Today’s Lecture

- Unsupervised Learning
  - Word Sense Induction
  - Topic Discovery
- K-Means Clustering
- Latent Dirichlet Allocation
- Approximate Inference
Supervised Learning
Unsupervised Learning
Unsupervised Learning
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
... the last kick of the match.
... the Duddon are no match, after all, for a route of any consequence.
... first or second round matches of any consequence.
... Tried soaking the matches in paint, he wrote, for Berowne; this is is very much a match for Berowne; this is.
... to win and the match is therefore needed to watch the ... of an elimination match.
... to lose you the match even though no ... match.
... needed to watch the drop in a burning match.
It was entertaining ... match, needed a diversion ...
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 last kick of the match.
... the Duddon are no match, after all, for a route of any consequence.
... first or second round matches of any consequence
... Tried soaking the matches in paint, he wrote, for Berowne; this is for Berowne; this is
... is very much a match for Berowne; this is
... to win and the match is therefore even though no ...
... 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 ...
... the last kick of the match. It was entertaining after all, for a route...
... 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 ...
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.
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.
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

Bayesian view of smoothing hyperparameters: Dirichlet prior
Recap: Multinomial Naive Bayes

\[ P(w_i | y_i) P(y_i) \]

Bayesian view of smoothing hyperparameters: Dirichlet prior
Recap: Multinomial Naive Bayes

Bayesian view of smoothing hyperparameters: Dirichlet prior
Recap: Multinomial Naive Bayes

Bayesian view of smoothing hyperparameters: Dirichlet prior
Recap: Multinomial Naive Bayes

$$P(w | y) = \prod_{i=1}^{N} P(w_i | y)$$

$$P(y)$$

Bayesian view of smoothing hyperparameters: Dirichlet prior

$$\alpha, \beta$$
Latent Dirichlet Allocation
Latent Dirichlet Allocation

\[ \prod_z P(\phi_z | \beta) \]
Latent Dirichlet Allocation

\[ \prod_z P(\phi_z | \beta) \prod_i P(\theta_i | \alpha) \]
Latent Dirichlet Allocation

\[
\prod_z P(\phi_z | \beta) \prod_i P(\theta_i | \alpha) \prod_j P(z_{i,j} | \theta_i) P(w_{i,j} | z_{i,j})
\]
Latent Dirichlet Allocation

\[ P(\phi_z, \theta_i | w_{i,j}, \alpha, \beta) \]
Latent Dirichlet Allocation

\[ P(\phi_z, \theta_i \mid w_{i,j}, \alpha, \beta) = \sum_{z_{i,j}} P(\phi_z, \theta_i, z_{i,j} \mid w_{i,j}, \alpha, \beta) \]
Latent Dirichlet Allocation

\[ P(\phi_z, \theta_i \mid w_{i,j}, \alpha, \beta) = \sum_{z_{i,j}} P(\phi_z, \theta_i, z_{i,j} \mid w_{i,j}, \alpha, \beta) \]
Approximate Inference

- Want to know global variables (e.g. $\phi$)
- 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

$\mathbb{E}f(x) \approx \frac{1}{N} \sum_{\text{samples}} f(x)$
Markov Chain Monte Carlo

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

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

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

\[ \sum_{z_{i,j}} P(\phi_z, \theta_i, z_{i,j} | w_{i,j}, \alpha, \beta) \] intractable
Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P(\phi_z, \theta_i, z_{i,j} \mid w_{i,j}, \alpha, \beta)$ intractable

- $P(z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta)$ tractable
Gibbs Sampling for LDA

- \[ \sum_{z_{i,j}} P(\phi_z, \theta_i, z_{i,j} \mid w_{i,j}, \alpha, \beta) \text{ intractable} \]
- \[ P(z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta) \text{ tractable} \]
  - Dirichlet prior \( \Rightarrow \) can marginalise out \( \phi, \theta \)
Gibbs Sampling for LDA

\[ \sum_{z_{i,j}} P(\phi_z, \theta_i, z_{i,j} \mid w_{i,j}, \alpha, \beta) \text{ intractable} \]

\[ P(z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta) \text{ tractable} \]

- Dirichlet prior ⇒ can marginalise out \( \phi, \theta \)

\[
P(z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta) \propto P(z_{i,j} \mid \theta_i) \ P(w_{i,j} \mid z_{i,j}, \phi_{z_{i,j}})
\]
Gibbs Sampling for LDA

- \[ \sum_{z_{i,j}} P(\phi_{z}, \theta_{i}, z_{i,j} \mid w_{i,j}, \alpha, \beta) \text{ intractable} \]

- \[ P(z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta) \text{ tractable} \]  
  - Dirichlet prior \( \Rightarrow \) can marginalise out \( \phi, \theta \)

\[
P(z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta) \\
\propto P(z_{i,j} \mid \theta_{i}) \ P(w_{i,j} \mid z_{i,j}, \phi_{z_{i,j}}) \\
\propto C_{i,z} \quad \propto C_{z,w}
\]
Gibbs Sampling for LDA

- $\sum_{z_{i,j}} P(\phi_z, \theta_i, z_{i,j} \mid w_{i,j}, \alpha, \beta)$ intractable

- $P(z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta)$ tractable
  - Dirichlet prior $\Rightarrow$ can marginalise out $\phi, \theta$

\[
P(\mathbf{z}_{i,j} \mid \mathbf{z}_{-i,j}, w_{i,j}, \alpha, \beta) \propto P(\mathbf{z}_{i,j} \mid \theta_i) \cdot P(w_{i,j} \mid \mathbf{z}_{i,j}, \phi_{z_{i,j}})
\]
\[
= \frac{C_{i,z}}{C_i} \cdot \frac{C_{z,w}}{C_z}
\]
Gibbs Sampling for LDA

\[ \sum_{z_{i,j}} P\left( \phi_z, \theta_i, z_{i,j} \mid w_{i,j}, \alpha, \beta \right) \text{ intractable} \]

\[ P\left( z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta \right) \text{ tractable} \]

- Dirichlet prior \( \Rightarrow \) can marginalise out \( \phi, \theta \)

\[
P\left( z_{i,j} \mid z_{-i,j}, w_{i,j}, \alpha, \beta \right) 
\propto P\left( z_{i,j} \mid \theta_i \right) P\left( w_{i,j} \mid z_{i,j}, \phi_{z_{i,j}} \right) 
\]

\[
= \frac{C_{i,z} + \alpha}{C_i + K\alpha} \quad \frac{C_{z,w} + \beta}{C_z + V\beta}
\]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[ \begin{array}{cccccc}
  a & a & b & a & b & b \\
  c & d & d & d & c \\
  b & a & c & b & d & d \\
  a & c \\
\end{array} \]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[ P(\mathbf{z}_1, 1 = 1) \cdot P(1 | \boldsymbol{\theta}_1) \cdot P(a | 1) \cdot P(\mathbf{z}_1, 1 = 2) \cdot P(2 | \boldsymbol{\theta}_1) \cdot P(a | 2) \]

| 1 2 2 1 1 2 |
| 2 2 1 1 1 |
| 1 2 1 1 1 2 |
| 1 2 1 2 |
| 1 2 |
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{array}{cccccc}
  a & a & b & a & b & b \\
  ? & 2 & 2 & 1 & 1 & 2 \\
  c & d & d & d & c & \\
  2 & 2 & 1 & 1 & 1 & \\
  b & a & c & b & d & d \\
  1 & 2 & 1 & 1 & 1 & 2 \\
  a & c \\
  1 & 2
\end{array}
\]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{array}{cccccc}
\text{a} & \text{a} & \text{b} & \text{a} & \text{b} & \text{b} \\
? & 2 & 2 & 1 & 1 & 2 \\
\text{c} & \text{d} & \text{d} & \text{d} & \text{c} & \\
2 & 2 & 1 & 1 & 1 & \\
\text{b} & \text{a} & \text{c} & \text{b} & \text{d} & \text{d} \\
1 & 2 & 1 & 1 & 1 & 2 \\
\text{a} & \text{c} & \\
1 & 2 \\
\end{array}
\]

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

\[ P(z_{1,1} = 2) \propto P(2 \mid \theta_1)P(a \mid 2) \]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\begin{align*}
\begin{array}{cccccc}
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 & \\
\end{array}
\end{align*}

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

\[ P(z_{1,1} = 2) \propto P(2 \mid \theta_1)P(a \mid 2) \]
Gibbs Sampling for LDA

\( K = 2, \ V = 4, \ \alpha = \beta = 1 \)

\[
\begin{array}{cccccc}
\text{a} & \text{a} & \text{b} & \text{a} & \text{b} & \text{b} \\
? & 2 & 2 & 1 & 1 & 2 \\
\text{c} & \text{d} & \text{d} & \text{d} & \text{c} & \\
2 & 2 & 1 & 1 & 1 & \\
\text{b} & \text{a} & \text{c} & \text{b} & \text{d} & \text{d} \\
1 & 2 & 1 & 1 & 1 & 2 \\
\text{a} & \text{c} & \\
1 & 2
\end{array}
\]

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

\[
P(z_{1,1} = 2) \propto \frac{3 + 1}{5 + 2} \quad P(a \mid 2)
\]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{array}{cccccc}
\text{a} & \text{a} & \text{b} & \text{a} & \text{b} & \text{b} \\
? & 2 & 2 & 1 & 1 & 2 \\
\text{c} & \text{d} & \text{d} & \text{d} & \text{c} & \\
2 & 2 & 1 & 1 & 1 & \\
\text{b} & \text{a} & \text{c} & \text{b} & \text{d} & \text{d} \\
1 & 2 & 1 & 1 & 1 & 2 \\
\text{a} & \text{c} & \\
1 & 2
\end{array}
\]

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

\[
P(z_{1,1} = 2) \propto \frac{3 + 1}{5 + 2} \quad P(a | 2)
\]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1\]

\[
\begin{array}{cccccc}
  a & a & b & a & b & b \\
  ? & 2 & 2 & 1 & 1 & 2 \\
  c & d & d & d & c \\
  2 & 2 & 1 & 1 & 1 \\
  b & a & c & b & d & d \\
  1 & 2 & 1 & 1 & 1 & 2 \\
  a & c \\
  1 & 2 \\
\end{array}
\]

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

\[
P(z_{1,1} = 2) \propto \frac{3 + 1}{5 + 2}
\]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{array}{cccccc}
  a & a & b & a & b & b \\
  c & d & d & d & c & ? \\
  2 & 2 & 1 & 1 & 2 & 1 \\
  b & a & c & b & d & d \\
  1 & 2 & 1 & 1 & 1 & 2 \\
  a & c & ? & 2 & 2 & 1 \\
  1 & 2 & 1 & 1 & 2 & 1 \\
\end{array}
\]

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

\[
P(z_{1,1} = 2) \propto \frac{3 + 1}{5 + 2} \quad \frac{P(a \mid 2)}{}
\]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[ \begin{align*}
    &a \ a \ b \ a \ b \ b \\
    &? \ 2 \ 2 \ 1 \ 1 \ 2 \\
    &c \ d \ d \ d \ c \\
    &2 \ 2 \ 1 \ 1 \ 1 \\
    &b \ a \ c \ b \ d \ d \\
    &1 \ 2 \ 1 \ 1 \ 1 \ 2 \\
    &a \ c \\
    &1 \ 2
\end{align*} \]

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

\[ P(z_{1,1} = 2) \propto \frac{3 + 1}{5 + 2} \frac{2 + 1}{8 + 4} \]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{array}{cccccc}
a & a & b & a & b & b \\
? & 2 & 2 & 1 & 1 & 2 \\
c & d & d & d & c & \\
2 & 2 & 1 & 1 & 1 & \\
b & a & c & b & d & d \\
1 & 2 & 1 & 1 & 1 & 2 \\
a & c \\
1 & 2
\end{array}
\]

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

\[ P(z_{1,1} = 2) \propto 0.143 \]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{align*}
\begin{array}{cccccc}
a & a & b & a & b & b \\
? & 2 & 2 & 1 & 1 & 2 \\
c & d & d & d & c \\
2 & 2 & 1 & 1 & 1 \\
b & a & c & b & d & d \\
1 & 2 & 1 & 1 & 1 & 2 \\
a & c \\
1 & 2 \\
\end{array}
\end{align*}
\]

\[ P(z_{1,1}=1) = 0.391 \]

\[ P(z_{1,1}=2) = 0.609 \]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{array}{cccc}
a & a & b & a & b & b \\
2 & 2 & 2 & 1 & 1 & 2 \\
c & d & d & d & c \\
2 & 2 & 1 & 1 & 1 \\
b & a & c & b & d & d \\
1 & 2 & 1 & 1 & 1 & 2 \\
a & c \\
1 & 2 \\
\end{array}
\]

\[ P(z_{1,1} = 1) = 0.391 \]

\[ P(z_{1,1} = 2) = 0.609 \]
Gibbs Sampling for LDA

\[ K = 2, \ V = 4, \ \alpha = \beta = 1 \]

\[
\begin{array}{cccccc}
a & a & b & a & b & b \\
2 & ? & 2 & 1 & 1 & 2 \\
c & d & d & d & c \\
2 & 2 & 1 & 1 & 1 \\
\end{array}
\]

\[
\begin{array}{cccc}
b & a & c & b & d & d \\
1 & 2 & 1 & 1 & 1 & 2 \\
a & c \\
1 & 2 \\
\end{array}
\]
Gibbs Sampling for LDA

- Given a sample:

\[
\hat{\theta}_i(z) = \frac{C_{i,z} + \alpha}{C_i + K\alpha} \quad \hat{\phi}_z(w) = \frac{C_{z,w} + \beta}{C_z + V\beta}
\]
Gibbs Sampling for LDA

- Given a sample:

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

- Can’t directly compare topics from different samples
Gibbs Sampling for LDA

- Given a sample:

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

- 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