Machine Learning for Data Science (CS4786)
Lecture 18

Probabilistic Models

- Set of models Θ consists of parameters s.t. P_θ for each $\theta \in \Theta$ is a distribution over data.

- Learning: Estimate $\theta^* \in \Theta$ that best models given data

Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

$$\theta_{MLE} = \text{argmax}_{\theta \in \Theta} \underbrace{\log P_\theta(x_1, \ldots, x_n)}_{\text{Likelihood}}$$

- A priori all models are equally good, data could have been generated by any one of them

Maximum A Posteriori

Pick $\theta \in \Theta$ that is most likely given data

Maximize a posteriori probability of model given data

$$\theta_{MAP} = \text{argmax}_{\theta \in \Theta} P(\theta | x_1, \ldots, x_n)$$

$$= \text{argmax}_{\theta \in \Theta} \; \log P(x_1, \ldots, x_n | \theta) + \log P(\theta)$$

EM Algorithm

Latent Variables

- We only observe $x_1, \ldots, x_n$, cluster assignments $c_1, \ldots, c_n$ are not observed
- Finding $\theta \in \Theta$ (even for 1-d GMM) that directly maximizes Likelihood or A Posteriori given $x_1, \ldots, x_n$ is hard!
- Given latent variables $c_1, \ldots, c_n$, the problem of maximizing likelihood (or a posteriori) became easy

Can we use latent variables to device an algorithm?

Expectation Maximization Algorithm

Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments

Expectation Maximization Algorithm

Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

(E step) For every t, define distribution Q_t over the latent variable c_t as:

$$Q_t^{(i)}(c_t) = P(c_t | x_t, \theta^{(i-1)})$$

(M step)

$$\theta^{(i)} = \mathrm{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta) \qquad \text{if MLE}$$

Expectation Maximization Algorithm

Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

(E step) For every t, define distribution Q_t over the latent variable c_t as:

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)}) \\ \propto P(x_t|c_t, \theta^{(i-1)})P(c_t|\theta^{(i-1)})$$

(M step)

$$\theta^{(i)} = \text{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta) \qquad \text{if MLE}$$

Expectation Maximization Algorithm

Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

(E step) For every t, define distribution Q_t over the latent variable c_t as:

$$Q_t^{(i)}(c_t) = P(c_t | x_t, \theta^{(i-1)}) \propto P(x_t | c_t, \theta^{(i-1)}) P(c_t | \theta^{(i-1)})$$

(M step)

$$\theta^{(i)} = \mathrm{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta) \qquad \text{if MLE}$$

$$\theta^{(i)} = \mathrm{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta) + \log P(\theta) \qquad \text{if MAP}$$

Why EM works?

- Every iteration of EM only improves log-likelihood (log a posteriori)

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n)$$

WHY SHOULD EM WORK?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)$$

Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)$$

$$= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right)$$

WHY SHOULD EM WORK?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\begin{aligned}
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) &= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} Q^{(i)}(c_t) \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right)
\end{aligned}$$

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$
\begin{aligned}
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) &= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} Q^{(i)}(c_t) \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right) \\
&\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
\end{aligned}
$$

WHY SHOULD EM WORK?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\begin{aligned}
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) &= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} Q^{(i)}(c_t) \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right) \\
&\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
\end{aligned}$$

Log(average) > average of Log

Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

M-step

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

M-step

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t | x_t)} \right)$$

E-step

WHY SHOULD EM WORK?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$ **M-step**

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t | x_t)} \right)$$ **E-step**

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t)$$

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

M-step

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t | x_t)} \right)$$

E-step

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t)$$

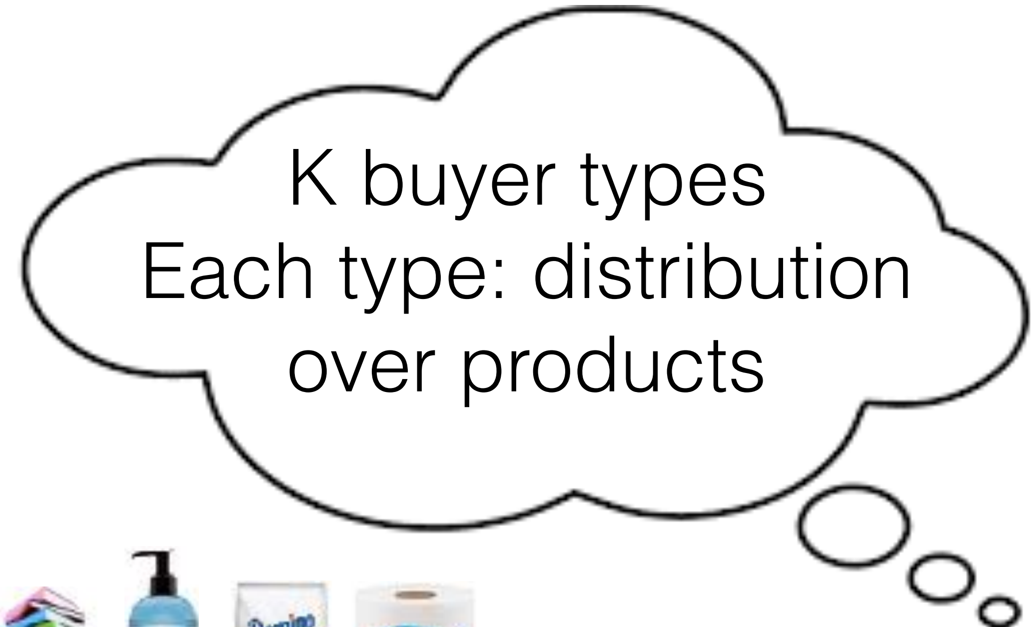
$$= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)$$

Mixture of Multinomials

Mixture of Multinomials

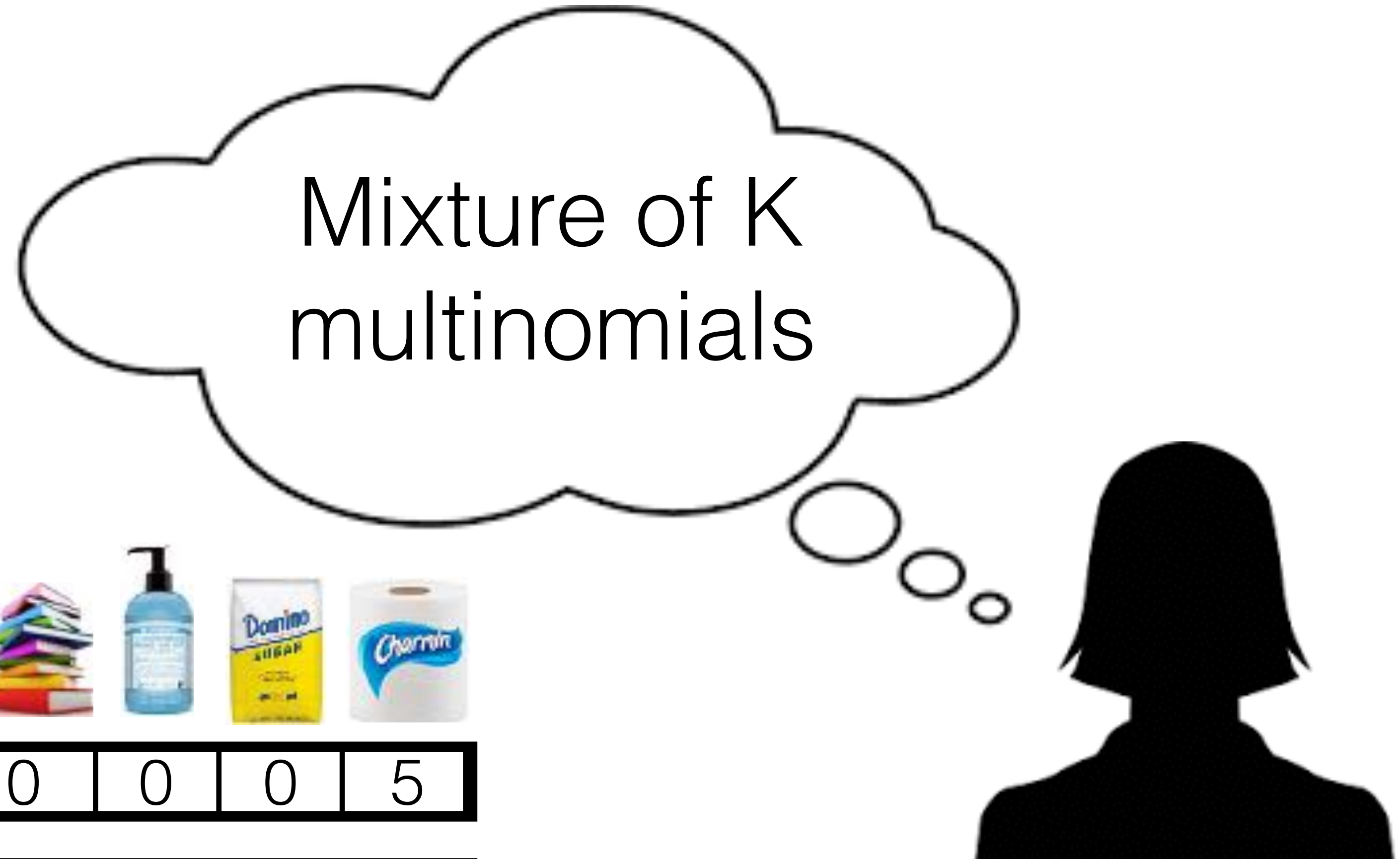
10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Mixture of Multinomials



10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Mixture of Multinomials

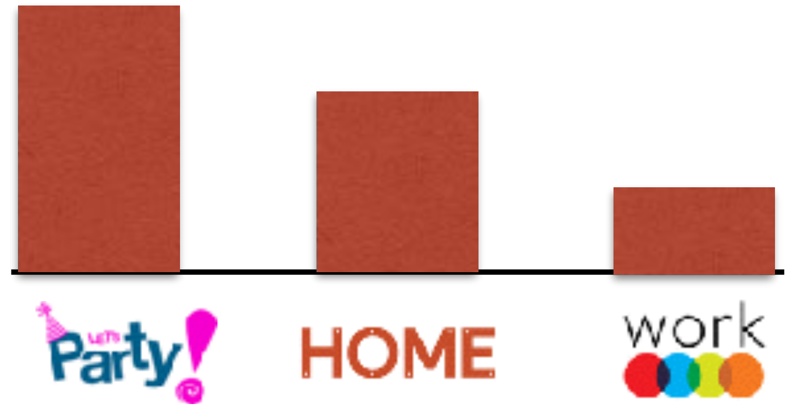


10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Mixture of Multinomials

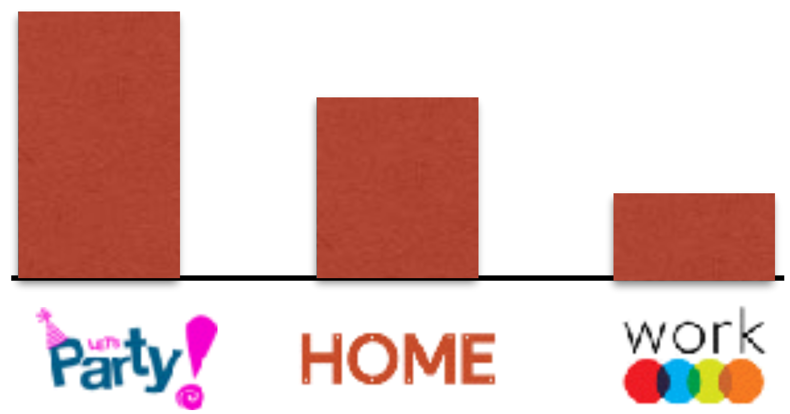
10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Mixture of Multinomials

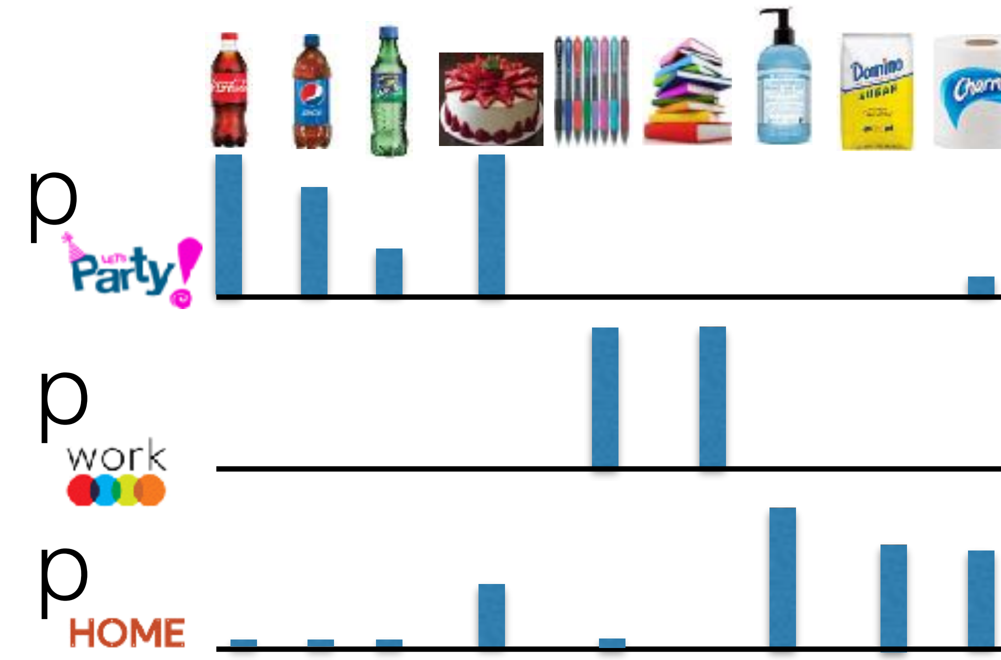
$\pi =$ 

10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

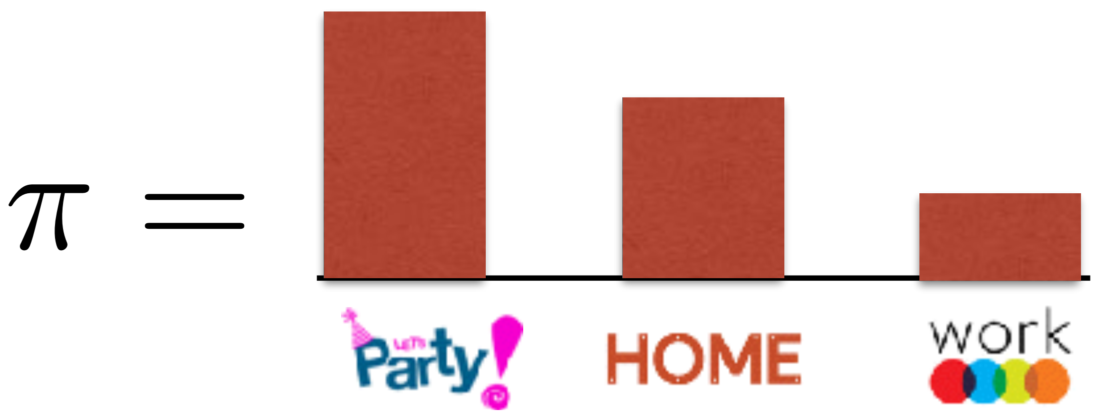
Mixture of Multinomials

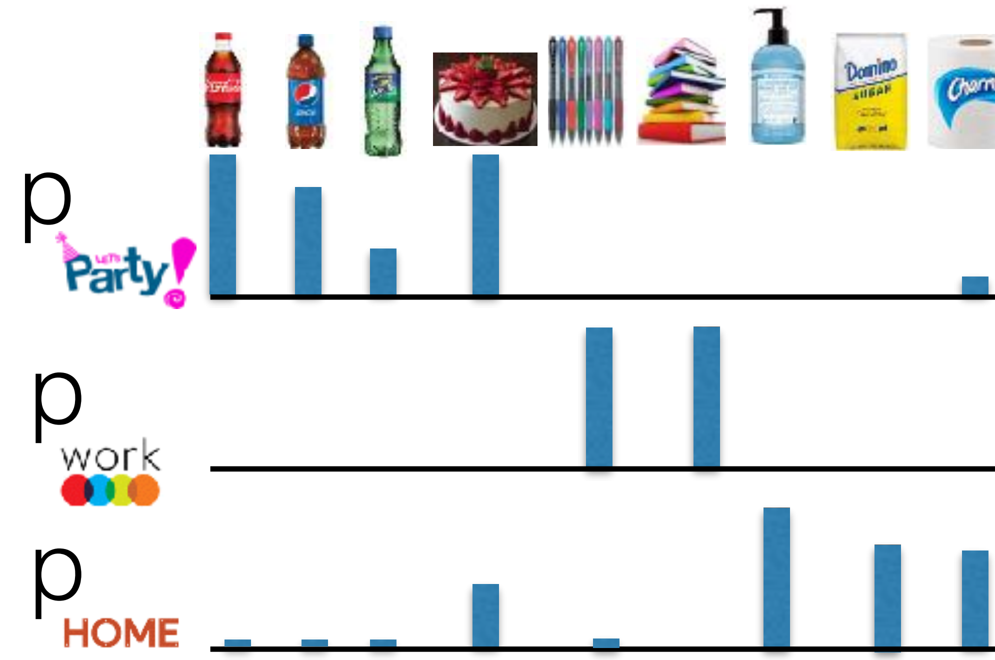
$\pi =$ 

10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5



Mixture of Multinomials



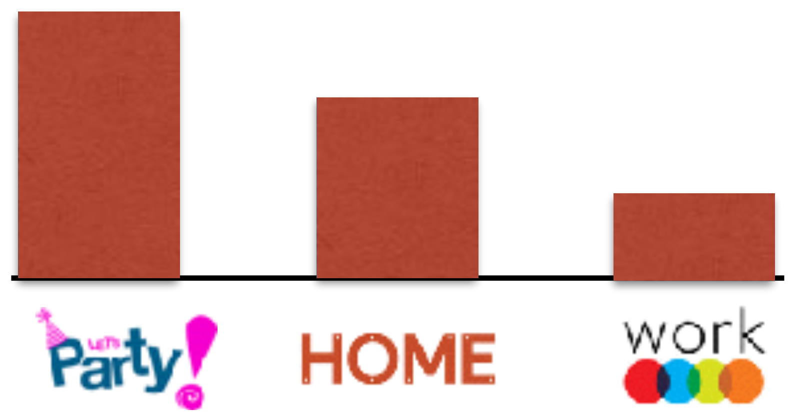


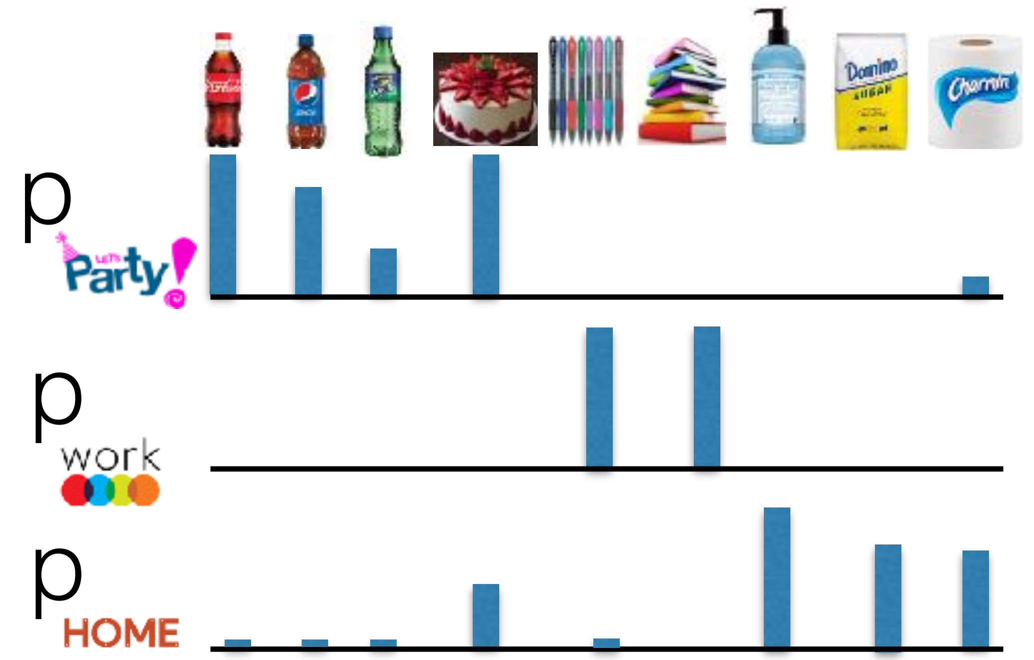
π ~

Party!	10	10	5	2	0	0	0	0	5
HOME	1	0	0	1	0	0	0	1	10
work	0	0	0	0	1	1	0	0	0
Party!	20	15	10	5	0	0	0	0	0
work	10	5	5	2	1	1	1	1	5

Mixture of Multinomials

$\pi =$





π ~									
Party! ~	10	10	5	2	0	0	0	0	5
HOME ~	1	0	0	1	0	0	0	1	10
work ~	0	0	0	0	1	1	0	0	0
Party! ~	20	15	10	5	0	0	0	0	0
work ~	10	5	5	2	1	1	1	1	5

Mixture of Multinomials

- Eg. Model purchases of each customer
- K-types of customers, each designated with distribution over the d items to buy
- Generative model:
  - π is mixture distribution over the K-types of buyers
  - $p_1, \ldots, p_K$ are the K distributions over the d items, one for each customer type
  - Generative process, each round draw customer type $c_t \sim \pi$
  - Next given c_t draw list of purchases as $x_t \sim \text{multinomial}(p_{c_t})$

Multinomial Distribution

$$P(x|p) = \frac{m!}{x[1]! \cdot \ldots \cdot x[d]!} p[1]^{x_t[1]} \cdot \ldots \cdot p[d]^{x_t[d]}$$

Probability of purchase vector x while drawing products independently m times from p

E-step

$$Q_t^{(i)}(c_t) \propto P(x_t | c_t, \theta^{(i-1)}) P(c_t | \theta^{(i-1)})$$

E-step

$$Q_t^{(i)}(c_t) \propto P(x_t | c_t, \theta^{(i-1)}) P(c_t | \theta^{(i-1)})$$

$$= \frac{P(x_t | p_{c_t}^{(i-1)}) \pi^{(i-1)}(c_t)}{\sum_{k=1}^{K} P(x_t | p_k^{(i-1)}) \pi^{(i-1)}(k)}$$

E-step

$$Q_t^{(i)}(c_t) \propto P(x_t|c_t, \theta^{(i-1)})P(c_t|\theta^{(i-1)})$$

$$= \frac{P(x_t|p_{c_t}^{(i-1)})\pi^{(i-1)}(c_t)}{\sum_{k=1}^K P(x_t|p_k^{(i-1)})\pi^{(i-1)}(k)}$$

$$= \frac{p_{c_t}[1]^{x_t[1]} \cdot \ldots \cdot p_{c_t}[d]^{x_t[d]} \cdot \pi_{c_t}^{(i-1)}}{\sum_{k=1}^K p_k[1]^{x_t[1]} \cdot \ldots \cdot p_{c_t}[d]^{x_t[d]} \cdot \pi_k^{(i-1)}}$$

M-step

$$\theta^{(i)} = \text{argmax}_\theta \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta) P(c_t = k | \theta) \right)$$

M-step

$$\theta^{(i)} = \text{argmax}_{\theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta) P(c_t = k | \theta) \right)$$

$$= \text{argmax}_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(\frac{m!}{x_t[1]! \cdot \ldots \cdot x_t[d]!} p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]} \right) + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \right\}$$

M-step

$$
\begin{aligned}
\theta^{(i)} &= \text{argmax}_\theta \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log\left(P(x_t|c_t = k, \theta) P(c_t = k|\theta)\right) \\
&= \text{argmax}_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log\left(\frac{m!}{x_t[1]! \cdot \ldots \cdot x_t[d]!} p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]} \right) \right. \\
&\qquad \left. + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \right\} \\
&= \text{argmax}_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log\left(p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]} \right) \right. \\
&\qquad \left. + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \right\}
\end{aligned}
$$

M-step

$$
\begin{aligned}
\theta^{(i)} &= \mathrm{argmax}_{\theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta) P(c_t = k | \theta) \right) \\
&= \mathrm{argmax}_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(\frac{m!}{x_t[1]! \cdot \ldots \cdot x_t[d]!} p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]} \right) \right. \\
&\qquad \left. + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \right\} \\
&= \mathrm{argmax}_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]} \right) \right. \\
&\qquad \left. + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \right\} \\
&= \mathrm{argmax}_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{d} Q_t^{(i)}(k) x_t[j] \log \left(p_k[j] \right) + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \right\}
\end{aligned}
$$

M-step

$$\pi_k^{(i)} = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n}$$

$$p_k[j] = \frac{\sum_{t=1}^{n} x_t[j] Q_t^{(i)}(k)}{m \sum_{t=1}^{n} Q_t^{(i)}(k)}$$

M-step

$$\pi_k^{(i)} = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n}$$

proportion of weights for each type

$$p_k[j] = \frac{\sum_{t=1}^{n} x_t[j] Q_t^{(i)}(k)}{m \sum_{t=1}^{n} Q_t^{(i)}(k)}$$

weighted number of jth product

Mixture of Multinomials

What is missing in this story?

Mixture of Multinomials

What is missing in this story?

10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Mixture of Multinomials

What is missing in this story?

10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Mixture of Multinomials

What is missing in this story?

10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Everyone is a bit of party and a bit of work!

Latent Dirichlet Allocation

- Generative story:

For $t = 1$ to n

For each customer draw mixture of types π_t

For $i = 1$ to m

For each item to purchase, first draw type $c_t[i] \sim \pi_t$

Next, given the type draw $x_t[i] \sim p_{c_t[i]}$

End For

End For

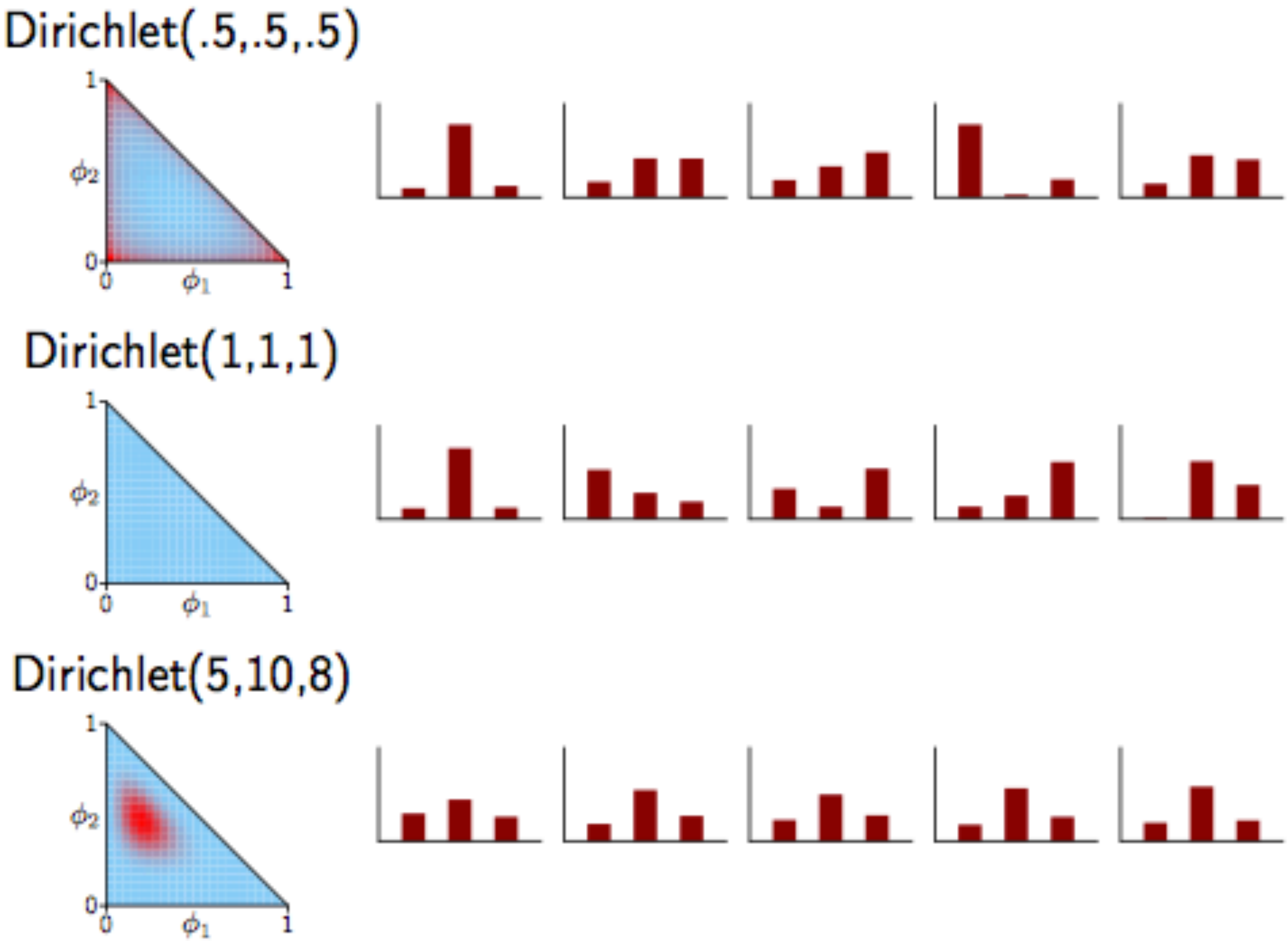
Dirichlet Distribution

- Its a distribution over distributions!
- Parameters $\alpha_1, \ldots, \alpha_K$ s.t. $\alpha_k > 0$
- The density function is given as

$$p(\pi; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k}$$

where $B(\alpha) = \prod_{k=1}^{K} \Gamma(\alpha_k) / \Gamma(\sum_{k=1}^{K} \alpha_k)$

Dirichlet Distribution



Latent Dirichlet Allocation

- Generative story:

  For $t = 1$ to n

  For each customer draw mixture of types $\pi_t \sim \text{Dirchlet}(\alpha)$

  For $i = 1$ to m

  For each item to purchase, first draw type $c_t[i] \sim \pi_t$

  Next, given the type draw $x_t[i] \sim p_{c_t[i]}$

  End For

  End For

- Parameters, α for the Dirichlet distribution and $p_1, \ldots, p_K$