L101: Machine Learning for Language Processing

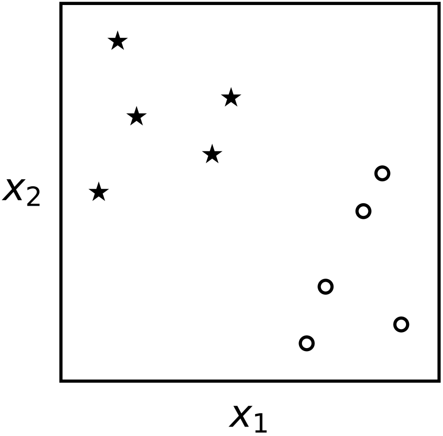
Lecture 5

Guy Emerson

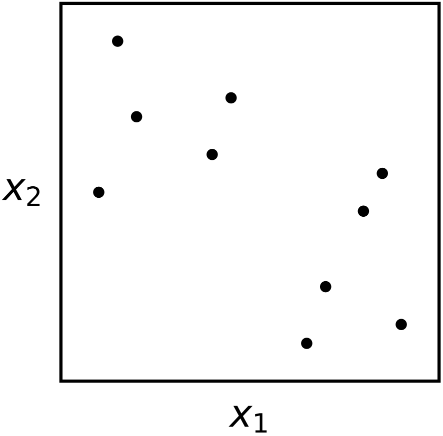
Today’s Lecture

- Unsupervised Learning
  - Word Sense Induction
  - Topic Discovery
- K-Means Clustering
- Latent Dirichlet Allocation
- Approximate Inference

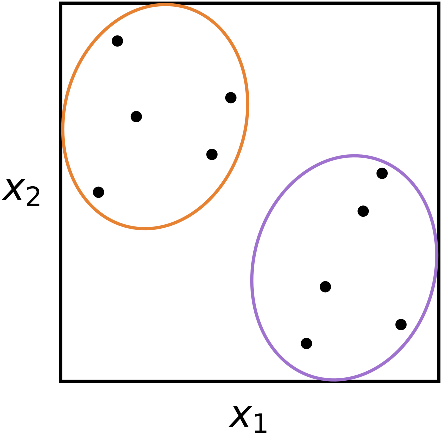
Supervised Learning



Unsupervised Learning



Unsupervised Learning



Word Senses

There was even closing drama when Shelford missed a penalty, and a chance to save the game, with the last kick of the match.

Micro-routes in the Duddon are no match, after all, for a route on any of the limestone crags in Yorkshire or Derbyshire.

Word Senses

- a thin piece of wood, ignites with friction
- a formal contest
- a burning piece of wood
- an exact duplicate
- the score needed to win
- a good matrimonial prospect
- a person of equal standing
- a pair of people who live together
- something that harmonizes

Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

Topics

This dissertation describes the measurement of angular diameters of compact radio sources by the technique of interplanetary scintillation. The design, construction and testing of a four acre radio aerial functioning at a frequency of 81.5 MHz is described, and its operation during a survey of the sky.

The stunning array of features and functions exhibited by proteins in nature should convince most scientists of the power of evolutionary design processes. Natural selection acting on populations over long periods of time has generated a vast number of proteins ideally suited to their biological functions.

Topics

This dissertation describes the measurement of angular diameters of compact radio sources by the technique of interplanetary scintillation. The design, construction and testing of a four acre radio aerial functioning at a frequency of 81.5 MHz is described, and its operation during a survey of the sky.

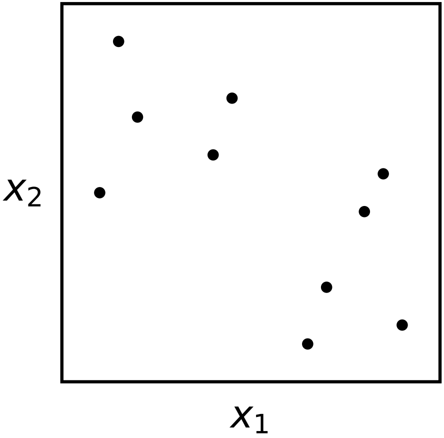
The stunning array of features and functions exhibited by proteins in nature should convince most scientists of the power of evolutionary design processes. Natural selection acting on populations over long periods of time has generated a vast number of proteins ideally suited to their biological functions.

Topics

This dissertation describes the measurement of angular diameters of compact radio sources by the technique of interplanetary scintillation. The design, construction and testing of a four acre radio aerial functioning at a frequency of 81.5 MHz is described, and its operation during a survey of the sky.

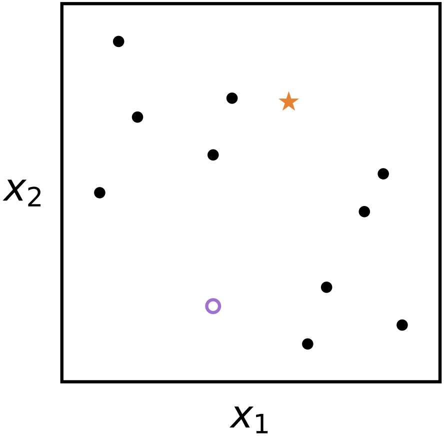
The stunning array of features and functions exhibited by proteins in nature should convince most scientists of the power of evolutionary design processes. Natural selection acting on populations over long periods of time has generated a vast number of proteins ideally suited to their biological functions.

K-Means Clustering



1. For each point, find closest cluster
2. For each cluster, find mean point

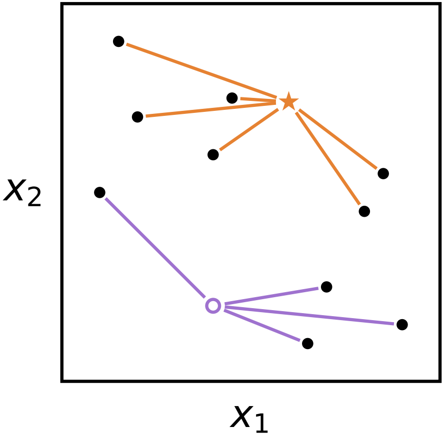
K-Means Clustering



1. For each point, find closest cluster

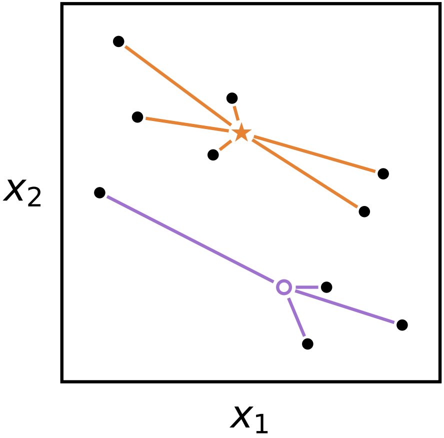
2. For each cluster, find mean point

K-Means Clustering



1. For each point, find closest cluster
2. For each cluster, find mean point

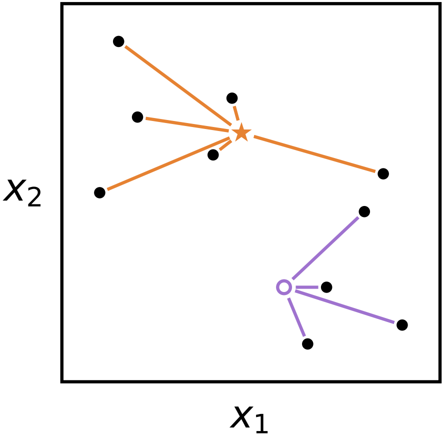
K-Means Clustering



1. For each point, find closest cluster

2. For each cluster, find mean point

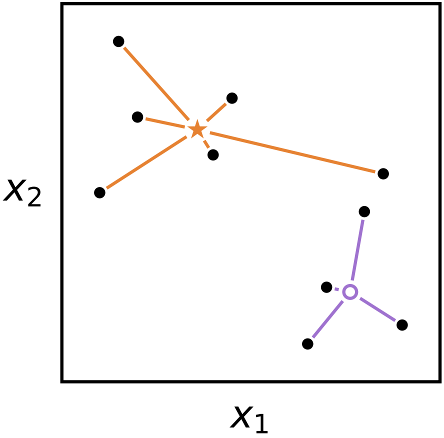
K-Means Clustering



1. For each point, find closest cluster

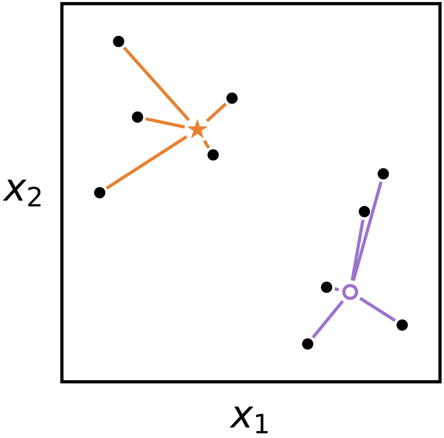
2. For each cluster, find mean point

K-Means Clustering



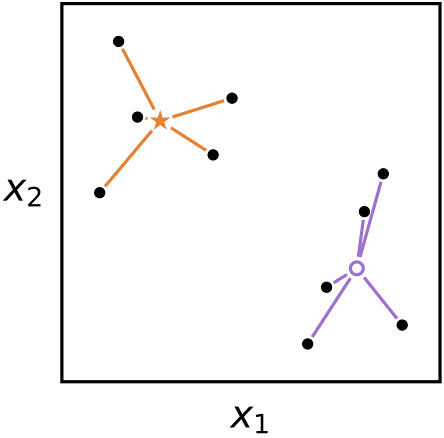
1. For each point, find closest cluster
2. For each cluster, find mean point

K-Means Clustering



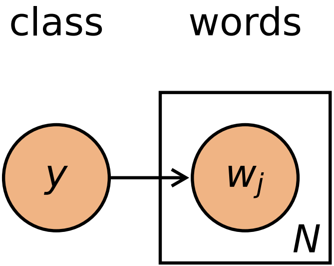
1. For each point, find closest cluster
2. For each cluster, find mean point

K-Means Clustering

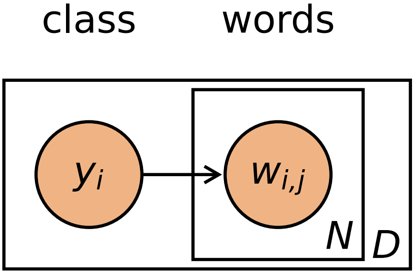


1. For each point, find closest cluster
2. For each cluster, find mean point

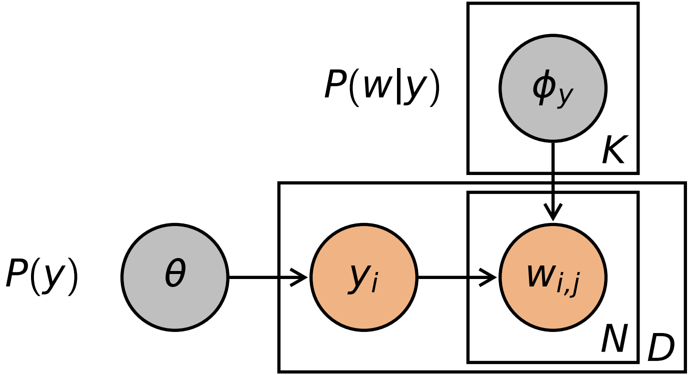
Recap: Multinomial Naive Bayes



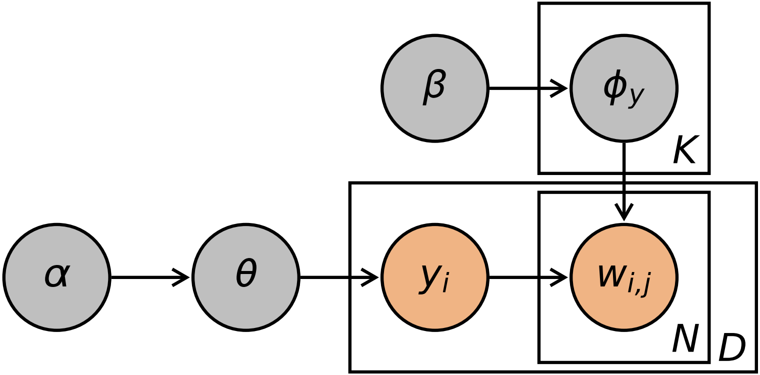
Recap: Multinomial Naive Bayes



Recap: Multinomial Naive Bayes

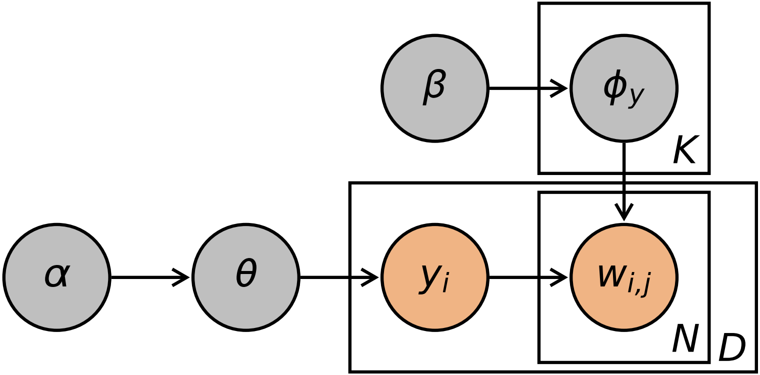


Recap: Multinomial Naive Bayes

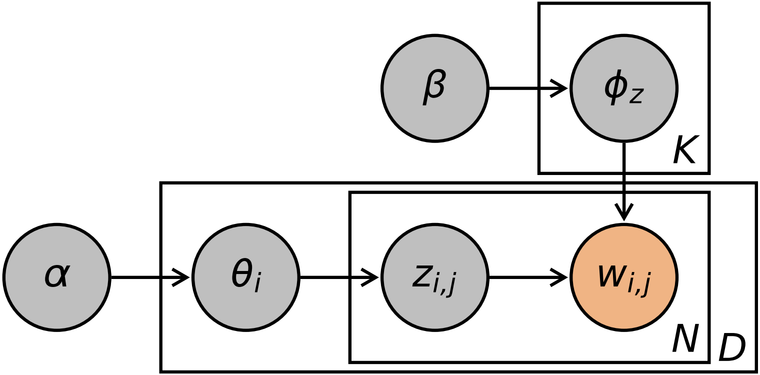


Bayesian view of smoothing hyperparameters: Dirichlet prior

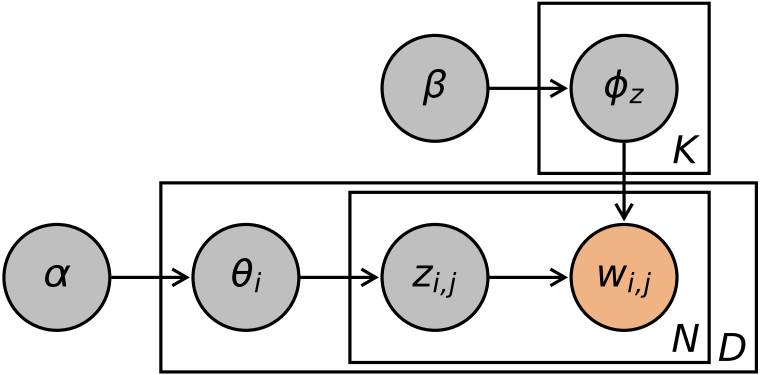
Recap: Multinomial Naive Bayes



Latent Dirichlet Allocation

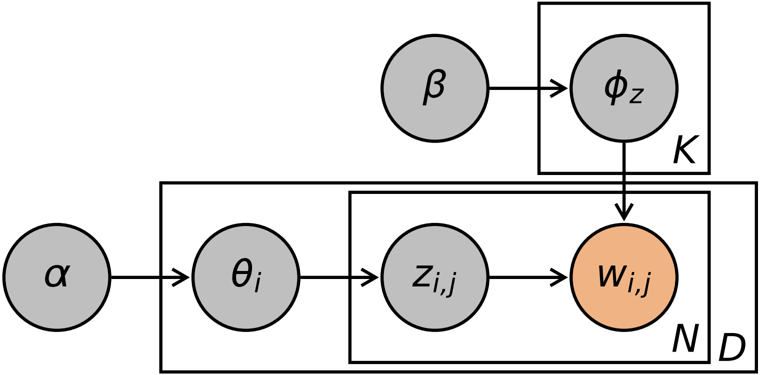


Latent Dirichlet Allocation



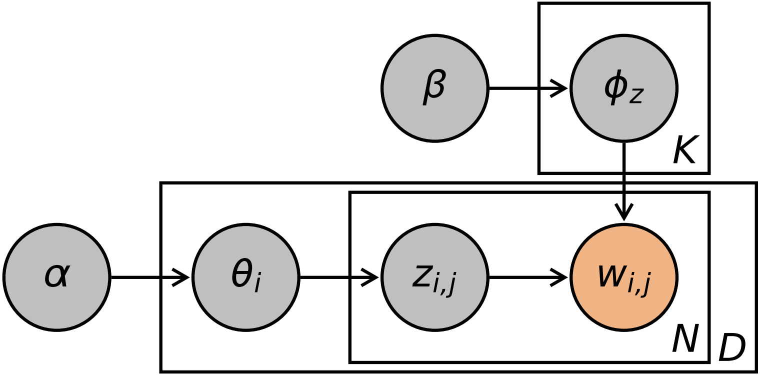
$$\prod_z P(\phi_z \mid \beta)$$

Latent Dirichlet Allocation



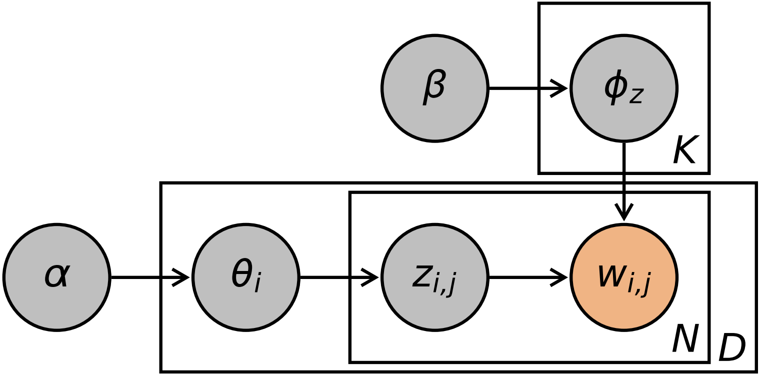
$$\prod_z P(\phi_z \mid \beta) \prod_i P(\theta_i \mid \alpha)$$

Latent Dirichlet Allocation



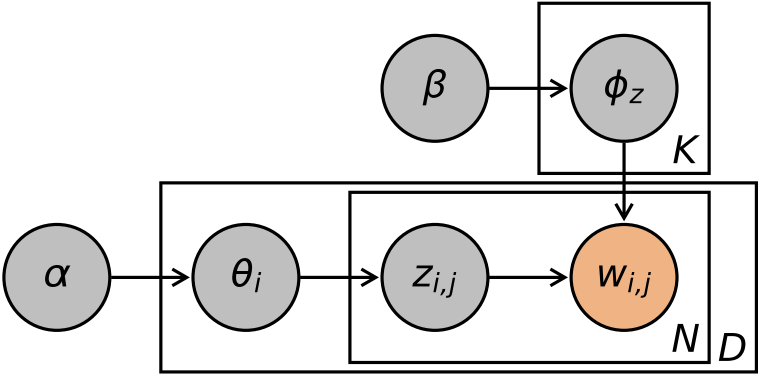
$$\prod_z P(\phi_z \mid \beta) \prod_i P(\theta_i \mid \alpha) \prod_j P\left(z_{i,j} \,\middle|\, \theta_i\right) P\left(w_{i,j} \,\middle|\, z_{i,j}\right)$$

Latent Dirichlet Allocation



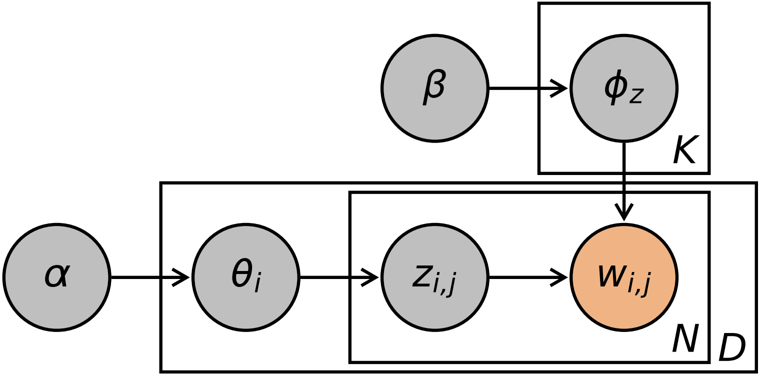
$$P\left(\phi_z, \theta_i \,\middle|\, w_{i,j}, \alpha, \beta\right)$$

Latent Dirichlet Allocation



$$P\left(\phi_z, \theta_i \,\middle|\, w_{i,j}, \alpha, \beta\right) = \sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \,\middle|\, w_{i,j}, \alpha, \beta\right)$$

Latent Dirichlet Allocation



$$P\left(\phi_z, \theta_i \middle| w_{i,j}, \alpha, \beta\right) = \sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \middle| w_{i,j}, \alpha, \beta\right)$$

Approximate Inference

- Want to know global variables (e.g. ϕ)
- Don't want to know local variables (e.g. z)
- Exact inference intractable

Markov Chain Monte Carlo

- $P(x)$ intractable

Markov Chain Monte Carlo

- $P(x)$ intractable
- Construct Markov chain converging to $P(x)$
- Sample from Markov chain

Markov Chain Monte Carlo

- $E_x\left[f(x)\right] = \sum_x P(x) f(x)$
- Construct Markov chain converging to $P(x)$
- Sample from Markov chain

Markov Chain Monte Carlo

- $E_x\left[f(x)\right] = \sum_x P(x) f(x)$
- Construct Markov chain converging to $P(x)$
- Sample from Markov chain
- $E_x\left[f(x)\right] \approx \frac{1}{N} \sum_{\text{samples}} f(x)$

Gibbs Sampling

- $P(x)$ intractable
- $P(x_1 | x_2, x_3, ...)$ tractable

Gibbs Sampling

- $P(x)$ intractable
- $P(x_1 | x_2, x_3, \ldots)$ tractable
- Markov chain:
  - Initialise x
  - Iteratively update $x_i \sim P(x_i | x_{-i})$

Gibbs Sampling

- $P(x)$ intractable
- $P(x_1 \mid x_2, x_3, \ldots)$ tractable
- Markov chain:
  - Initialise x
  - Iteratively update $x_i \sim P(x_i \mid x_{-i})$
- Distribution converges to $P(x)$

Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \middle| w_{i,j}, \alpha, \beta\right)$ intractable

Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \middle| w_{i,j}, \alpha, \beta\right)$ intractable
- $P\left(z_{i,j} \middle| z_{-i,j}, w_{i,j}, \alpha, \beta\right)$ tractable

Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \,\middle|\, w_{i,j}, \alpha, \beta\right)$ intractable
- $P\left(z_{i,j} \,\middle|\, z_{-i,j}, w_{i,j}, \alpha, \beta\right)$ tractable
  - Dirichlet prior $\Rightarrow$ can marginalise out ϕ, θ

Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \,\middle|\, w_{i,j}, \alpha, \beta\right)$ intractable
- $P\left(z_{i,j} \,\middle|\, z_{-i,j}, w_{i,j}, \alpha, \beta\right)$ tractable
  - Dirichlet prior ⇒ can marginalise out ϕ, θ

$$\begin{aligned} & P\left(z_{i,j} \,\middle|\, z_{-i,j}, w_{i,j}, \alpha, \beta\right) \\ \propto \quad & P\left(z_{i,j} \,\middle|\, \theta_i\right) \; P\left(w_{i,j} \,\middle|\, z_{i,j}, \phi_{z_{i,j}}\right) \end{aligned}$$

Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \middle| w_{i,j}, \alpha, \beta\right)$ intractable
- $P\left(z_{i,j} \middle| z_{-i,j}, w_{i,j}, \alpha, \beta\right)$ tractable
  - Dirichlet prior ⇒ can marginalise out ϕ, θ

$$
\begin{aligned}
& P\left(z_{i,j} \middle| z_{-i,j}, w_{i,j}, \alpha, \beta\right) \\
\propto \;\; & \underbrace{P\left(z_{i,j} \middle| \theta_i\right)}_{\propto C_{i,z}} \;\; \underbrace{P\left(w_{i,j} \middle| z_{i,j}, \phi_{z_{i,j}}\right)}_{\propto C_{z,w}}
\end{aligned}
$$

Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \,\middle|\, w_{i,j}, \alpha, \beta\right)$ intractable
- $P\left(z_{i,j} \,\middle|\, z_{-i,j}, w_{i,j}, \alpha, \beta\right)$ tractable
  - Dirichlet prior ⇒ can marginalise out ϕ, θ

$$
\begin{aligned}
& P\left(z_{i,j} \,\middle|\, z_{-i,j}, w_{i,j}, \alpha, \beta\right) \\
\propto\ & P\left(z_{i,j} \,\middle|\, \theta_i\right) \quad P\left(w_{i,j} \,\middle|\, z_{i,j}, \phi_{z_{i,j}}\right) \\
=\ & \frac{C_{i,z}}{C_i} \qquad\qquad \frac{C_{z,w}}{C_z}
\end{aligned}
$$

Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P\left(\phi_z, \theta_i, z_{i,j} \,\middle|\, w_{i,j}, \alpha, \beta\right)$ intractable
- $P\left(z_{i,j} \,\middle|\, z_{-i,j}, w_{i,j}, \alpha, \beta\right)$ tractable
  - Dirichlet prior ⇒ can marginalise out ϕ, θ

$$
\begin{aligned}
& P\left(z_{i,j} \,\middle|\, z_{-i,j}, w_{i,j}, \alpha, \beta\right) \\
\propto \quad & P\left(z_{i,j} \,\middle|\, \theta_i\right) \quad P\left(w_{i,j} \,\middle|\, z_{i,j}, \phi_{z_{i,j}}\right) \\
= \quad & \frac{C_{i,z} + \alpha}{C_i + K\alpha} \qquad \frac{C_{z,w} + \beta}{C_z + V\beta}
\end{aligned}
$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a a b a b b

c d d d c

b a c b d d

a c

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
1	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2

c	d	d	d	c
2	2	1	1	1

b	a	c	b	d	d
1	2	1	1	1	2

a	c
1	2

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$$P(z_{1,1} = 1) \propto P(1 \mid \theta_1) P(a \mid 1)$$

$$P(z_{1,1} = 2) \propto P(2 \mid \theta_1) P(a \mid 2)$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$$P(z_{1,1} = 1) \propto P(1 \mid \theta_1) P(a \mid 1)$$

$$P(z_{1,1} = 2) \propto P(2 \mid \theta_1) P(a \mid 2)$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$$P(z_{1,1}=1) \propto \frac{2+1}{5+2} P(a|1)$$

$$P(z_{1,1}=2) \propto \frac{3+1}{5+2} P(a|2)$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$$P(z_{1,1} = 1) \propto \frac{2+1}{5+2}\ P(a \mid 1)$$

$$P(z_{1,1} = 2) \propto \frac{3+1}{5+2}\ P(a \mid 2)$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$$P(z_{1,1} = 1) \propto \frac{2+1}{5+2} \quad \frac{2+1}{10+4}$$

$$P(z_{1,1} = 2) \propto \frac{3+1}{5+2} \quad P(a \mid 2)$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$$P(z_{1,1} = 1) \propto \frac{2+1}{5+2} \quad \frac{2+1}{10+4}$$

$$P(z_{1,1} = 2) \propto \frac{3+1}{5+2} \quad P(a \mid 2)$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$$P(z_{1,1} = 1) \propto \frac{2+1}{5+2} \quad \frac{2+1}{10+4}$$

$$P(z_{1,1} = 2) \propto \frac{3+1}{5+2} \quad \frac{2+1}{8+4}$$

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$P(z_{1,1} = 1) \propto$ 0.092

$P(z_{1,1} = 2) \propto$ 0.143

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
?	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$P(z_{1,1} = 1) =$ 0.391

$P(z_{1,1} = 2) =$ 0.609

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
2	2	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

$P(z_{1,1} = 1) =$ 0.391

$P(z_{1,1} = 2) =$ 0.609

Gibbs Sampling for LDA

$$K = 2,\ V = 4,\ \alpha = \beta = 1$$

a	a	b	a	b	b
2	?	2	1	1	2
c	d	d	d	c	
2	2	1	1	1	
b	a	c	b	d	d
1	2	1	1	1	2
a	c				
1	2				

Gibbs Sampling for LDA

- Given a sample:

$$\hat{\theta}_i(z) = \frac{C_{i,z} + \alpha}{C_i + K\alpha} \qquad \hat{\phi}_z(w) = \frac{C_{z,w} + \beta}{C_z + V\beta}$$

Gibbs Sampling for LDA

- Given a sample:

$$\hat{\theta}_i(z) = \frac{C_{i,z} + \alpha}{C_i + K\alpha} \qquad \hat{\phi}_z(w) = \frac{C_{z,w} + \beta}{C_z + V\beta}$$

- Can't directly compare topics from different samples

Gibbs Sampling for LDA

- Given a sample:

$$\hat{\theta}_i(z) = \frac{C_{i,z} + \alpha}{C_i + K\alpha} \qquad \hat{\phi}_z(w) = \frac{C_{z,w} + \beta}{C_z + V\beta}$$

- Can't directly compare topics from different samples

- Can compare e.g. $D_{KL}(\text{doc } 1 || \text{doc } 2)$, as distributions over words

Summary

- Tasks:
  - Word Sense Induction
  - Topic Discovery
- Models:
  - K-Means
  - Latent Dirichlet Allocation
- Training:
  - Gibbs Sampling