

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

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

- 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

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

MAXIMUM A POSTERIORI

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

Maximize a posteriori probability of model given data

$$\begin{aligned}\theta_{MAP} &= \operatorname{argmax}_{\theta \in \Theta} P(\theta | x_1, \dots, x_n) \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{P(x_1, \dots, x_n | \theta) P(\theta)}{\sum_{\theta \in \Theta} P(x_1, \dots, x_n | \theta) P(\theta)}\end{aligned}$$

MAXIMUM A POSTERIORI

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

Maximize a posteriori probability of model given data

$$\begin{aligned}\theta_{MAP} &= \operatorname{argmax}_{\theta \in \Theta} P(\theta | x_1, \dots, x_n) \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{P(x_1, \dots, x_n | \theta) P(\theta)}{\sum_{\theta \in \Theta} P(x_1, \dots, x_n | \theta) P(\theta)} \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{P(x_1, \dots, x_n | \theta) P(\theta)}{P(x_1, \dots, x_n)}\end{aligned}$$

MAXIMUM A POSTERIORI

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

Maximize a posteriori probability of model given data

$$\begin{aligned}\theta_{MAP} &= \operatorname{argmax}_{\theta \in \Theta} P(\theta | x_1, \dots, x_n) \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{P(x_1, \dots, x_n | \theta) P(\theta)}{\sum_{\theta \in \Theta} P(x_1, \dots, x_n | \theta) P(\theta)} \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{P(x_1, \dots, x_n | \theta) P(\theta)}{P(x_1, \dots, x_n)} \\ &= \operatorname{argmax}_{\theta \in \Theta} \underbrace{P(x_1, \dots, x_n | \theta)}_{\text{red}} \underbrace{P(\theta)}_{\text{green}}\end{aligned}$$

MAXIMUM A POSTERIORI

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

Maximize a posteriori probability of model given data

$$\begin{aligned}\theta_{MAP} &= \operatorname{argmax}_{\theta \in \Theta} P(\theta | x_1, \dots, x_n) \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{P(x_1, \dots, x_n | \theta) P(\theta)}{\sum_{\theta \in \Theta} P(x_1, \dots, x_n | \theta) P(\theta)} \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{P(x_1, \dots, x_n | \theta) P(\theta)}{P(x_1, \dots, x_n)} \\ &= \operatorname{argmax}_{\theta \in \Theta} \underbrace{P(x_1, \dots, x_n | \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}} \\ &= \operatorname{argmax}_{\theta \in \Theta} \log P(x_1, \dots, x_n | \theta) + \log P(\theta)\end{aligned}$$

EM Algorithm

LATENT VARIABLES

- We only observe x_1, \dots, x_n , cluster assignments c_1, \dots, c_n are not observed
- Finding $\theta \in \Theta$ (even for 1-d GMM) that directly maximizes Likelihood or A Posteriori given x_1, \dots, x_n is hard!
- Given latent variables c_1, \dots, 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, \dots, c_n are Latent variables. Eg. cluster assignments

EXPECTATION MAXIMIZATION ALGORITHM

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

- Initialize $\theta^{(0)}$ arbitrarily, repeat until 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)

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

EXPECTATION MAXIMIZATION ALGORITHM

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

- Initialize $\theta^{(0)}$ arbitrarily, repeat until 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)

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

EXPECTATION MAXIMIZATION ALGORITHM

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

- Initialize $\theta^{(0)}$ arbitrarily, repeat until 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)

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

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

Why EM works?

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

WHY SHOULD EM WORK?

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

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

WHY SHOULD EM WORK?

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

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

WHY SHOULD EM WORK?

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

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

WHY SHOULD EM WORK?

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

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

WHY SHOULD EM WORK?

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

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

WHY SHOULD EM WORK?

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

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

Log(average) > average of Log

WHY SHOULD EM WORK?

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

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

WHY SHOULD EM WORK?

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

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

M-step

WHY SHOULD EM WORK?

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

$$\begin{aligned} \log P_{\theta^{(i)}}(x_1, \dots, x_n) &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \\ &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) && \mathbf{M\text{-step}} \\ &= \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right) && \mathbf{E\text{-step}} \end{aligned}$$

WHY SHOULD EM WORK?

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

$$\begin{aligned} \log P_{\theta^{(i)}}(x_1, \dots, x_n) &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \\ &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) && \mathbf{M\text{-step}} \\ &= \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right) && \mathbf{E\text{-step}} \\ &= \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t) \end{aligned}$$

WHY SHOULD EM WORK?

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

$$\begin{aligned} \log P_{\theta^{(i)}}(x_1, \dots, x_n) &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \\ &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) && \mathbf{M\text{-step}} \\ &= \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right) && \mathbf{E\text{-step}} \\ &= \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t) \\ &= \sum_{t=1}^n \log P_{\theta^{(i)}}(x_t) \end{aligned}$$

EXAMPLE: EM FOR GMM

For every $k \in [K]$, the maximization step yields,

$$\mu_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t(k)}, \quad \Sigma_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) (x_t - \mu_k^{(i)}) (x_t - \mu_k^{(i)})^\top}{\sum_{t=1}^n Q_t(k)}$$

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

Let us derive this!

EXAMPLE: EM FOR GMM

For every $k \in [K]$, the maximization step yields,

$$\mu_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t(k)}, \quad \Sigma_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) (x_t - \mu_k^{(i)}) (x_t - \mu_k^{(i)})^\top}{\sum_{t=1}^n Q_t(k)}$$

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

Let us derive this!

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

K buyer types
Each type: distribution
over products



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

Mixture of K 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
----	----	---	---	---	---	---	---	---

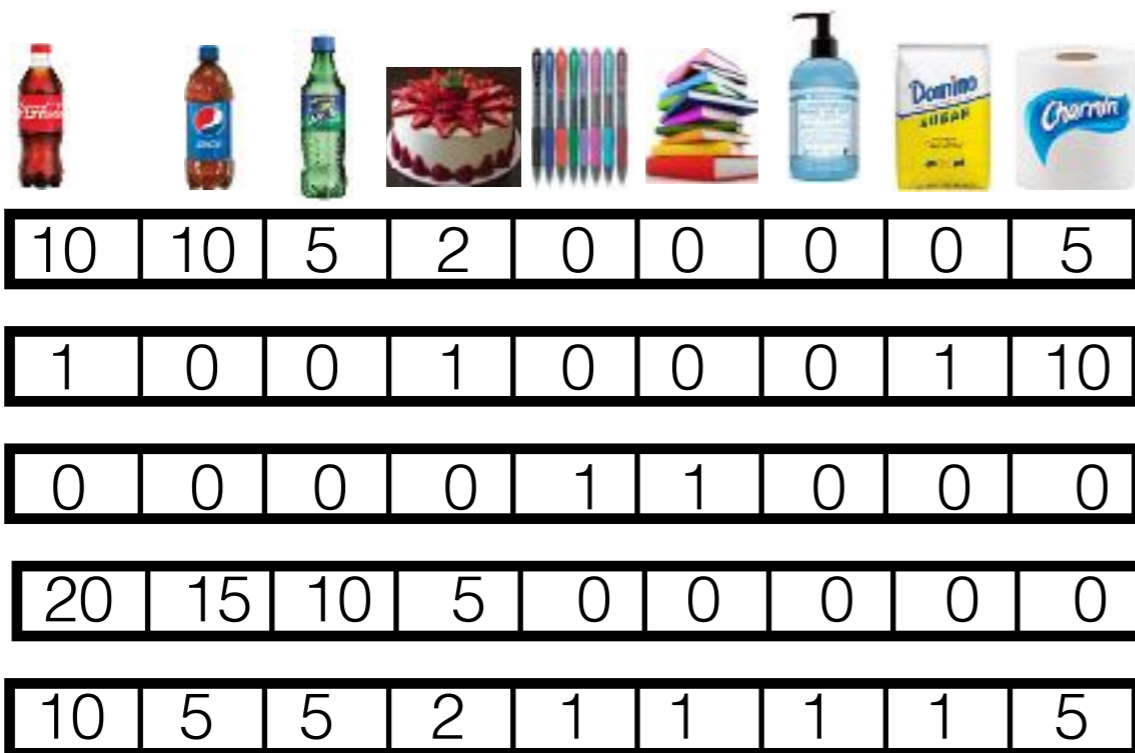
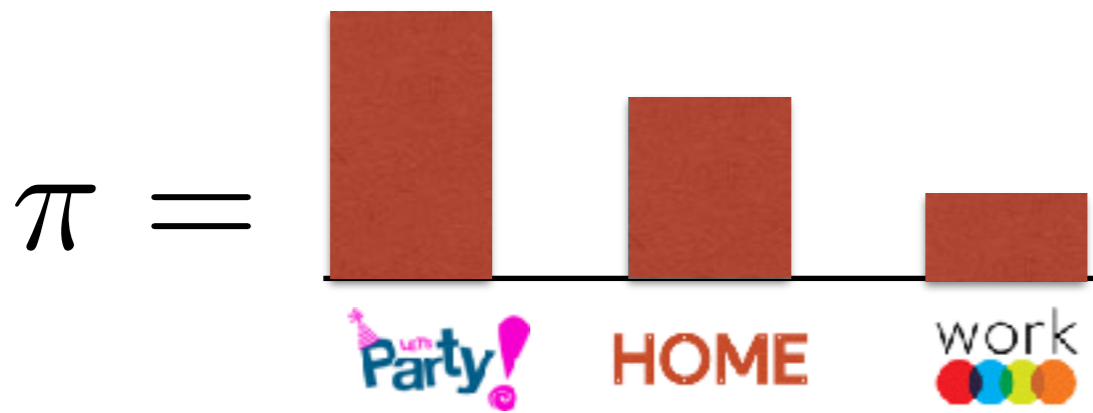
1	0	0	1	0	0	0	1	10
---	---	---	---	---	---	---	---	----

0	0	0	0	1	1	0	0	0
---	---	---	---	---	---	---	---	---

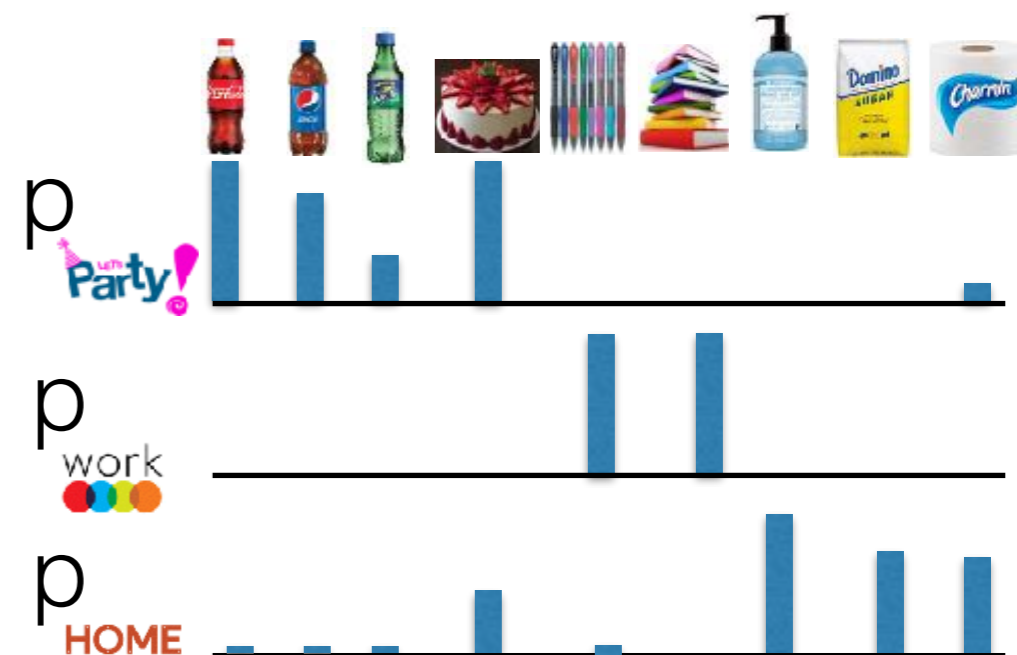
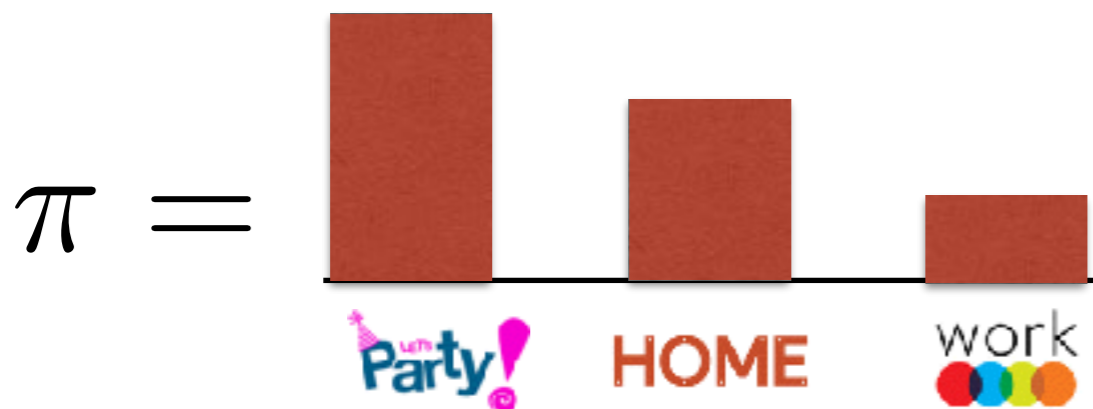
20	15	10	5	0	0	0	0	0
----	----	----	---	---	---	---	---	---

10	5	5	2	1	1	1	1	5
----	---	---	---	---	---	---	---	---

Mixture of Multinomials



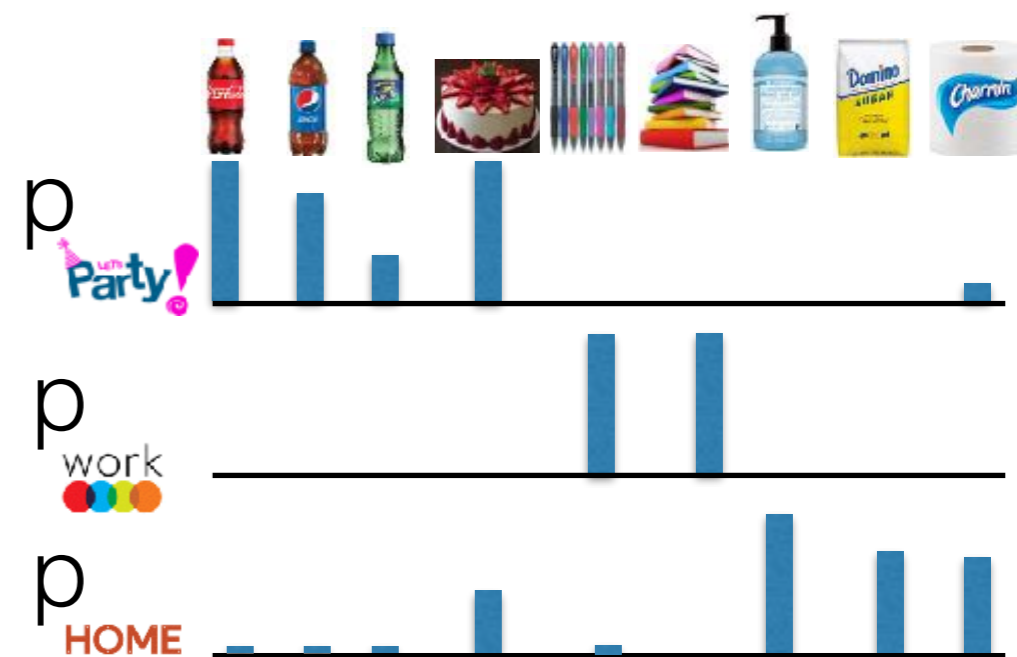
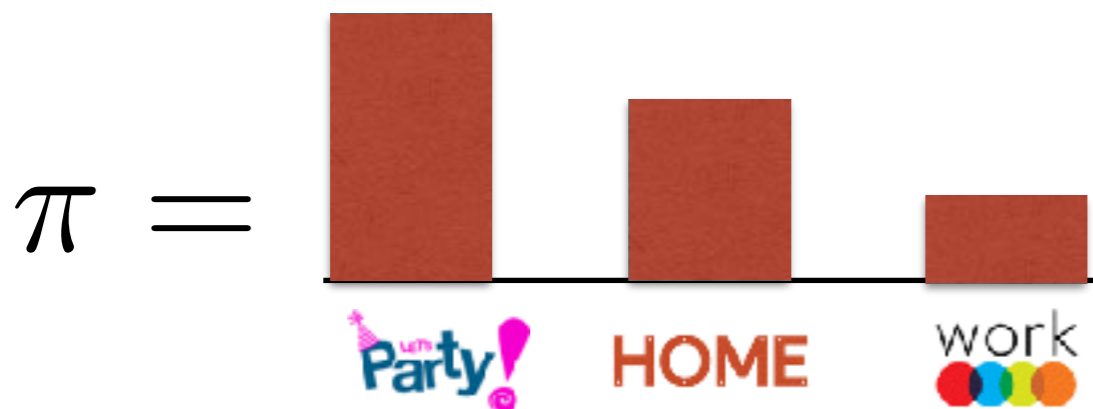
Mixture of Multinomials



A row of nine icons corresponding to the columns of the table below: Coca-Cola, Pepsi, Sprite, Cake, Pens, Books, Lotion, Domino Sugar, and Charmin.

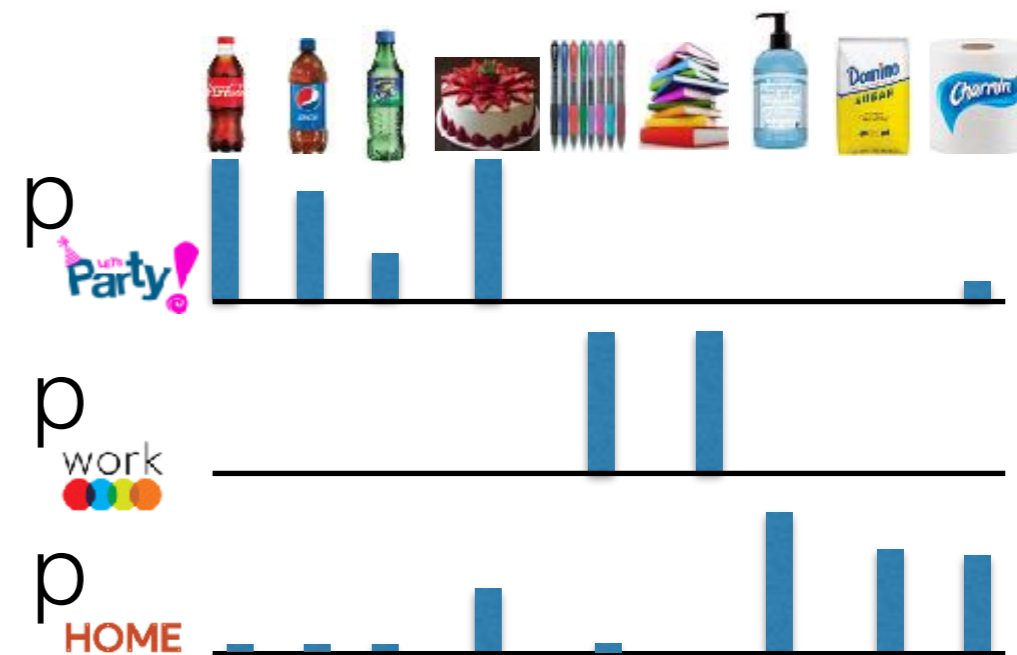
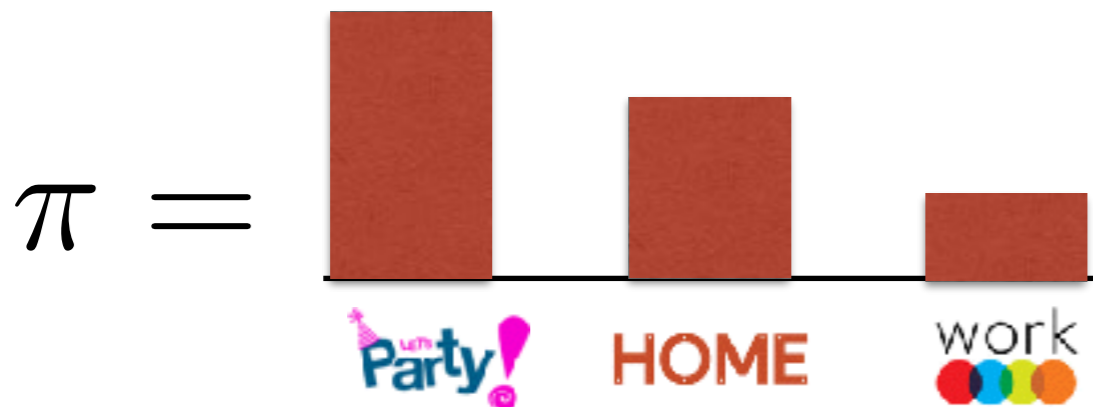
10	10	5	2	0	0	0	0	5
1	0	0	1	0	0	0	1	10
0	0	0	0	1	1	0	0	0
20	15	10	5	0	0	0	0	0
10	5	5	2	1	1	1	1	5

Mixture of Multinomials



	Party	Pepsi	Sprite	Cake	Markers	Books	Hand Sanitizer	Domino Sugar	Charmin
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
π_{Party}	20	15	10	5	0	0	0	0	0
work	10	5	5	2	1	1	1	1	5

Mixture of Multinomials



	Coca-Cola	Pepsi	Sprite	Cake	Markers	Books	Hand Sanitizer	Domino Sugar	Charmin
π	~0.10	~0.10	~0.05	~0.02	~0.00	~0.00	~0.00	~0.00	~0.05
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, \dots, 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

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

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

$$\begin{aligned} Q_t^{(i)}(c_t) &\propto P(x_t | c_t, \theta^{(i-1)}) P(c_t | \theta^{(i-1)}) \\ &= \frac{P(x_t | p_{c_t}^{(i-1)}) \pi^{(i-1)}(c_t)}{\sum_{k=1}^K P(x_t | p_k^{(i-1)}) \pi^{(i-1)}(k)} \end{aligned}$$

E-step

$$\begin{aligned} Q_t^{(i)}(c_t) &\propto P(x_t | c_t, \theta^{(i-1)}) P(c_t | \theta^{(i-1)}) \\ &= \frac{P(x_t | p_{c_t}^{(i-1)}) \pi^{(i-1)}(c_t)}{\sum_{k=1}^K P(x_t | p_k^{(i-1)}) \pi^{(i-1)}(k)} \\ &= \frac{p_{c_t}[1]^{x_t[1]} \cdot \dots \cdot p_{c_t}[d]^{x_t[d]} \cdot \pi_{c_t}^{(i-1)}}{\sum_{k=1}^K p_k[1]^{x_t[1]} \cdot \dots \cdot p_k[d]^{x_t[d]} \cdot \pi_k^{(i-1)}} \end{aligned}$$

M-step

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

M-step

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

M-step

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

M-step

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

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

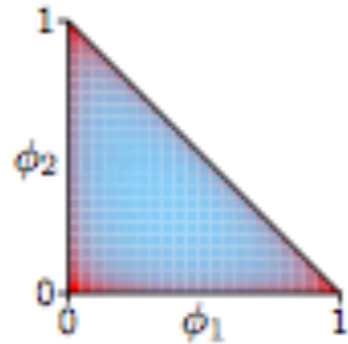
- Its a distribution over distributions!
- Parameters $\alpha_1, \dots, \alpha_K$ s.t. $\alpha_k > 0$
- The density function is given as

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

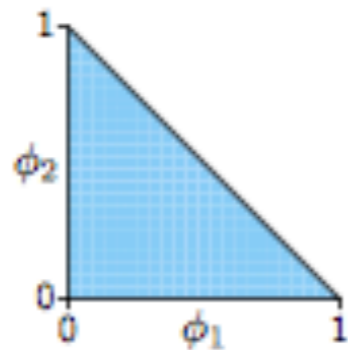
where $B(\alpha) = \prod_{k=1}^K \Gamma(\alpha_k) / \Gamma(\sum_{k=1}^K \alpha_k)$

DIRICHLET DISTRIBUTION

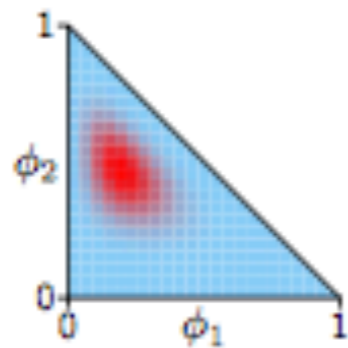
Dirichlet(.5,.5,.5)



Dirichlet(1,1,1)



Dirichlet(5,10,8)



LATENT DIRICHLET ALLOCATION

- Generative story:
 - For $t = 1$ to n
 - For each customer draw mixture of types $\pi_t \sim \text{Dirichlet}(\alpha)$
 - 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, \dots, p_K