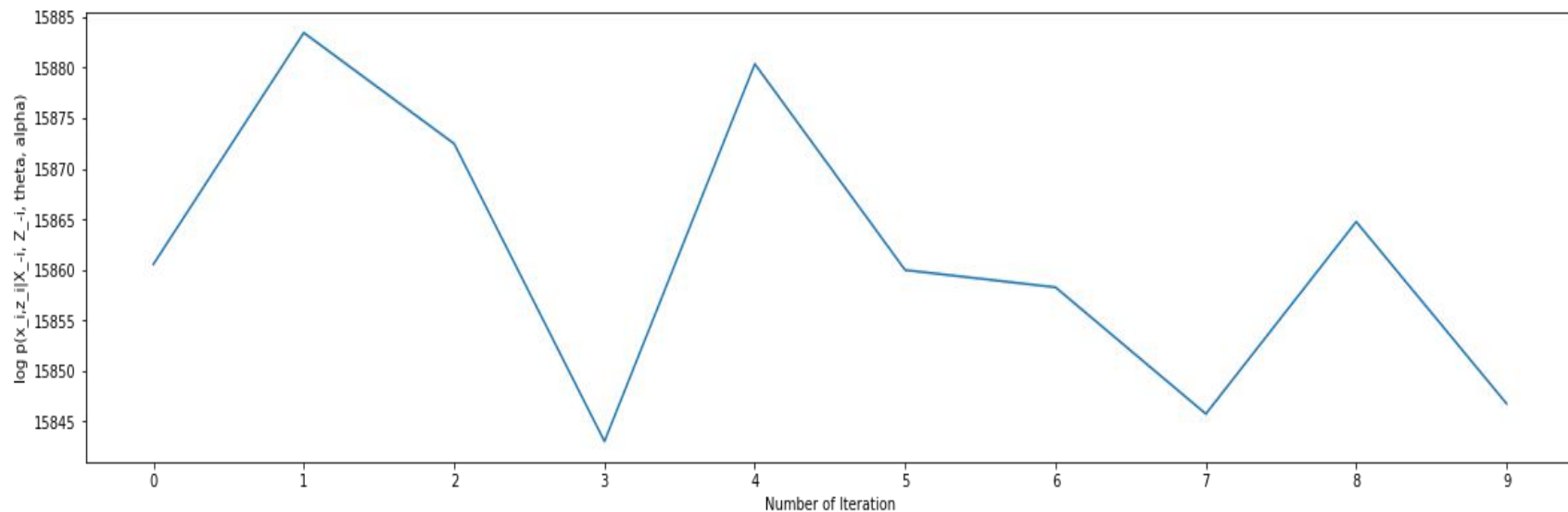
- Finite Mixtures with multinomial prior: (6 mixture components)

<https://drive.google.com/file/d/1MPwjr8DZk1b2Kxs4IsflhUN0xkrJN56g/view?usp=sharing>

Known parameters (2 mixture components) (after 20 iterations)



## Unknown parameters (3 mixture components) (after 10 iterations)



# Fully Collapsed Gibbs Sampling

- **Multinomial Mixture Model**

The within class parameters  $\theta_k$  can be integrated out using the Dirichlet-multinomial distribution

Thus, instead of

$$p(x_i | \theta_{zi})$$

we need to compute

$$\frac{p(X | Z, \beta)}{p(X_{-i} | Z_{-i}, \beta)}$$

Here, 'Z' is the mixture component, ' $\beta$ ' is the shape parameter, ' $\theta$ ' is the biasedness.



# Fully Collapsed Sampler

- **Multinomial Mixture Model**

If we assume that for  $(\mathbf{x}_i, \mathbf{z}_i)$ , the target distribution class is  $\mathbf{k}$ , i.e,  $\mathbf{z}_i = \mathbf{k}$ :

$$\frac{p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\beta})}{p(\mathbf{X}_{-i} | \mathbf{Z}_{-i}, \boldsymbol{\beta})} = \frac{\prod_k p(\mathbf{X}_k | \boldsymbol{\beta})}{\prod_k p(\mathbf{X}_{k,-i} | \boldsymbol{\beta})} = \frac{p(\mathbf{X}_k | \boldsymbol{\beta})}{p(\mathbf{X}_{k,-i} | \boldsymbol{\beta})}$$

Here,  $\mathbf{X}_k = \{\mathbf{x}_i \in \mathbf{X} \mid \mathbf{z}_i = \mathbf{k}\}$ . If we assume  $\mathbf{x}_i$  is a side of the dice, using the Dirichlet prior  $\boldsymbol{\beta}$  in the Dirichlet-multinomial distribution, we can estimate the within class likelihood  $\mathbf{p}(\mathbf{X}_k | \boldsymbol{\beta})$ .

# Fully Collapsed Sampler

## Derivation: Multinomial Mixture Model

Suppose we are tossing a dice with two sides and our trial=4. Let us consider one side of the dice be denoted by 'H'(head) and another side be denoted by 'T'(tail). If we consider total trials and experiments in one cluster, then,

$$X_k = \{6H5T, 3H4T, 10H9T, 1H1T\}$$

$$\text{Let } x_j = 6H5T$$

$$\text{So, } X_{k,-j} = \{3H4T, 10H9T, 1H1T\}$$

# Fully Collapsed Sampler

## Derivation: Multinomial Mixture Model

$$\frac{P(X_k | \beta)}{P(X_{k,-i} | \beta)} = \frac{P(20H19T | \beta)}{P(14H14T | \beta)}$$

Using Dirichlet-multinomial distribution,

$$= \frac{\frac{\tau(20+\beta_1)\tau(19+\beta_2)}{\tau(39+\beta_1+\beta_2)}}{\frac{\tau(14+\beta_1)\tau(14+\beta_2)}{\tau(28+\beta_1+\beta_2)}} \quad \text{equation (2)}$$

Using gamma function,  $\tau(N) = (N-1)!$ ,

$$\frac{P(X_k | \beta)}{P(X_{k,-i} | \beta)} = \frac{\frac{(19+\beta_1)! (18+\beta_2)!}{(38+\beta_1+\beta_2)!}}{\frac{(13+\beta_1)! (13+\beta_2)!}{(27+\beta_1+\beta_2)!}} \quad \text{equation(3)}$$

# Fully Collapsed Sampler

## Derivation: Multinomial Mixture Model

Equation 2 can also be written as:

$$\frac{P(X_k | \beta)}{P(X_{k,-i} | \beta)} = \frac{\frac{\tau(14+6+\beta_1)\tau(14+5+\beta_2)}{\tau(28+11+\beta_1+\beta_2)}}{\frac{\tau(14+\beta_1)\tau(14+\beta_2)}{\tau(28+\beta_1+\beta_2)}}$$

Generalizing,

$$\frac{P(X_k | \beta)}{P(X_{k,-i} | \beta)} = \frac{\frac{\tau(x+a+\beta_1)\tau(x+b+\beta_2)}{\tau(x+y+a+b+\beta_1+\beta_2)}}{\frac{\tau(x+\beta_1)\tau(y+\beta_2)}{\tau(x+y+\beta_1+\beta_2)}}$$

$$\text{Using gamma function, } \frac{P(X_k | \beta)}{P(X_{k,-i} | \beta)} = \frac{\frac{(x+a+\beta_1-1)! (y+b+\beta_2-1)!}{(x+y+a+b+\beta_1+\beta_2-1)!}}{\frac{(x+\beta_1-1)! (y+\beta_2-1)!}{(x+y+\beta_1+\beta_2)!}}$$

# Fully Collapsed Sampler

## Derivation: Multinomial Mixture Model

$$\text{or, } \frac{P(X_k | \beta)}{P(X_{k,-i} | \beta)} = \frac{\prod_{a=0}^{a-1} (x+a+\beta_1) \prod_{b=0}^{b-1} (y+b+\beta_2)}{\prod_{a+b=0}^{a+b-1} (x+y+\beta_1+\beta_2)}$$

$$\text{More generally, } \frac{P(X_k | \beta)}{P(X_{k,-i} | \beta)} = \frac{c_{xi}^k + \beta_{xi}}{C^k + B}$$

Here,  $c_{xi}$  is the count of the number of observations having the same side  $x_i$  in class  $k$ ,  $n$  is the number of .

# Fully Collapsed Sampler

- **Multinomial Mixture Model**

1. Randomly assign values to cluster, i.e,  $z_i = k$  with uniform probability.

2. For each cluster  $k$ , compute  $\frac{c_{xi}^k + \beta_{xi}}{C^k + B}$

3. Add data point  $(x_i, z_i)$  back by sampling  $z_i = k$ , using:

$$z_i = k \sim \frac{c_k + A/k}{N + A - 1} \left( \frac{c_{xi}^k + \beta_{xi}}{C^k + B} \right)$$

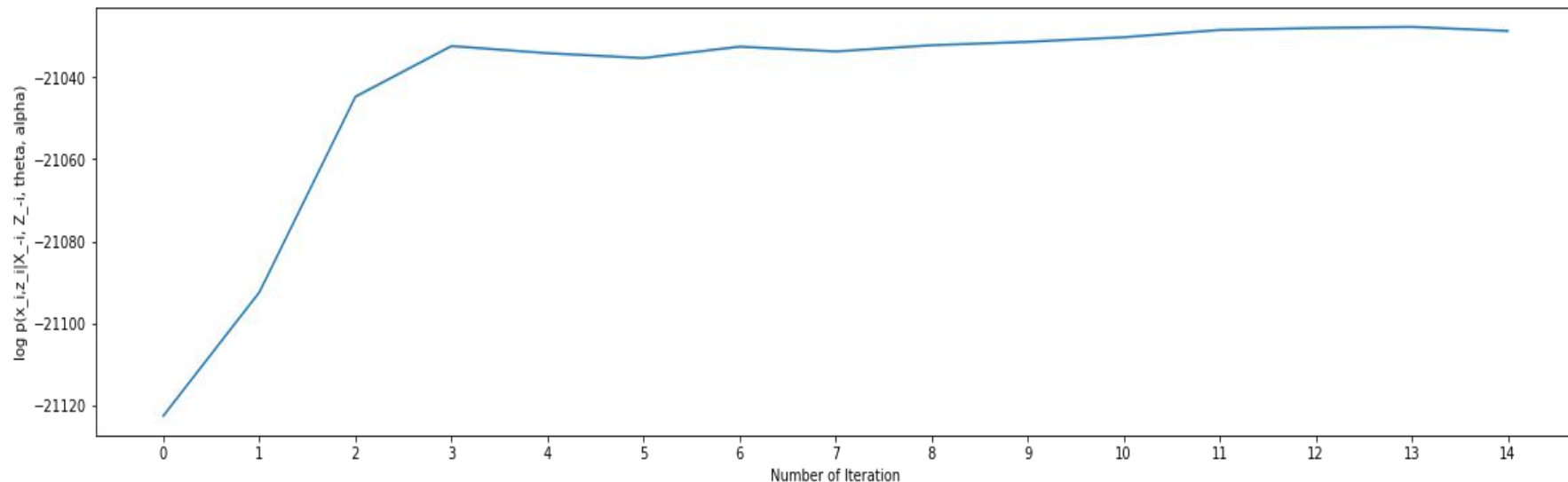
Mixture proportions can be estimated at the end.

- Suppose we have two mixture components with the following parameters:

```
dice_sides = 3
param = {0: {'pi1': 0.4, 'theta1': np.random.randint(1, 100,
size=dice_sides)},
         1: {'pi2': 0.6, 'theta2': np.random.randint(1, 100,
size=dice_sides)},
}
```

# Collapsed sampler in single dataset (after 50 iterations)

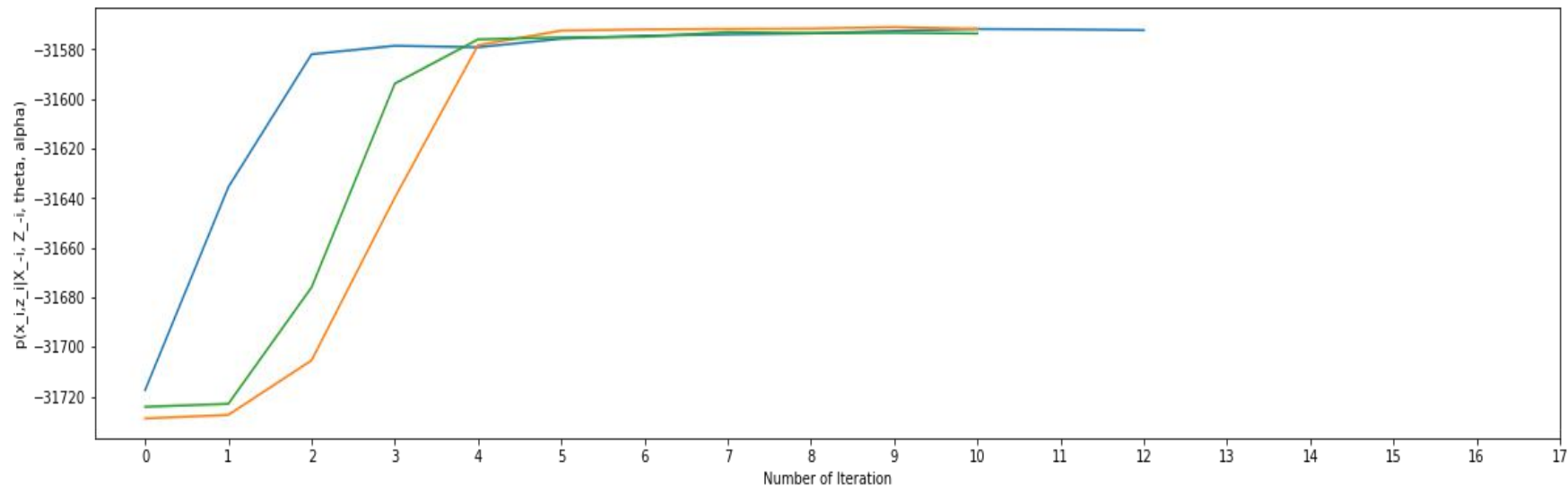
(As algorithm converge at 14 iterations, iterations after 14 are skipped)





# Collapsed sampler in multiple datasets (after 50 iterations)

(As algorithm converge at 13 iterations, iterations after 13 are skipped)



# Sample code

- Fully collapsed sampler (Multinomial Mixture Model):

<https://colab.research.google.com/drive/1H8a0ISumianazWHcBS3We6wIbS2qLC7T>

# Takeaways:

- We learned how to represent knowledge using graphical models.
- We learned about multinomial mixture model.
- We learned the way to model multinomial mixture model using gibbs sampling in the cases when parameters are known and unknown.

# References:

- Suresh Manandhar. *Bayesian ML : Posterior Distributions and Mixture Models Continuous Probability Density Function*
- Christopher Bishop. *Pattern Recognition and Machine Learning*

Thank you