Machine Learning for Data Science (CS4786)
Lecture 17

Latent Dirichlet Allocation & Intro to Graphical Models

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016sp/

Announcements

- Assignment A2 is out.
- Due date: April 15th, 11:59pm
- Consists of two simple problems
- Group size: 1-4, groups not transferred from last time
- Competition I will begin soon and you will have 3 weeks.

Probabilistic Models

- Set Θ consists of parameters s.t. P_θ is the distribution over the random variables by each $\theta \in \Theta$
- Data is generated by one of the $\theta \in \Theta$
- Learning: Estimate value or distribution for $\theta^* \in \Theta$ given data (we saw MLE and talked about MAP)

Maximum Likelihood Principal

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

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

EM Algorithm

(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)

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

- x_t observation, c_t latent variable.

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 $x_t \sim \mathrm{multinomial}(p_{c_t})$

Mixture of Multinomials

What is missing in this story?

- Every customer could be a bit of every type, or at least a few types
- Another example is modeling documents based on words contained in them.
- A document could belong to multiple topics (unline clustering)

Latent Dirichlet Allocation

- Each document has a mixture of topics
- Every word in each document is assigned a specific topic
- How do we model this?

n. # of documents
d. size of lexicon
m. # of words in each document
k. # of topics

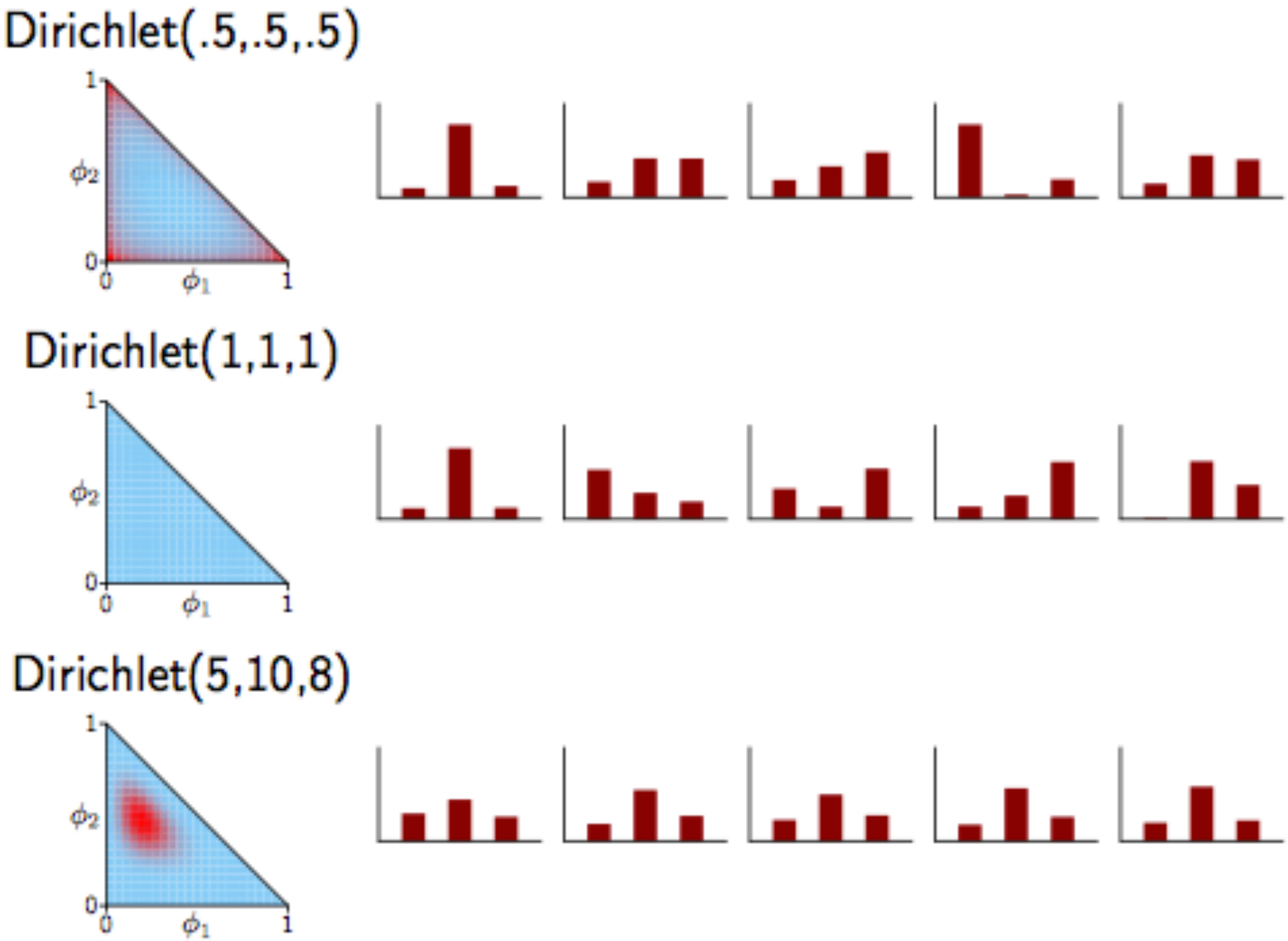
Dirichlet Distribution

- 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 $B(\alpha) = \prod_{k=1}^{K} \Gamma(\alpha_k) / \Gamma(\sum_{k=1}^{K} \alpha_k)$
- $K = 2$ its called β distribution
- For each document we draw π from a Dirchlet distribution (each customer is a mixture of the various types)

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 $x_t[i] \sim p_{c_t[i]}$

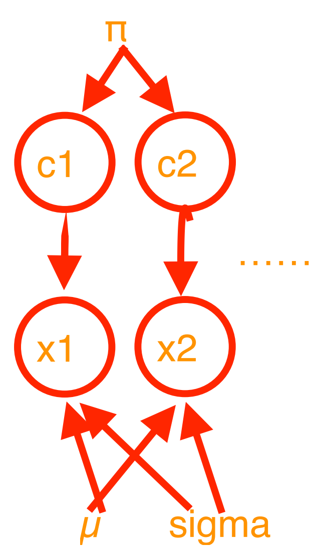
  End For

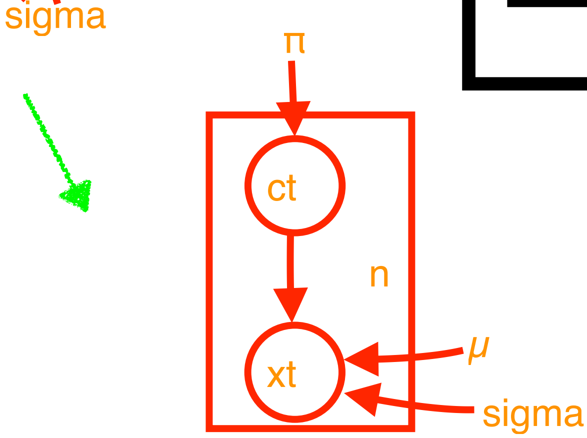
  End For

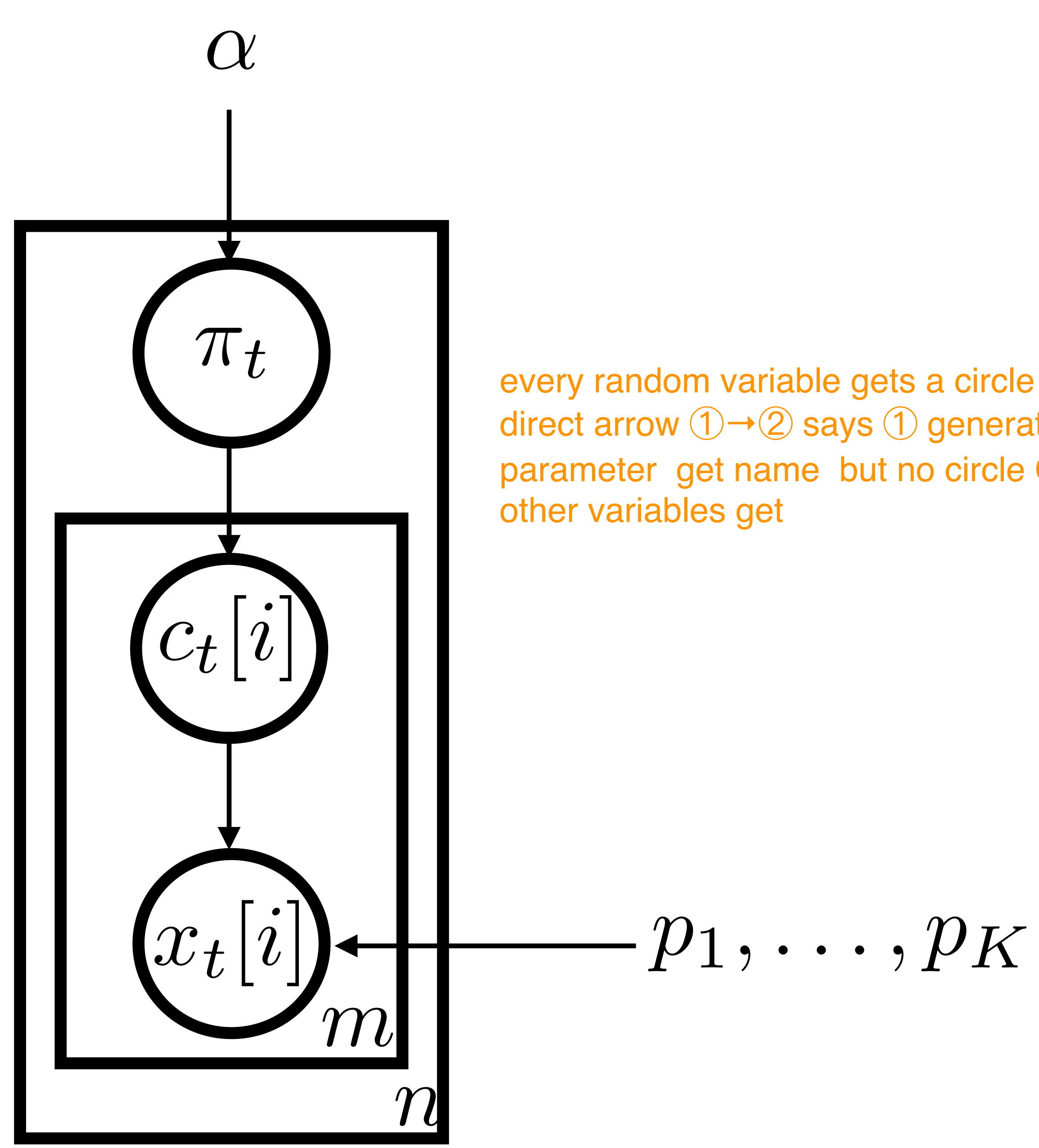
- Parameters, α for the Dirichlet distribution and $p_1, \ldots, p_K$ the distributions for each time over the d items.

Latent Dirichlet Allocation

Gausion distribution
π
μ ~ μ1 to μk
sigma ~ sigma1 to sigmak







every random variable gets a circle O with name in it
direct arrow ①→② says ① generates ②
parameter get name but no circle O
other variables get