Nonparametric Bayesian Methods

Finite and infinite mixtures

Dirichlet Process (DP) and stick breaking

Gibbs sampling

December 19, 2019

Recall - Beta Distribution

$$\text{Beta}(x|a,b) = \frac{1}{B(a,b)} \cdot x^{a-1}(1-x)^{b-1},$$

where $x \in [0,1]; a, b > 0$

$B(a,b)$ is the **Beta function**:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \text{with}$$

$$\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$$

Interpretation: probability of a Bernoulli process after observing $a-1$ successes and $b-1$ failures

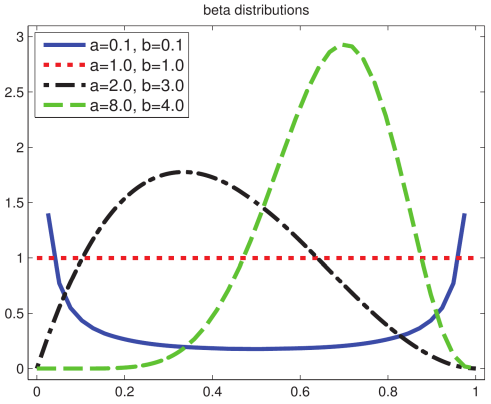


Figure: Some Beta distributions.
Kevin Murphy (2012). Machine Learning - A Probabilistic Perspective.

Dirichlet Distribution

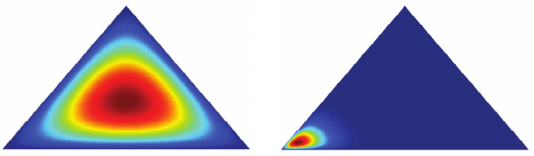
Multivariate generalization of the beta distribution.

Given $\boldsymbol{x} = x_1, ..., x_n$ and $\boldsymbol{\alpha} = \alpha_1, ..., \alpha_n$, where $x_i \in [0, 1]; \alpha_i > 0$,

$$\text{Dir}(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \cdot \prod_{k=1}^{n} x_k^{\alpha_k - 1},$$

where $B(\boldsymbol{\alpha})$ is the multivariate generalization of the beta function:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{n} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{n} \alpha_k)}$$



(a) $\boldsymbol{\alpha} = (2, 2, 2)$ (b) $\boldsymbol{\alpha} = (20, 2, 2)$

Figure: Dirichlet distributions.
Kevin Murphy (2012). Machine Learning - A Probabilistic Perspective.

Recall - Finite Gaussian Mixture Model

Fixed, finite number of clusters K.

Centers of the clusters:

$$\mu_k \sim \mathcal{N}(\mu_0, \sigma_0)$$

Probabilities of clusters (parameters):

$$\rho_{1...K} \sim \text{Dir}(\alpha_{1...K})$$

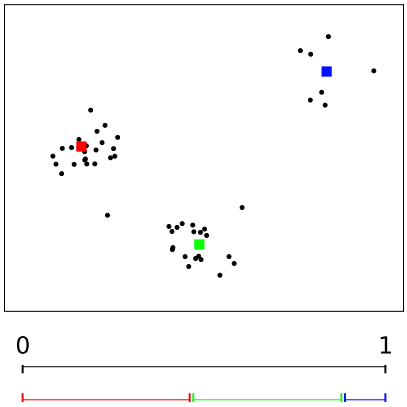
Assignments to clusters:

$$z_i \sim \text{Categorical}(\rho_{1...K})$$

Coordinates of data points:

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i})$$

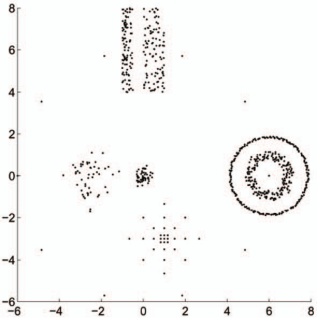
Note that Dir() is finite $\rightarrow$ realize K clusters with probability 1



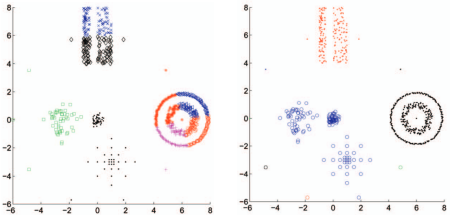
Selecting K - What are the issues?

One of the main issues is what K to select.

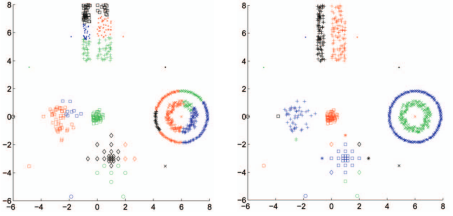
How many clusters are here?



Example from Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6), 835-850.



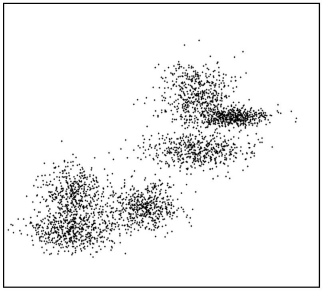
(a) $K = 8$ (b) $K = 8$



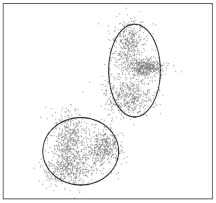
(c) $K = 22$ (d) $K = 27$

Selecting K - More Issues

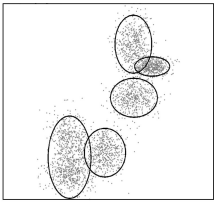
And how many here?



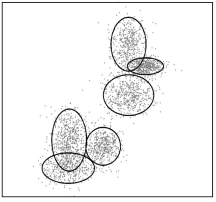
Example from Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.



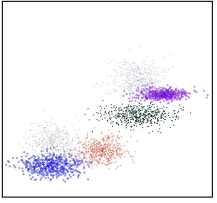
(a) $K = 2$



(b) $K = 5$



(c) $K = 6$



(d) True Labels

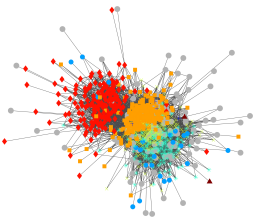
Selecting K - adaptation

Sometimes unknown in advance how many clusters:

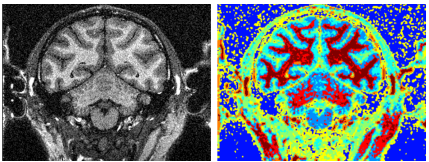
- Example: movie genres. How many are there? Action, adventure, comedy, ...? French, Italian, British movies,...? Movies with a specific actor?
- Other examples: topics of documents, communities in graphs, image segmentation
- Streaming data: difficult to estimate how many clusters if I don't see all data in advance

Naive solution: select a K, cluster with EM, evaluate the result, iterate...

Other naive solution: Can we just select a large enough K?



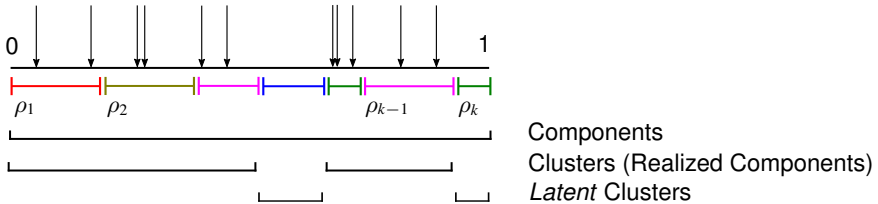
(a) Social Network communities. *Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of Facebook networks. Physica A: Stat Mech Appl, 391(16), 4165-4180.*



(b) Image segmentation. *Orbanz, P. (2008). Diss. ETH No. 17822*

Latent Clusters

For a finite number of drawings N, we do not have to realize **all** K clusters



All K components will be realized with probability 1, but only when $N \to \infty$

Selecting K

- Handle the need of new clusters: select a large K
  - Will only realize *some* components, and get more when needed
- Is the problem solved then? Only partially:
  - How large should this K be? Any specific K might have problems
  - Our belief in K could change as we observe more data points
  - Still have issue with streaming /growing data
- Better solution: select $K = \infty$

Selecting $K = \infty$

- When $K = \infty$ we have nonparametric Bayesian methods
  - Where "nonparametric" means we have *infinitely* many parameters
  - We can keep drawing new parameters

- But how? Recall, parameters = cluster probabilities: $\rho_{1...K} \sim \text{Dir}(\alpha_{1...K})$
  - Cannot draw infinite points from Dir()
  - How to get infinite probabilities that sum to 1?
  - We need: (i) a suitable distribution, (ii) a way to sample from it

Dirichlet Processes

A Dirichlet Process $DP(\alpha, H)$ is a distribution over probability distributions on a space Θ. Here

- $\alpha \in \mathbb{R}_{>0}$ is the *concentration parameter*.
- H is the *base measure* on Θ.

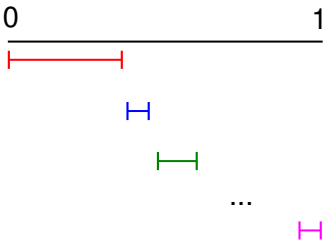
Hence, a sample $G \sim DP(\alpha, H)$ is a function $G : \Theta \to \mathbb{R}_{\geq 0}$ s.t. $\int_\Theta G(\theta) d\theta = 1$.

$DP(\alpha, H)$ is characterized by the following property: For every partition $(T_1, \ldots, T_k)$ of Θ and $G \sim DP(\alpha, H)$, we have

$$(G(T_1), \ldots, G(T_K)) \sim \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K)).$$

Stick–Breaking Process

Observation: Sampling $(\rho_1, \ldots, \rho_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$ is equivalent to sampling $\rho_1 \sim \text{Beta}(\alpha_1, \alpha_2 \cdots \alpha_K)$ and $(\rho_2, \ldots, \rho_K) \sim \text{Dir}(\alpha_2, \ldots, \alpha_K)$



$$\beta_1 \sim \text{Beta}(\alpha_1, \alpha_{2\ldots K}) \qquad \rho_1 = \beta_1$$

$$\beta_2 \sim \text{Beta}(\alpha_2, \alpha_{3\ldots K}) \qquad \rho_2 = \beta_2(1-\beta_1)$$

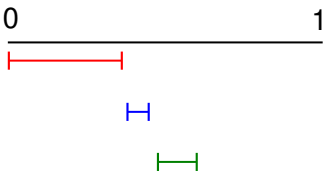
$$\beta_3 \sim \text{Beta}(\alpha_3, \alpha_{4\ldots K}) \qquad \rho_3 = \beta_3(1-\beta_1)(1-\beta_2)$$

$$\rho_K = 1 - \textstyle\sum_{i=1}^{K-1} \rho_i$$

We can continue doing this – but only with $\alpha = (\alpha_1, \ldots, \alpha_K)$ of finite length K.

Stick–Breaking

Solution: fix α s.t. $\beta_i \sim \text{Beta}(1, \alpha) \quad \forall i$. Revised construction:



$$\beta_1 \sim \text{Beta}(1, \alpha) \qquad \rho_1 = \beta_1$$
$$\beta_2 \sim \text{Beta}(1, \alpha) \qquad \rho_2 = \beta_2(1 - \beta_1)$$
$$\beta_3 \sim \text{Beta}(1, \alpha) \qquad \rho_3 = \beta_3(1 - \beta_1)(1 - \beta_2)$$
$$\beta_k \sim \text{Beta}(1, \alpha) \qquad \rho_k = \beta_k(1 - \sum_{i=1}^{k-1} \rho_i)$$

We call this the GEM distribution: $\rho \sim \text{GEM}(\alpha), \rho = \{\rho_k\}_{k=1}^{\infty}$
(Griffiths–Engen–McCloskey distribution)

Stick–Breaking Construction of the Dirichlet Process

Connection to DP: If $\rho \sim GEM(\alpha)$ and $\theta_k \sim H$ for $k = 1, 2, \ldots$, then

$$G(\theta) = \sum_{k=1}^{\infty} \rho_k \delta_{\theta_k}(\theta)$$

is a sample from $DP(\alpha, H)$. In particular, samples from a DP are almost surely discrete measures on Θ, supported on a countable set $\{\theta_k\}_{k=1}^{\infty} \subset \Theta$.

If we repeatedly sample $\theta^{(1)}, \theta^{(2)}, \ldots$ from $G \sim DP(\alpha, H)$ (recall: G is a probability distribution on Θ), we have

$$\theta^{(i)} = \theta_{k_i}$$

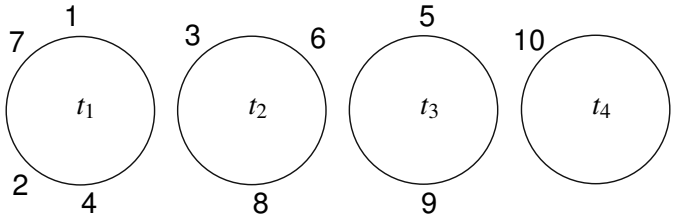
for some k_i. We sometimes get a value not observed before (i.e., $k_i \neq k_j$ for all $j < i$), and sometimes a repetition of a previous value (i.e., $k_i = k_j$ for some $j < i$).

Think of $\theta^{(i)}, \theta^{(j)}$ with $k_i = k_j$ as data points belonging to the same cluster.

Chinese Restaurant Process - Overview

- Technique to draw samples from a DP
- Metaphor: customers are observations $\theta^{(i)}$, tables are clusters θ_k
- When a new customer arrives, he can either:
  - Join an existing table with probability $\propto$ the number of people sitting there
  - Start a new table with probability $\propto \alpha$

Chinese Restaurant Process - Formulation



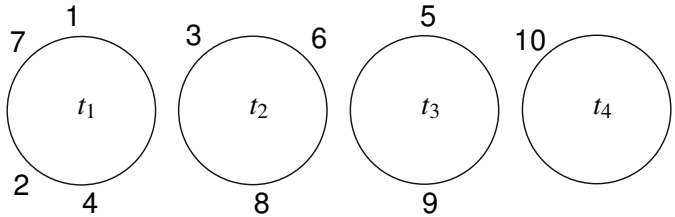
Sample table assignment above is a **partition over the integers**:

$\mathcal{P} = \{\{1, 2, 4, 7\}, \{3, 6, 8\}, \{5, 9\}, \{10\}\}$

Given this notation, define the CRP as follows:

$$P(\text{customer } n+1 \text{ joins table } \tau \mid \mathcal{P}) = \begin{cases} \frac{|\tau|}{\alpha+n} & \text{if } \tau \in \mathcal{P}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

Chinese Restaurant Process - Example

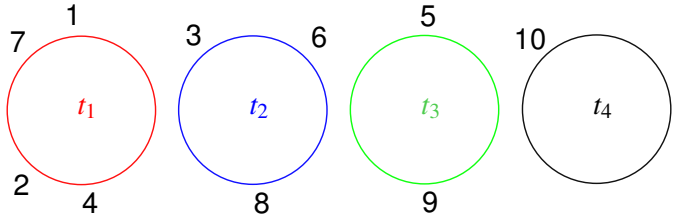


$$P(\text{customer } n+1 \text{ joins table } \tau \mid \mathcal{P}) = \begin{cases} \frac{|\tau|}{\alpha+n} & \text{if } \tau \in \mathcal{P}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

$$\mathcal{P} = \{\{1,2,4,7\},\{3,6,8\},\{5,9\},\{10\}\}$$

$$P(\mathcal{P}) = \frac{\alpha}{\alpha}\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{2}{\alpha+3}\right)\left(\frac{\alpha}{\alpha+4}\right)\left(\frac{1}{\alpha+5}\right)\left(\frac{3}{\alpha+6}\right)\left(\frac{2}{\alpha+7}\right)\left(\frac{1}{\alpha+8}\right)\left(\frac{\alpha}{\alpha+9}\right)$$

Chinese Restaurant Process - Exchangeability

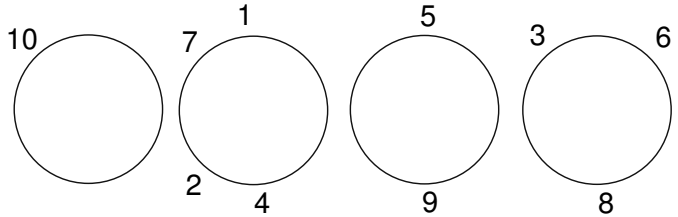


$$\mathcal{P} = \{\{1,2,4,7\},\{3,6,8\},\{5,9\},\{10\}\}$$

$$P(\mathcal{P}) = \frac{\alpha}{\alpha}\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{2}{\alpha+3}\right)\left(\frac{\alpha}{\alpha+4}\right)\left(\frac{1}{\alpha+5}\right)\left(\frac{3}{\alpha+6}\right)\left(\frac{2}{\alpha+7}\right)\left(\frac{1}{\alpha+8}\right)\left(\frac{\alpha}{\alpha+9}\right)$$

$$P(\mathcal{P}) = \frac{\alpha^{|\mathcal{P}|}}{\alpha^{(n)}} \prod_{\tau \in \mathcal{P}} (|\tau| - 1)!$$

Chinese Restaurant Process - Exchangeability



$$\mathcal{P} = \{\{10\}, \{1, 2, 4, 7\}, \{5, 9\}, \{3, 6, 8\}\}$$

$$P(\mathcal{P}) = \frac{\alpha}{\alpha} \left(\frac{\alpha}{\alpha+1}\right) \left(\frac{1}{\alpha+2}\right) \left(\frac{2}{\alpha+3}\right) \left(\frac{3}{\alpha+4}\right) \left(\frac{\alpha}{\alpha+5}\right) \left(\frac{1}{\alpha+6}\right) \left(\frac{\alpha}{\alpha+7}\right) \left(\frac{1}{\alpha+8}\right) \left(\frac{2}{\alpha+9}\right)$$

$$P(\mathcal{P}) = \frac{\alpha^{|\mathcal{P}|}}{\alpha^{(n)}} \prod_{\tau \in \mathcal{P}} (|\tau| - 1)!$$

Exchangeable: it is order– and labeling–independent!

Chinese Restaurant Process - Clusters

How many tables does the process create for N?

Equivalent: how many clusters are realized?

$$P(\mathcal{P}) = \frac{\alpha}{\alpha}\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)\left(\frac{2}{\alpha+3}\right)\left(\frac{3}{\alpha+4}\right)\left(\frac{\alpha}{\alpha+5}\right)\left(\frac{1}{\alpha+6}\right)\left(\frac{\alpha}{\alpha+7}\right)\left(\frac{1}{\alpha+8}\right)\left(\frac{2}{\alpha+9}\right)$$

Mark 1 when a customer creates a table, 0 otherwise.

$$E\mathbb{1} = \sum_{i=1}^{N} \frac{\alpha}{\alpha + i} \sim O(\alpha \log(N))$$

Rich-get-richer effect *(preferential attachment)*: already "popular" clusters attract more new data points

Expected cluster count is the property of the DP as a prior (without seeing data)

Exchangeability - Definition

- It's clear how CRP relates to clustering
  - Less clear how it precisely relates to the DP
- Let's see some more details on the notion of exchangeability
- Formal definition: let $(X_1, X_2, ...)$ be a sequence of random variables. The sequence is exchangeable when, for every permutation π of $\mathbb{N}$, the random vectors

$$(X_1, X_2, \dots) \quad \text{and} \quad (X_{\pi(1)}, X_{\pi(2)}, \dots)$$

have the same distribution.

Exchangeability - De Finetti's Theorem

De Finetti's Theorem:

Let $(X_1, X_2, ...)$ be an infinitely exchangeable sequence of random variables.

Then, $\forall n$:

$$p(X_1, ..., X_n) = \int \left(\prod_{i=1}^{n} p(x_i|G) \right) dP(G),$$

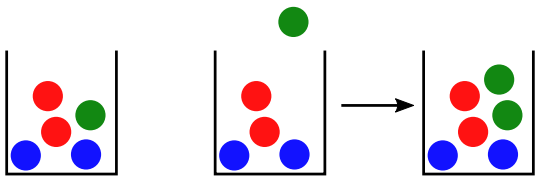
for some random variable G.

Note: an infinitely exchangeable sequence can be represented by conditionally independent random variables

Exchangeability - Pólya Urn

Consider the (multicolored) Pólya urn model:

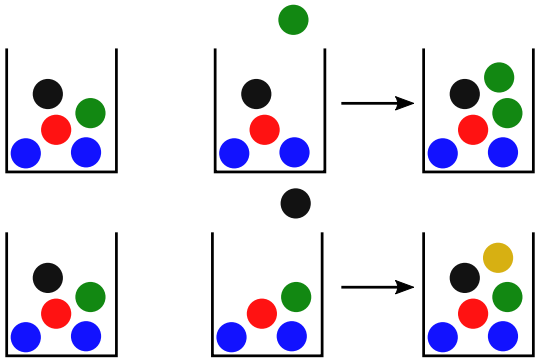
- We have an urn with colored balls
- We can draw balls from the urn at random
  - After drawing a ball, put it back in the urn **together** with a new ball of the same color



Exchangeability - Hoppe Urn

Hoppe urn: a Pólya urn with a special *black* ball. After drawing:

- A **non-black ball**: put it back in the urn with a new ball of the same color
- The **black** ball, put it back in the urn with a new ball of a **new** color



Exchangeability - Results

- Observe: the CRP is identical to the Hoppe urn process
  - We just need to add colors to the tables
- Hoppe urn and CRP are **exchangeable** $\rightarrow$ can apply De Finetti's theorem
- The DP is the r.v. G in De Finetti's theorem for Hoppe urn / CRP
- Therefore, if the prior of G is the DP, then CRP is how we assign points to clusters when we integrate out G

The DP Mixture Model

Let now Θ be a set that parametrizes a set of probability distributions, and fix a base measure H on Θ. Concrete example:

- $\Theta = \mathbb{R}$ with $\mu \in \Theta$ corresponding to $\mathcal{N}(\mu, \sigma)$ for some fixed $\sigma > 0$,
- $H = \mathcal{N}(\mu_0, \sigma_0)$ for some fixed $\mu_0 \in \mathbb{R}$, $\sigma_0 \in \mathbb{R}_{>0}$.

Based on that, we can define a generative model, the **DP Mixture Model**:

Probabilities of clusters ("mixture weights"):

$$\rho = (\rho_1, \rho_2, \dots) \sim \text{GEM}(\alpha)$$

Centers of the clusters:

$$\mu_k \sim \mathcal{N}(\mu_0, \sigma_0), \quad k = 1, 2, 3, \dots$$

Assignments of data points to clusters:

$$z_i \sim \text{Categorical}(\rho), \quad i = 1, \dots, N$$

Coordinates of data points:

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma), \quad i = 1, \dots, N$$

Fitting - Introduction

Intuition: leverage exchangeability.

- Any point can be considered the “last arrived”
- Prior: probabilities of table assignments w.r.t. people seating (= cluster size)
- Posterior: probability of the point given the cluster centers

By exchangeability, we can change the assignment of the element without influencing other assignments

Fitting - Technique

- To fit the DPMM we use a technique called Gibbs sampling
  - EM considered “difficult” for nonparametric distributions
- Idea: sample each variable in turn, conditioned on the values of all the other variables in the distribution
  - *(This is why we need exchangeability)*

Fitting - Probability Distribution

Collapsed Gibbs sampling formulation:

$$
\begin{aligned}
p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}, \alpha, \boldsymbol{\mu}) &\propto p(z_i = k|\boldsymbol{z}_{-i}, \alpha, \cancel{\boldsymbol{\mu}})\ p(\boldsymbol{x}|z_i = k, \boldsymbol{z}_{-i}, \cancel{\alpha}, \boldsymbol{\mu}) \\
&\propto p(z_i = k|\boldsymbol{z}_{-i}, \alpha)\ p(x_i|\boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\mu})\ p(\boldsymbol{x}_{-i}|\boldsymbol{z}_{-i}, \boldsymbol{\mu}) \\
&\propto \underbrace{p(z_i = k|\boldsymbol{z}_{-i}, \alpha)}_{\text{Prior}}\ \underbrace{p(x_i|\boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\mu})}_{\text{Likelihood}}
\end{aligned}
$$

For every cluster $k \in \mathcal{P}$,

where $\boldsymbol{z}_{-i}$ and $\boldsymbol{x}_{-i}$ are assignments and points *excluding* the considered point i

Fitting - Prior

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}, \alpha, \boldsymbol{\mu}) \propto p(z_i = k|\boldsymbol{z}_{-i}, \alpha)\, p(x_i|\boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\mu})$$

Prior is simple: We saw that this is the CRP; thanks to exchangeability, we can view i as the *last* client to enter the restaurant

$$p(z_i = k|\boldsymbol{z}_{-i}, \alpha) = \begin{cases} \frac{N_{k,-i}}{\alpha+N-1} & \text{for existing } k, \\ \frac{\alpha}{\alpha+N-1} & \text{otherwise,} \end{cases}$$

where $N_{k,-i}$ is the number of elements sitting at table k, excluding i

(identical to our previous $|\tau|$, excluding the considered member)

Fitting - Posterior

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}, \alpha, \boldsymbol{\mu}) \propto p(z_i = k|\boldsymbol{z}_{-i}, \alpha)\, p(x_i|\boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\mu})$$

For the posterior observe that under the condition $z_i = k$, we don't need to considerpoints in $\boldsymbol{x}$ that aren't in k.

Define $\boldsymbol{x}_{-i,c} = \{x_j : z_j = c, j \neq i\}$ the data assigned to cluster c. Then,

$$p(x_i|\boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\mu}) = \begin{cases} p(x_i|\boldsymbol{x}_{-i,k}, \boldsymbol{\mu}) = \frac{p(x_i, \boldsymbol{x}_{-i,k}|\boldsymbol{\mu})}{p(\boldsymbol{x}_{-i,k}|\boldsymbol{\mu})}, & \text{for existing } k, \\ p(x_i|\boldsymbol{\mu}), & \text{otherwise.} \end{cases}$$

Fitting - Result

Final collapsed Gibbs sampler

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}, \alpha, \boldsymbol{\mu}) = \text{Prior} \times \text{Likelihood}$$
$$= \begin{cases} \frac{N_{k,-i}}{\alpha+N-1}\, p(x_i|\boldsymbol{x}_{-i,k}, \boldsymbol{\mu}) & \text{for existing } k, \\ \frac{\alpha}{\alpha+N-1}\, p(x_i|\boldsymbol{\mu}) & \text{otherwise.} \end{cases}$$

Collapsed Gibbs sampler for DP mixtures

for $i = 1$ to N in random order **do**
  Remove x_i's sufficient statistics from old cluster z_i;
  for $k = 1$ to K **do**
    Compute $p_k(x_i) = p_k(x_i|\boldsymbol{x}_{-i,k})$;
    Set $N_{k,-i} = |\boldsymbol{x}_{-i,k}|$;
    Compute $p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}) = \frac{N_{k,-i}}{\alpha+N-1}$;
  end for
  Compute $p_\star(x_i) = p(x_i|\boldsymbol{\mu})$;
  Compute $p(z_i = \star|\boldsymbol{z}_{-i}, \boldsymbol{x})$;
  Normalize $p(z_i|\cdot)$;
  Sample $z_i \sim p(z_i|\cdot)$;
  Add x_i's sufficient statistics to new cluster z_i;
  If any cluster is empty, remove it and decrease K;
end for

Latent Dirichlet Allocation

- One of the most popular nonparametric Bayesian method
- Extension of the model we just defined
- So far we considered points generated from a univariate distribution
  - We always assumed Gaussian, but it could have been any univariate
  - Every point belonged to one and only one cluster
- What if data is generated from a *multivariate* distribution? (e.g., Dirichlet)

Latent Dirichlet Allocation – Motivation

- Motivation: topic modeling on documents
  - Each document belongs to more than one topic (is a mixture over topics $\rightarrow$ multivariate)
  - We don't know in advance how many topics there are ($\rightarrow$ nonparametric)
  - Topics are distributions over all the words
- Assumes the vocabulary is finite
- Captures the fact that words can belong to different topics (polysemy)

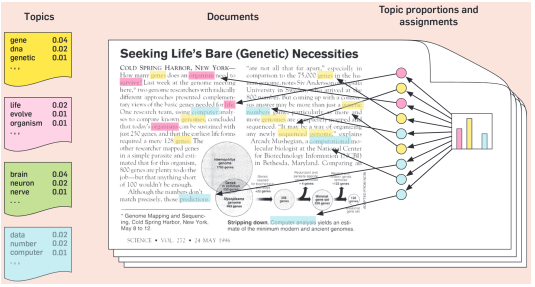


Figure: Topic Modeling.
From Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Latent Dirichlet Allocation – Model

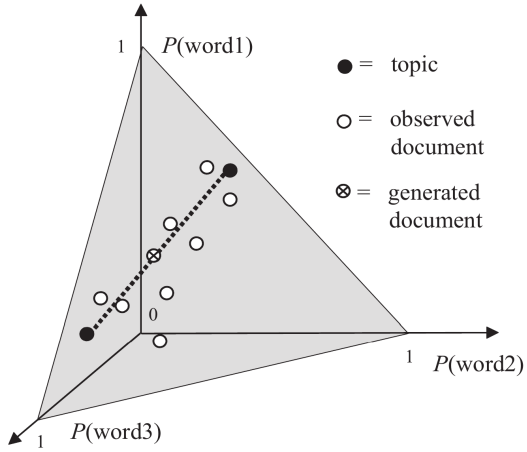


Figure: Geometric interpretation of LDA, with $K = 2$ topics and $V = 3$ words.
From Murphy (2012). Machine Learning - A Probabilistic Perspective.

Latent Dirichlet Allocation – Model

Given K topics and V words in the vocabulary, for M documents with N words each,

Distribution of topics in document d:

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})$$

What topic the word w belongs to in document d:

$$z_{d,w} \sim \text{Categorical}(\boldsymbol{\theta}_d)$$

Distribution of words in topic k:

$$\boldsymbol{\varphi}_k \sim \text{Dir}(\boldsymbol{\beta})$$

What is the word w in document d:

$$w_{d,w} \sim \text{Categorical}(\boldsymbol{\varphi}_{z_{d,w}})$$

Where α controls prior weights of topics in documents, and β controls prior weights of words in topics