CS598JHM: Advanced NLP (Spring 2013)
http://courses.engr.illinois.edu/cs598jhm/

Lecture 13: Dirichlet Processes

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

Office hours: by appointment

Finite mixture model

Mixing proportions:
The prior probability of each component (assuming uniform α)

$\pi|\alpha \sim \mathit{Dirichlet}(\alpha/K, ..., \alpha/K)$

Mixture components:
The distribution over observations for each component

$\theta^*_k|H \sim H$ (H is typically a Dirichlet distribution)

Indicator variables:
Which component is observation i drawn from?

$z_i|\pi \sim \mathit{Multinomial}(\pi)$

The observations:
The probability of observation i under component z_i

$x_i|z_i, \{ \theta^*_k \} \sim F(\theta^*_{z_i})$ (F is typically a categorical distribution)

Dirichlet Process DP(α, H)

The **Dirichlet process** DP(α, H) defines a distribution over distributions over a probability space $\boldsymbol{\Theta}$.

Draws $G \sim \mathrm{DP}(\alpha, H)$ from this DP are **random distributions** over $\boldsymbol{\Theta}$

DP(α, H) has two parameters:

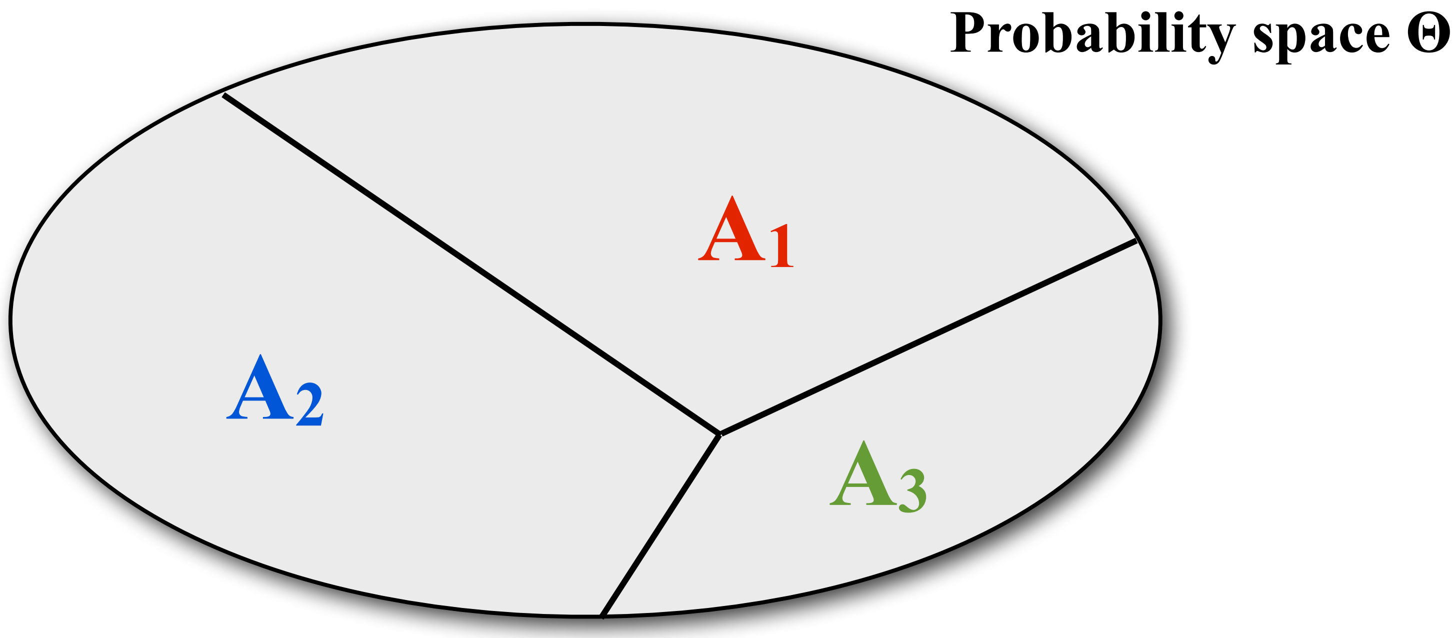
Base distribution H:
a distribution over the probability space $\boldsymbol{\Theta}$

Concentration parameter α:
a positive real number

If $G \sim \mathrm{DP}(\alpha, H)$, then for any finite measurable partition $A_1 \ldots A_r$ of $\boldsymbol{\Theta}$:

$$(G(A_1), \ldots, G(A_r)) \sim \mathit{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_r))$$

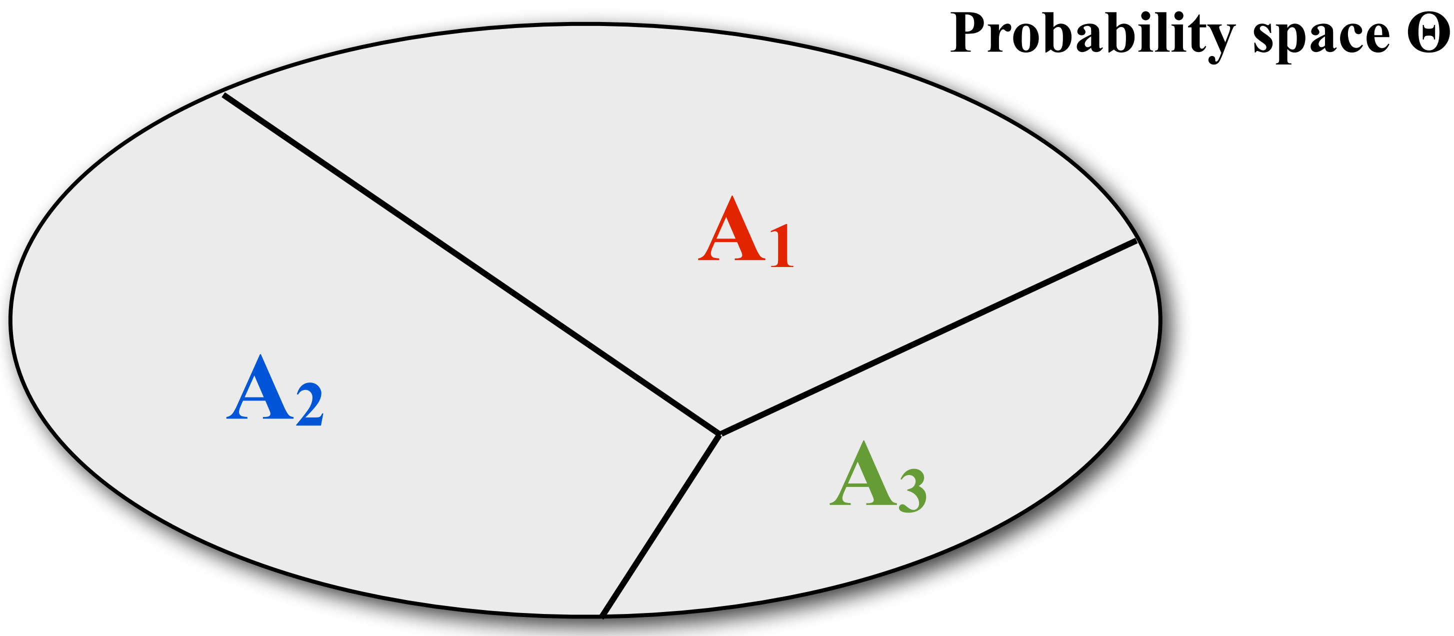
The base distribution H



Since A_1, A_2, A_3 partition $\mathbf{\Theta}$, we can use the base distribution H to define a categorical distribution over A_1, A_2, A_3:
$H(A_1) + H(A_2) + H(A_3) = 1$

Note that we can use H to define a categorical distribution over *any* finite partition $A_1 \ldots A_r$ of $\mathbf{\Theta}$, even if H is smooth

Draws from the DP: $G \sim \mathrm{DP}(\alpha, H)$



Every individual draw G from $\mathrm{DP}(\alpha, H)$ is also a distribution over $\boldsymbol{\Theta}$
G also defines a categorical distribution over any partition of $\boldsymbol{\Theta}$

For *any* finite partition $A_1 \ldots A_r$ of $\boldsymbol{\Theta}$, this categorical distribution is drawn from a Dirichlet prior defined by α and H:

$$(G(A_1),\; G(A_2),\; G(A_3)) \sim Dir(\alpha H(A_1),\, \alpha H(A_2),\; \alpha H(A_3))$$

The role of H and α

The base distribution H defines the **mean** (expectation) of G:

For any measurable set $A \subseteq \boldsymbol{\Theta}$, $E[G(A)] = H(A)$

The concentration parameter α is **inversely** related to the **variance** of G :

$V[G(A)] = H(A)(1 - H(A))/(\alpha + 1)$

α specifies how much mass is around the mean

The larger α, the smaller the variance

α is also called the **strength parameter:** If we use $DP(\alpha, H)$ as a prior, α tells us how much we can deviate from the prior:

As $\alpha \rightarrow \infty$, $G(A) \rightarrow H(A)$

The posterior of G: $G|\theta_1, \ldots \theta_n$

Assume the distribution G is drawn from a DP: $G \sim \mathrm{DP}(\alpha, H)$

The **prior** of G:

$$(G(\mathrm{A}_1),\ldots, G(\mathrm{A}_K)) \sim \mathit{Dirichlet}(\alpha H(\mathrm{A}_1), \ldots, \alpha H(\mathrm{A}_K))$$

Given a sequence of observations $\theta_1 \ldots \theta_n$ from $\boldsymbol{\Theta}$
that are drawn from this G: $\theta_i|G \sim G$
What is the **posterior of** G given the observed $\theta_1 \ldots \theta_n$?

For any finite partition $A_1 \ldots A_K$ of $\boldsymbol{\Theta}$**,**
define the number of observations in A_k : $n_k = \#\{\, i: \theta_i \in \mathrm{A}_k \,\}$

The **posterior** of G given observations $\theta_1 \ldots \theta_n$

$$(G(\mathrm{A}_1),\ldots, G(\mathrm{A}_K))|\theta_1, \ldots \theta_n \sim \mathit{Dirichlet}(\alpha H(\mathrm{A}_1) + n_1, \ldots, \alpha H(\mathrm{A}_K) + n_K)$$

The posterior of G: $G|\theta_1, \ldots \theta_n$

The observations $\theta_1 \ldots \theta_n$ define an **empirical distribution** over $\boldsymbol{\Theta}$:

$$\frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}$$

←This is just a fancy way of saying $P(A_k) = n_k/n$

The **posterior** of G given observations $\theta_1 \ldots \theta_n$

$$(G(A_1), \ldots, G(A_K))|\theta_1, \ldots \theta_n \sim \mathit{Dirichlet}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_K) + n_K)$$

The posterior is a DP with:

- **concentration parameter** $\alpha + n$
- a **base distribution** that is a weighted average of H and the empirical distribution.

$$G|\theta_1, \ldots \theta_n \sim DP(\alpha + n, \quad \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n})$$

The weight of the empirical distribution is proportional to the amount of data. The weight of H is proportional to α

The Blackwell MacQueen urn

Assume each value in Θ has a unique color.

$\theta_1 ... \theta_n$ is a sequence of colored balls.

With probability $\alpha / (\alpha + n)$, the $n+1$th ball is drawn from H

With probability $n/(\alpha+n)$ the $n+1$th ball is drawn from an urn that contains all previously drawn balls.

Note that this implies that G is a discrete distribution, even if H is not.

The clustering property of DPs

$\theta_1 \ldots \theta_n$ induces a partition of the set $1 \ldots n$ into k unique values.

This means that the DP defines a distribution over such partitions.

The expected number of clusters k increases with α but grows only logarithmically in n:

$$E[\, k \mid n] \simeq \alpha \log(1 + n/\alpha)$$

NLP 101: language modeling

Task: Given a stream of words $w_1 \ldots w_n$, predict the next word w_{n+1} with a unigram model $P(w)$

Answer:

If w_{n+1} is a word w we've seen before:

$$P(w_{n+1} = w) \propto \mathrm{Freq}(w)$$

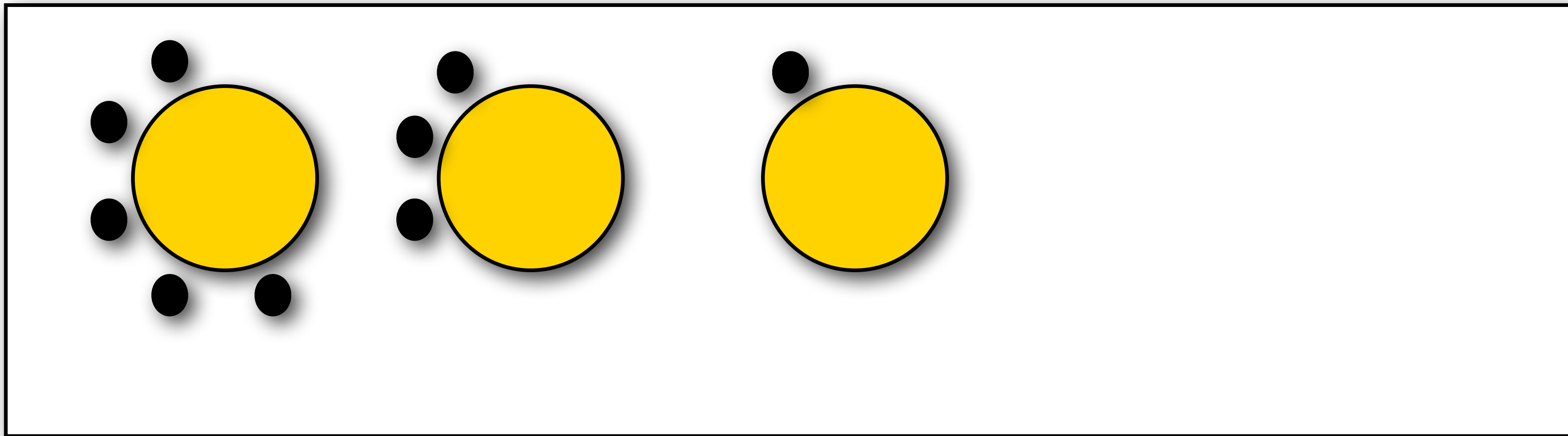
But what if w_{n+1} has never been seen before?

We need to reserve some mass for new events

$$P(w_{n+1} \text{ is a new word}) \propto \alpha$$

$$P(w_{n+1} = w) = \begin{cases} \mathrm{Freq}(w)/(n+\alpha) & \text{if } \mathrm{Freq}(w) > 0 \\ \alpha/(n+\alpha) & \text{if } \mathrm{Freq}(w) = 0 \end{cases}$$

The Chinese restaurant processs



The (i+1)th customer c_{i+1} sits:

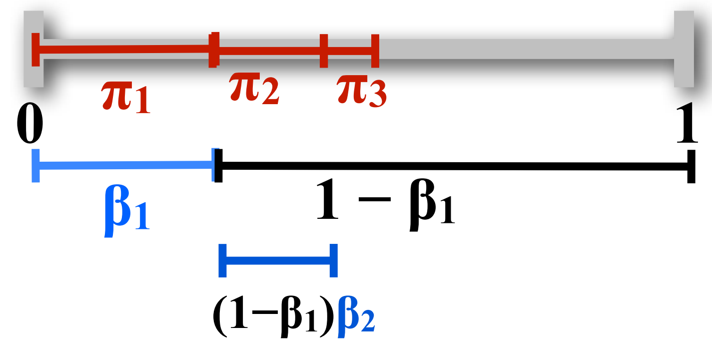
- at an *existing* table t_k that already has n_k customers with probability $n_k/(i+\alpha)$
- at *new* table with probability $\alpha/(i+\alpha)$

The predictive distribution $\theta_{n+1} | \theta_1, \ldots, \theta_n$

The predictive distribution of θ_{n+1} given a sequence of i.i.d. draws $\theta_1, \ldots, \theta_n \sim G$, with $G \sim DP(\alpha, H)$ and G marginalized out is given by the posterior base distribution given $\theta_1, \ldots, \theta_n$

$$
\begin{aligned}
P(\theta_{n+1} \in A) &= E[G(A) | \theta_1, \ldots, \theta_n] \\
&= \frac{\alpha}{\alpha + n} H(A) + \frac{\sum_{i=1}^{n} \delta_{\theta_i}(A)}{\alpha + n}
\end{aligned}
$$

The stick-breaking representation



$G \sim DP(\alpha, H)$ if:

- The component parameters are drawn from the base distribution: $\theta^*_k \sim H$
- The weights of each cluster are defined by a stick-breaking process:

$$\beta_k \sim Beta(1, \alpha) \qquad \pi_k \sim \beta_k \prod_{l=1...k-1} (1-\beta_l)$$

also written as $\pi \sim \mathrm{GEM}(\alpha)$ (Griffiths/Engen/McCloskey)

$$G = \sum_{k=1...\infty} \pi_k \delta$$

θ^*_k

Dirichlet Process Mixture Models

Each observation x_i is associated with a latent parameter θ_i
Each θ_i is drawn i.i.d. from G; each x_i is drawn from $F(\theta_i)$

$$G|\alpha,H \sim DP(\alpha, H) \qquad \theta_i|G \sim G \qquad x_i|\theta_i \sim F(\theta_i)$$

Since G is discrete, θ_i can be equal to θ_j
All x_i, x_j with $\theta_i = \theta_j$ belong to the same mixture component
There are a countably infinite number of mixture components.

Stick-breaking representation:

Mixing proportions: $\pi|\alpha \sim GEM(\alpha)$
Indicator variables: $z_i|\pi \sim Mult(\pi)$
Component parameters: $\theta^*_k|H \sim H$
Observations: $x_i|z_i,\{\theta^*_k\} \sim F(\theta^*_{z_i})$

Hierarchical Dirichlet Processes

Since both H and G are distributions over the same space $\mathbf{\Theta}$, the base distribution of a DP can be a draw from another DP. This allows us to specify hierarchical Dirichlet Processes, where each group of data is generated by its own DP.

Assume a global measure G_0 drawn from a DP:

$G_0 \sim \mathrm{DP}(\gamma, H)$

For each group j, define another DP G_j with base measure G_0:

$G_j \sim \mathrm{DP}(\alpha_0, G_0)$

(or $G_j \sim \mathrm{DP}(\alpha_j, G_0)$), but it is common to assume all α_j are the same)

α_0 specifies the amount of variability around the prior G_0

Since all groups share the same base G_0, all G_j use the same atoms (balls of the same colors)