Finite Mixture Model: Multinomial

Digya Acharya

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.

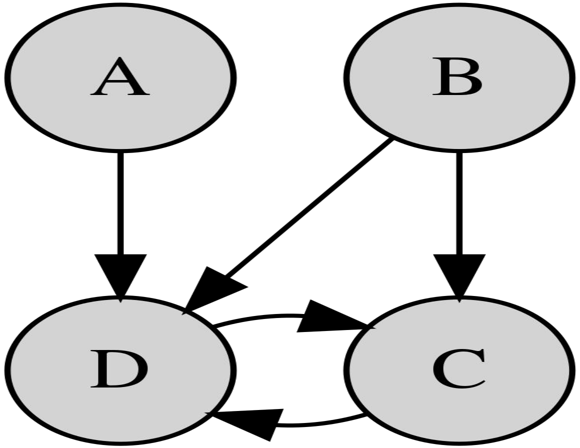
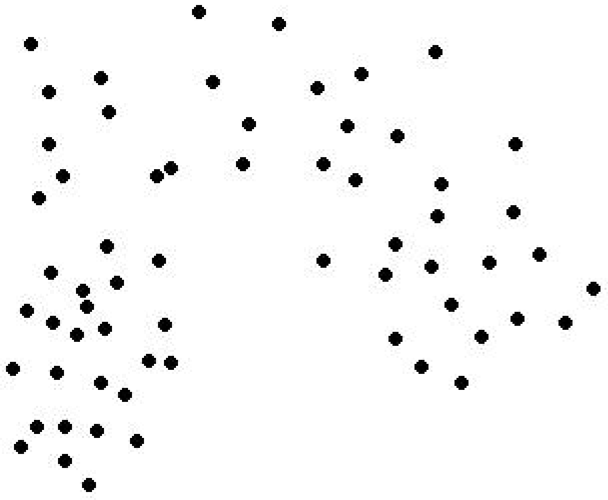


Figure 1: An example of a graphical model. Each arrow indicates a dependency. In this example: D depends on A, B, and C; and C depends on B and D; whereas A and B are each independent.
[Image Source: Wikipedia]

Mixture models

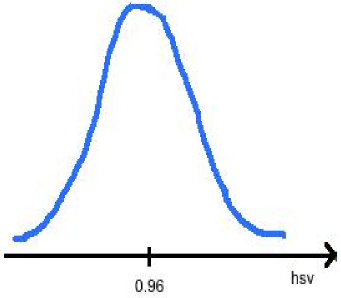
Clustering:
- Involves grouping of similar objects into a set known as cluster.
- Applications: creating newsfeeds, customer segmentation, social network analysis and so on.

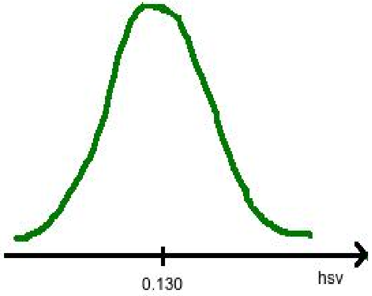


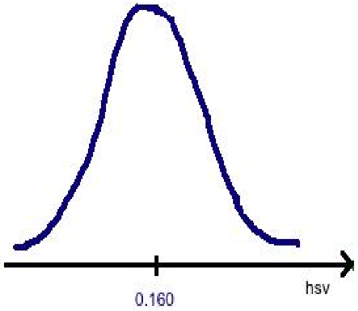
What is the picture representing (sky, river or forest)?

Picture from: Unsplash

HSV color distribution

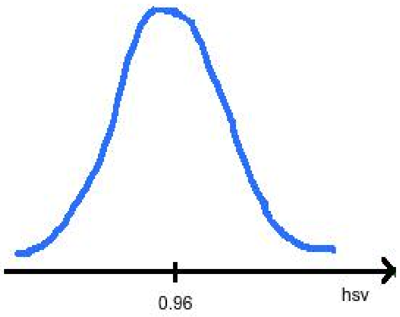




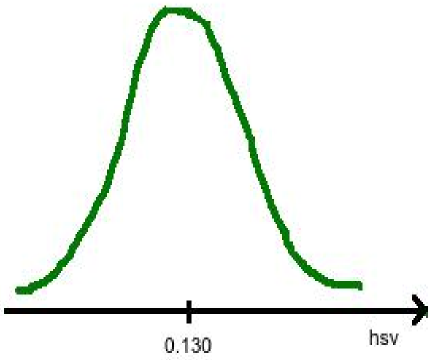


Weight assignment

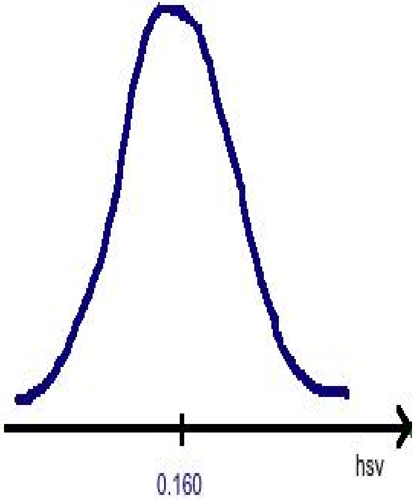
$\pi_k = [0.2, 0.3, 0.5]$



P(x | μ_1)



P(x | μ_2)



P(x | μ_3)

- Each value of π_k lies between 0 and 1.
- The sum of all the values of π_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

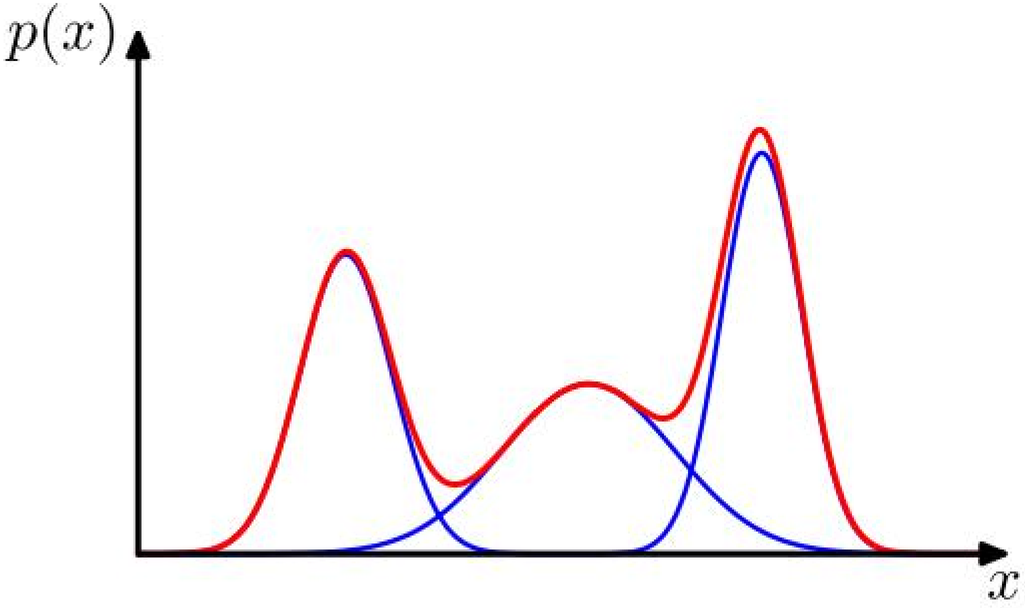


Figure 3: Illustration of mixture distribution in one dimension showing three distributions (each scaled by a coefficient) in blue and their sum in red

Bishop's PRML (Mixture of Gaussians)

Bimodal density function

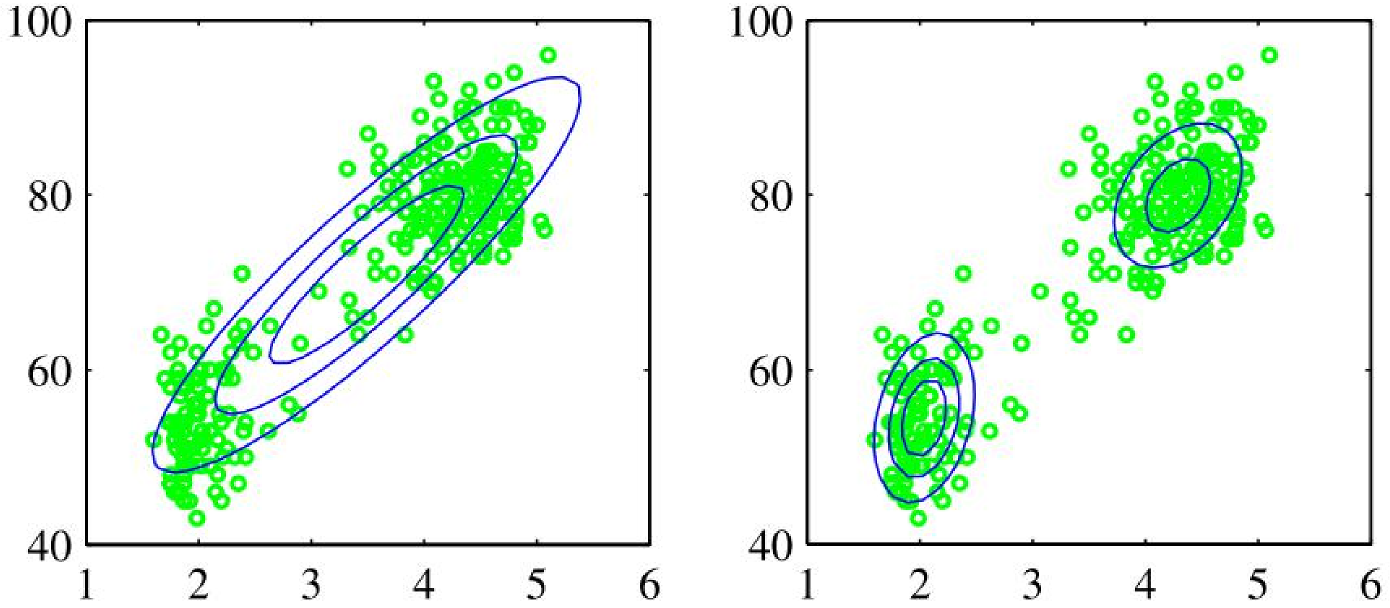


Figure 2: Left - single distribution representing data, Right- two distribution to model different set of data

Bishop’s PRML (Mixture of Gaussians)

Mixture models

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

Latent Variable Models

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

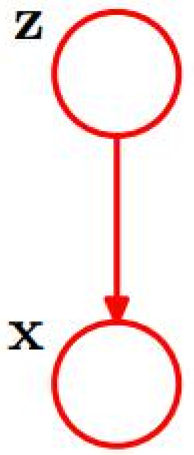
$$Z = \{z_1, z_2, \ldots, z_n\}$$

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, π_k is the mixture component.

Posterior Dirichlet Categorical Conditional Distribution

- Find the probability of belonging to observed class k, i.e, **$p(z_i=k)$.**

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

Biasedness (θ): Math: English = 0.25: 0.75, i.e, θ = [0.25, 0.75]

Random data points: X1, X2, X3, X4 **To find:** P(Z=k|θ)

P(Z_X1 = Math | [0.25, 0.75]) = 1

P(Z_X2 = Math | [0.25, 0.75]) = 0

P(Z_X3 = Math | [0.25, 0.75]) = 0

P(Z_X4 = Math | [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 ***(c_1,...,c_k)*** .

We will use an indicator variable ***z = k*** to indicate that the observed class is ***k***:

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

Using dirichlet prior,

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

Posterior Dirichlet Categorical Conditional Distribution

Here, **C-1** is the total number of items before adding new **z**. Simplifying the above equation, we get,

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

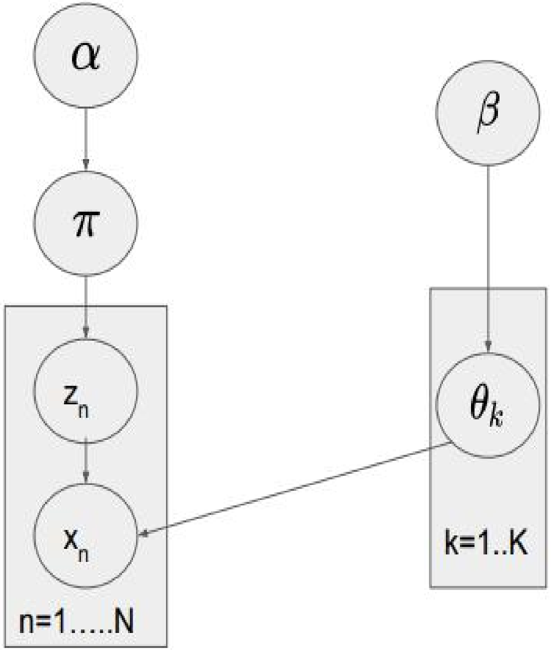
Using $\tau(N+1) = N\tau(N)$,

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

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

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

Modeling Mixture Component: Finite Mixture Model



To model mixture components $\boldsymbol{\pi}$, we use:

$$p(X, Z, \pi|\theta, \alpha) = p(X, Z|\theta, \pi)p(\pi|\alpha)$$
$$= p(X|Z, \theta)p(Z|\pi)p(\pi|\alpha)$$

Here,

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

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

- $p(Z|\pi)$ ~ Categorical(π)
- $p(\theta_k|\beta)$ ~ Beta(β)

We can integrate out mixture proportions ‘π’ using dirichlet categorical distribution:

$$
\begin{aligned}
\mathrm{p}(\mathrm{X}, \mathrm{Z} | \theta, \alpha) &= \int \mathrm{p}(\mathrm{X}, \mathrm{Z}, \pi | \theta, \alpha) \mathrm{d}\pi \\
&= \mathrm{p}(\mathrm{X}|\mathrm{Z}, \theta) \int \mathrm{p}(\mathrm{Z}|\pi) \mathrm{p}(\pi|\alpha) \mathrm{d}\pi \\
&= \mathrm{p}(\mathrm{X}|\mathrm{Z}, \theta) \mathrm{p}(\mathrm{Z}|\alpha) \\
&= \mathrm{p}(\mathrm{Z}|\alpha) \prod_{\mathrm{i}} \mathrm{p}(\mathrm{x_i}|\mathrm{Z}, \theta) \\
&= \mathrm{p}(\mathrm{Z}|\alpha) \prod_{\mathrm{i}} \mathrm{p}(\mathrm{x_i}|\theta_{\mathrm{zi}})
\end{aligned}
$$

Gibbs Sampling for Finite Mixtures

As conditional distribution converges to joint distribution in the limit,

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

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

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

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: {π: 0.2, θ:0.1},
          1: {π: 0.8, θ:0.9}}
```

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

Cases

1. We know the value of ‘θ’.
2. We do not know the value of ‘θ’.

Case 1 ('θ' 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:

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

iii) Compute multinomial distribution for each cluster k:

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

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) *To obtain mixture proportion, 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.*

Case 2 ('θ' 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:

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

iii) Estimate multinomial parameters θ_k for each cluster k using:

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

iv) Compute multinomial distribution for each cluster k:

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

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) *To obtain mixture proportion, 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.*

Sample code

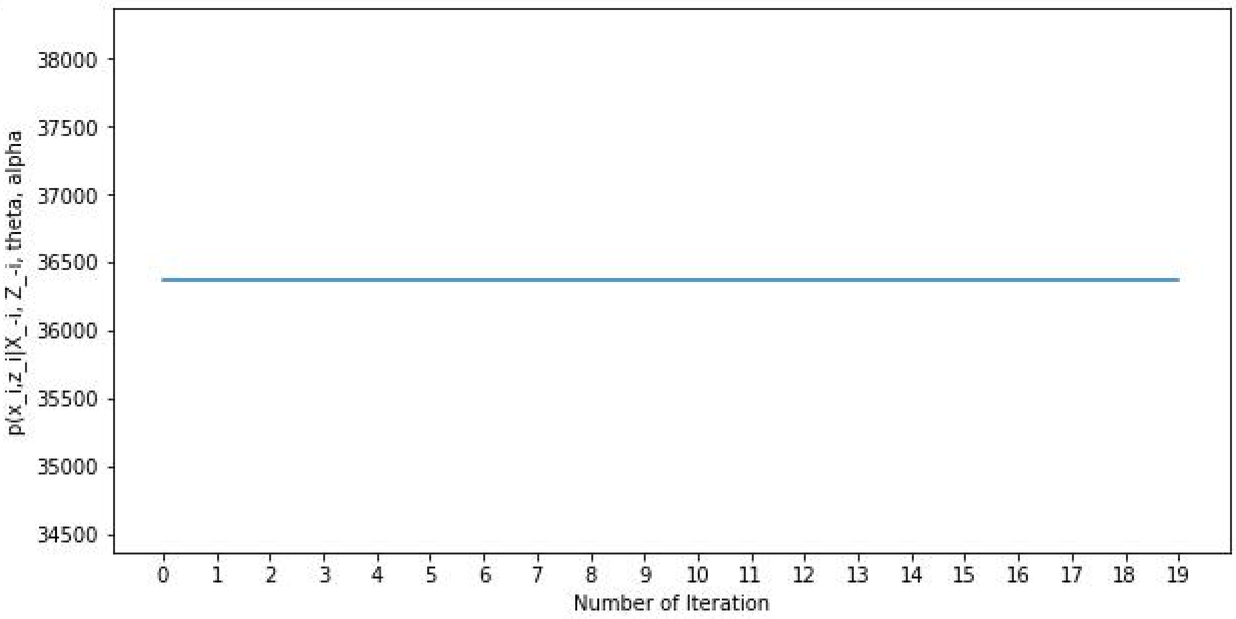
- Finite Mixtures with multinomial prior: (2 mixture components)

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

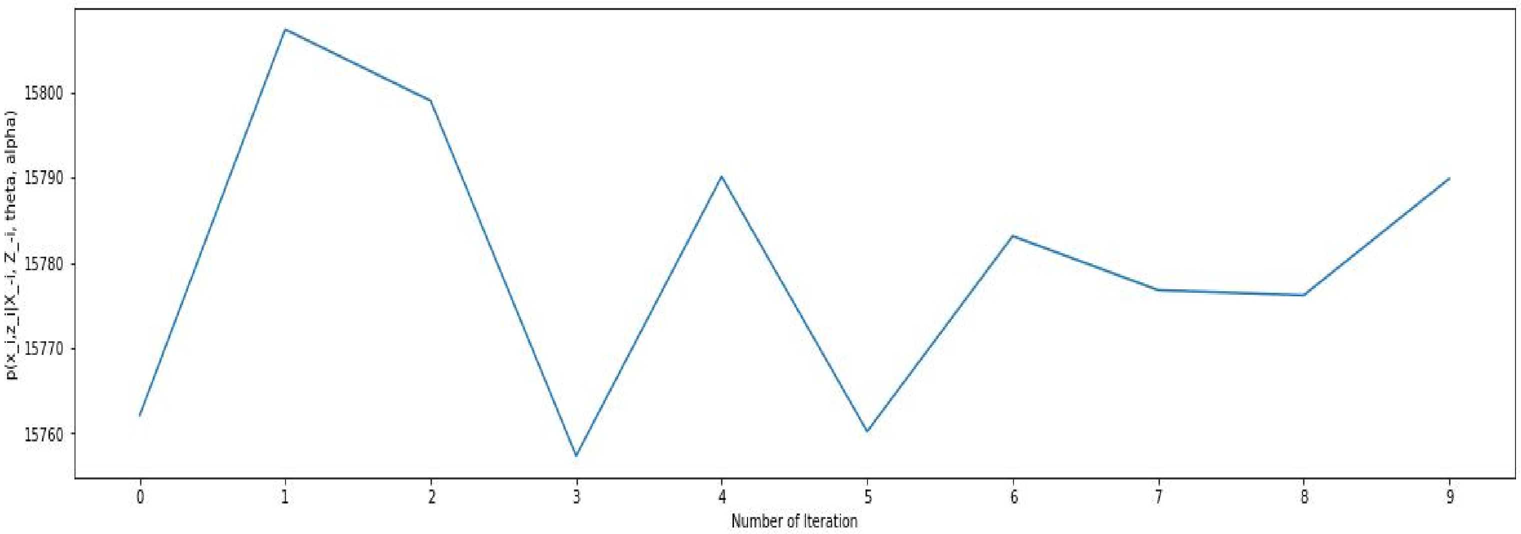
- Finite Mixtures with multinomial prior: (6 mixture components)

  https://drive.google.com/file/d/1MPwjr8DZk1b2Kxs4IsfIhUN0xkrJN56g/view?usp=sharing

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



Unknown parameters (3 mixture components) (after 20 iterations)



Fully Collapsed Gibbs Sampling

- **Multinomial Mixture Model**

The within class parameters θ_k can be integrated out using the Dirichlet-multinomial distribution

Thus, instead of

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

we need to compute

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

Here, ‘Z’ is the mixture component, ‘β’ is the shape parameter, ‘θ’ is the biasedness.

Fully Collapsed Sampler

- **Multinomial Mixture Model**

If we assume that for $(\boldsymbol{x_i}, \boldsymbol{z_i})$, the target distribution class is $\boldsymbol{k}$, i.e, $\boldsymbol{z_i} = \boldsymbol{k}$:

$$\frac{p(X \mid Z,\ \beta)}{p(X_{-i} \mid Z_{-i},\ \beta\ )} = \frac{\prod_k p(X_k \mid \beta)}{\prod_k p(X_{k,-i} \mid \beta)} = \frac{p(X_k \mid \beta)}{p(X_{k,-i} \mid \beta)}$$

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

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}*
Let x_i = *6H5T*
So, X_{k-i} = *{3H4T, 10H9T, 1H1T}*

Fully Collapsed Sampler

Derivation: Multinomial Mixture Model

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

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, $\tau(\mathrm{N}) = (\mathrm{N}\text{-}1)!$,

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

Fully Collapsed Sampler

Derivation: Multinomial Mixture Model

Equation 2 can also be written as:

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

Generalizing,

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

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

Fully Collapsed Sampler

Derivation: Multinomial Mixture Model

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

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

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 zi = k, using:

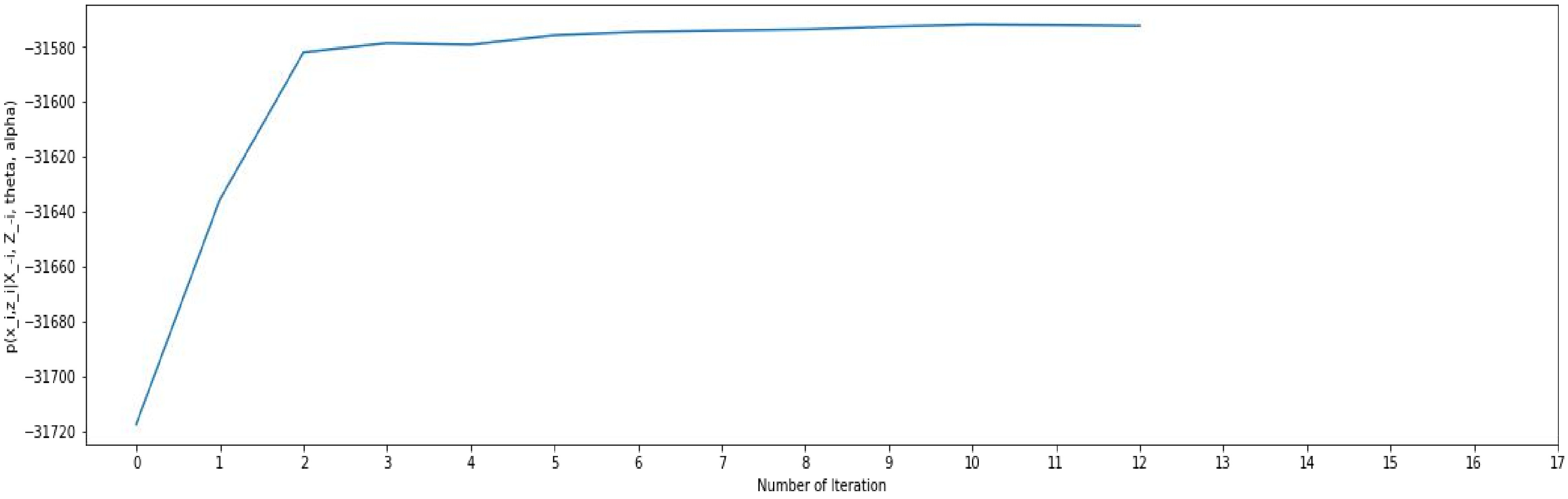
$$z_i = k \sim \frac{c_k + A/k}{N+A-1} \left(\frac{c_{xi}{}^{k} + \beta_{xi}}{C^{k}+B}\right)$$

Mixture proportions can be estimated at the end.

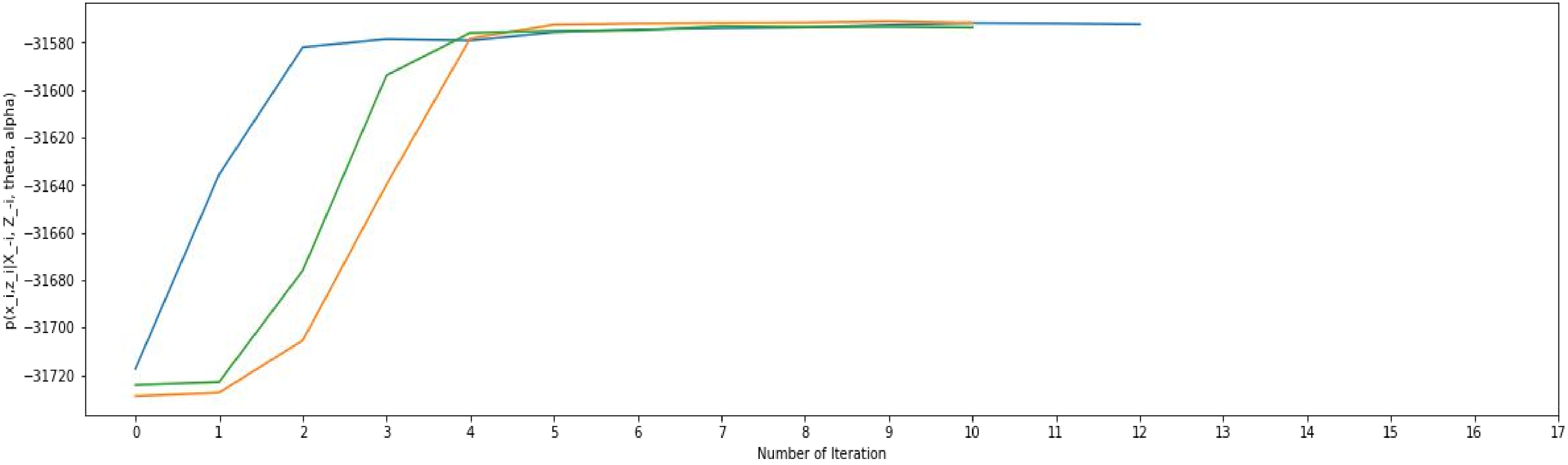
- Suppose we have two mixture components with the following parameters:

```
dice_sides = 3
param = {0: {'pi': 0.4, 'theta': np.random.randint(1, 100, size=dice_sides)},
             1: {'pi': 0.6, 'theta': np.random.randint(1, 100,
size=dice_sides)},
}
```

Single collapsed sampler (after 50 iterations)



Multiple collapsed sampler (after 50 iterations)



Sample code

- Fully collapsed sampler (Multinomial Mixture Model):

https://colab.research.google.com/drive/1H8a0lSumianazWHcBS3We6wIbS2qLC7T

References:

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

Thank you