Clustering and the $k$-means Algorithm

David M. Blei

COS424
Princeton University

March 11, 2007
Clustering

• Goal: Automatically segment data into groups of similar points
Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
Clustering

- **Goal:** Automatically segment data into groups of similar points
- **Question:** When and why would we want to do this?
- **Useful for:**
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
  - Examples:
    - Customers according to purchase histories
    - Genes according to expression profile
    - Search results according to topic
    - MySpace users according to interests
    - A museum catalog according to image similarity
Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data

Examples:
- Customers according to purchase histories
- Genes according to expression profile
- Search results according to topic
- MySpace users according to interests
- A museum catalog according to image similarity
Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space

Examples:
- Customers according to purchase histories
- Genes according to expression profile
- Search results according to topic
- MySpace users according to interests
- A museum catalog according to image similarity
Clustering

• Goal: Automatically segment data into groups of similar points
• Question: When and why would we want to do this?
• Useful for:
  • Automatically organizing data
  • Understanding hidden structure in some data
  • Representing high-dimensional data in a low-dimensional space
• Examples:
Clustering

• Goal: Automatically segment data into groups of similar points
• Question: When and why would we want to do this?
• Useful for:
  • Automatically organizing data
  • Understanding hidden structure in some data
  • Representing high-dimensional data in a low-dimensional space
• Examples:
  • Customers according to purchase histories
Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories
  - Genes according to expression profile
Clustering

• Goal: Automatically segment data into groups of similar points

• Question: When and why would we want to do this?

• Useful for:
  • Automatically organizing data
  • Understanding hidden structure in some data
  • Representing high-dimensional data in a low-dimensional space

• Examples:
  • Customers according to purchase histories
  • Genes according to expression profile
  • Search results according to topic
Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories
  - Genes according to expression profile
  - Search results according to topic
  - MySpace users according to interests
Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories
  - Genes according to expression profile
  - Search results according to topic
  - MySpace users according to interests
  - A museum catalog according to image similarity
Clustering set-up

• Our data are

\[ \mathcal{D} = \{x_1, \ldots, x_N\}. \]
Clustering set-up

- Our data are
  \[ \mathcal{D} = \{ \mathbf{x}_1, \ldots, \mathbf{x}_N \} \].
- Each data point is \( p \)-dimensional, i.e.,
  \[ \mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle \].
Clustering set-up

- Our data are
  \[ \mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}. \]

- Each data point is \( p \)-dimensional, i.e.,
  \[ \mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle. \]

- Define a *distance function* between data, \( d(\mathbf{x}_n, \mathbf{x}_m) \).
Clustering set-up

- Our data are \( \mathcal{D} = \{x_1, \ldots, x_N\} \).
- Each data point is \( p \)-dimensional, i.e.,
  \[ x_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle. \]
- Define a distance function between data, \( d(x_n, x_m) \).
- Goal: segment the data into \( k \) groups
  \[ \{z_1, \ldots, z_N\} \text{ where } z_i \in \{1, \ldots, K\}. \]
Example data

500 2-dimensional data points: \( x_n = \langle x_{n,1}, x_{n,2} \rangle \)
What is a good distance function here?
What is a good distance function here?

Squared Euclidean distance is reasonable

\[
d(x_n, x_m) = \sum_{i=1}^{p} (x_{n,i} - x_{m,i})^2 = ||x_n - x_m||^2
\]
Example data

- Goal: segment this data into $k$ groups.
• Goal: segment this data into $k$ groups.

• What should $k$ be?
Goal: segment this data into $k$ groups.

What should $k$ be?

Automatically choosing $k$ is complicated; for now, 4.
Different clustering algorithms use the data and distance measurements in different ways.
• Different clustering algorithms use the data and distance measurements in different ways

• Begin with \textit{k}-means, the simplest clustering algorithm
• Different clustering algorithms use the data and distance measurements in different ways
• Begin with $k$-means, the simplest clustering algorithm
The basic idea is to describe each cluster by its mean value.
The basic idea is to describe each cluster by its mean value. (Note: this works only for distances such that a mean is well-defined.)
The basic idea is to describe each cluster by its mean value.

(Note: this works only for distances such that a mean is well-defined.)

The goal of $k$-means is to assign data to clusters and define these clusters with their means.
$k$-means algorithm

1. Initialization

- Data are $x_1: N$.
- Choose initial cluster means $m_1: k$ (same dimension as data).

2. Repeat

- Assign each data point to its closest mean $z_n = \arg\min_{i \in \{1, \ldots, k\}} d(x_n, m_i)$.
- Compute each cluster mean to be the coordinate-wise average over data points assigned to that cluster, $m_k = \frac{1}{N_k} \sum_{n: z_n = k} x_n$.

3. Until assignments $z_1: N$ do not change.
1. **Initialization**
   - Data are $x_{1:N}$
**k-means algorithm**

1. **Initialization**
   - Data are $x_{1:N}$
   - Choose initial cluster means $m_{1:k}$ (same dimension as data).
$k$-means algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).

2. Repeat

   Assign each data point to its closest mean
   \[ z_n = \arg \min_{i \in \{1, \ldots, k\}} d(\mathbf{x}_n, \mathbf{m}_i) \]

   Compute each cluster mean to be the coordinate-wise average over data points assigned to that cluster,
   \[ \mathbf{m}_k = \frac{1}{N_k} \sum_{n: z_n = k} \mathbf{x}_n \]

   Until assignments $z_1: N$ do not change.
$k$-means algorithm

1. Initialization
   - Data are $x_{1:N}$
   - Choose initial cluster means $m_{1:k}$ (same dimension as data).

2. Repeat
   1. Assign each data point to its closest mean
      
      \[
      z_n = \arg \min_{i \in \{1, \ldots, k\}} d(x_n, m_i)
      \]
\textbf{k-means algorithm}

1. \textbf{Initialization}
   - Data are $x_{1:N}$
   - Choose initial cluster means $m_{1:k}$ (same dimension as data).

2. \textbf{Repeat}
   1. Assign each data point to its closest mean
      \[ z_n = \arg \min_{i \in \{1, \ldots, k\}} d(x_n, m_i) \]
   2. Compute each cluster mean to be the coordinate-wise average over data points assigned to that cluster,
      \[ m_k = \frac{1}{N_k} \sum_{\{n : z_n = k\}} x_n \]
**k-means algorithm**

1. **Initialization**
   - Data are $x_{1:N}$
   - Choose initial cluster means $m_{1:k}$ (same dimension as data).

2. **Repeat**
   1. Assign each data point to its closest mean
      \[
      z_n = \arg \min_{i \in \{1, \ldots, k\}} d(x_n, m_i)
      \]
   2. Compute each cluster mean to be the coordinate-wise average over data points assigned to that cluster,
      \[
      m_k = \frac{1}{N_k} \sum_{\{n : z_n = k\}} x_n
      \]

3. **Until assignments $z_{1:N}$ do not change**
$k$-means example
k-means example
$k$-means example
$k$-means example
$k$-means example
$k$-means example
$k$-means example
Objective function

• How can we measure how well our algorithm is doing?

The $k$-means objective function is the sum of the squared distances of each point to each assigned mean:

$$F(z_1:N, m_1:k) = \frac{1}{2} N \sum_{n=1}^{N} ||x_n - m_{z_n}||^2$$
Objective function

- How can we measure how well our algorithm is doing?
- The $k$-means objective function is the sum of the squared distances of each point to each assigned mean

$$F(z_{1:N}, m_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||x_n - m_{z_n}||^2$$
$k$-means example (look at the objective)
$k$-means example (look at the objective)
$k$-means example (look at the objective)
k-means example (look at the objective)
$k$-means example (look at the objective)
k-means example (look at the objective)
$k$-means example (look at the objective)
Coordinate descent

\[ F(z_{1:N}, m_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||x_n - m_{z_n}||^2 \]

- Holding the means fixed, assigning each point to its closest mean minimizes \( F \) with respect to \( z_{1:N} \).
Coordinate descent

\[ F(z_{1:N}, m_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||x_n - m_{z_n}||^2 \]

- Holding the means fixed, assigning each point to its closest mean minimizes \( F \) with respect to \( z_{1:N} \).
- Holding the assignments fixed, computing the centroids of each cluster minimizes \( F \) with respect to \( m_{1:k} \).
Coordinate descent

\[ F(z_{1:N}, m_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||x_n - m_{z_n}||^2 \]

- Holding the means fixed, assigning each point to its closest mean minimizes \( F \) with respect to \( z_{1:N} \).
- Holding the assignments fixed, computing the centroids of each cluster minimizes \( F \) with respect to \( m_{1:k} \).
- Thus, \( k \)-means is a coordinate descent algorithm.
Coordinate descent

\[ F(z_{1:N}, m_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||x_n - m_{z_n}||^2 \]

- Holding the means fixed, assigning each point to its closest mean minimizes \( F \) with respect to \( z_{1:N} \).
- Holding the assignments fixed, computing the centroids of each cluster minimizes \( F \) with respect to \( m_{1:k} \).
- Thus, \( k \)-means is a coordinate descent algorithm.
- It finds a local minimum. (Multiple restarts are often necessary.)
Objective for the example data

![Graph showing the objective for the example data over rounds of k-means clustering.](image-url)
Compressing images

- Each pixel is associated with a red, green, and blue value.
Compressing images

- Each pixel is associated with a red, green, and blue value
- A $1024 \times 1024$ image is a collection of 1048576 values $\langle x_1, x_2, x_3 \rangle$, which requires 3M of storage
Compressing images

• Each pixel is associated with a red, green, and blue value
• A $1024 \times 1024$ image is a collection of 1048576 values $\langle x_1, x_2, x_3 \rangle$, which requires 3M of storage
• How can we use $k$-means to compress this image?
Vector quantization

- Replace each pixel $x_n$ with its assignment $m_{zn}$ (“paint by numbers”).

- The $k$ means are called the codebook.

- With $k = 100$, we need 7 bits per pixel plus $100 \times 3$ bits $\approx 897K$. 

D. Blei Clustering 01
Vector quantization

- Replace each pixel $x_n$ with its assignment $m_{zn}$ (“paint by numbers”).
- The $k$ means are called the codebook.
Vector quantization

• Replace each pixel \( \mathbf{x}_n \) with its assignment \( \mathbf{m}_{z_n} \) (“paint by numbers”).
• The \( k \) means are called the codebook.
• With \( k = 100 \), we need 7 bits per pixel plus \( 100 \times 3 \) bits \( \approx 897 \text{K} \).
Charlie Brown and Linus VQ

2 means
4 means
Charlie Brown and Linus VQ

8 means
Charlie Brown and Linus VQ

16 means
Charlie Brown and Linus VQ

32 means
Charlie Brown and Linus VQ

64 means
Charlie Brown and Linus VQ

128 means
Charlie Brown and Linus VQ

256 means
The objective gives a measure of how distorted the compressed picture is relative to the original picture.
The objective gives a measure of how distorted the compressed picture is relative to the original picture.

For more clusters, the picture is less distorted.
In many practical settings, Euclidean distance is not appropriate. When?

- Euclidean distance requires the definition of a mean, which may not be meaningful in discrete or positive data.
- $k$-medoids only requires knowing distances between data points, $d_{nm} = d(x_n, x_m)$.
- No need to define the mean.
- Each cluster is associated with its most typical example.
In many practical settings, Euclidean distance is not appropriate. When?

For example,
**k-medoids**

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
  - Discrete multivariate data, such as purchase histories
$k$-medoids

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
  - Discrete multivariate data, such as purchase histories
  - Positive data, such as time spent on a web-page


- **In many practical settings, Euclidean distance is not appropriate. When?**
- **For example,**
  - Discrete multivariate data, such as purchase histories
  - Positive data, such as time spent on a web-page
- **$k$-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.**
In many practical settings, Euclidean distance is not appropriate. When?

For example,
- Discrete multivariate data, such as purchase histories
- Positive data, such as time spent on a web-page

$k$-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.

No need to define the mean.
In many practical settings, Euclidean distance is not appropriate. When?

For example,
- Discrete multivariate data, such as purchase histories
- Positive data, such as time spent on a web-page

$k$-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.

No need to define the mean.

Each of the clusters is associated with its most typical example
$k$-medoids algorithm

1 Initialization
**k-medoids algorithm**

1. **Initialization**
   - Data are $x_{1:N}$
**k-medoids algorithm**

1. **Initialization**
   - Data are $x_{1:N}$
   - Choose initial cluster identities $m_{1:k}$
**k-medoids algorithm**

1. **Initialization**
   - Data are \( x_{1:N} \)
   - Choose initial cluster identities \( m_{1:k} \)

2. **Repeat**
**k-medioids algorithm**

1. **Initialization**
   - Data are \( x_{1:N} \)
   - Choose initial cluster identities \( m_{1:k} \)

2. **Repeat**
   1. **Assign each data point to its closest center**

   \[
z_n = \arg \min_{i \in \{1, \ldots, k\}} d(x_n, m_i)
   \]
**k-medoids algorithm**

1. **Initialization**
   - Data are \( x_{1:N} \)
   - Choose initial cluster identities \( m_{1:k} \)

2. **Repeat**
   1. Assign each data point to its closest center
      
      \[
      z_n = \arg \min_{i \in \{1, \ldots, k\}} \: d(x_n, m_i)
      \]
   2. For each cluster, find the data point in that cluster that is closest to the other points in that cluster
      
      \[
      i_k = \arg \min_{n \atop z_n = k} \sum_{m \atop z_m = k} d(x_n, x_m)
      \]
**k-medoids algorithm**

1. **Initialization**
   - Data are $x_{1:N}$
   - Choose initial cluster identities $m_{1:k}$

2. **Repeat**
   1. Assign each data point to its closest center
      \[
      z_n = \arg \min_{i \in \{1, \ldots, k\}} d(x_n, m_i)
      \]
   2. For each cluster, find the data point in that cluster that is closest to the other points in that cluster
      \[
      i_k = \arg \min_{\{n: z_n = k\}} \sum_{\{m: z_m = k\}} d(x_n, x_m)
      \]
   3. Set each cluster center equal to their closest data points
      \[
      m_k = x_{i_k}
      \]
**k-medoids algorithm**

1. **Initialization**
   - Data are $x_{1:N}$
   - Choose initial cluster identities $m_{1:k}$

2. **Repeat**
   1. Assign each data point to its closest center
      
      $$z_n = \arg \min_{i \in \{1,...,k\}} d(x_n, m_i)$$

   2. For each cluster, find the data point in that cluster that is closest to the other points in that cluster
      
      $$i_k = \arg \min \{ n : z_n=k \} \sum \{ m : z_m=k \} d(x_n, x_m)$$

   3. Set each cluster center equal to their closest data points
      
      $$m_k = x_{i_k}$$

3. **Until** assignments $z_{1:N}$ do not change
Choosing $k$

- Choosing $k$ is a nagging problem in cluster analysis
Choosing $k$

- Choosing $k$ is a nagging problem in cluster analysis
- Sometimes, the problem determines $k$
Choosing $k$

- Choosing $k$ is a nagging problem in cluster analysis
- Sometimes, the problem determines $k$
  - A certain required compression in VQ
Choosing $k$

- Choosing $k$ is a nagging problem in cluster analysis
- Sometimes, the problem determines $k$
  - A certain required compression in VQ
  - Clustering customers for $k$ salespeople in a business
Choosing $k$

- Choosing $k$ is a nagging problem in cluster analysis
- Sometimes, the problem determines $k$
  - A certain required compression in VQ
  - Clustering customers for $k$ salespeople in a business
- Usually, we seek the “natural” clustering, but what does this mean?
Choosing \( k \)

- Choosing \( k \) is a nagging problem in cluster analysis
- Sometimes, the problem determines \( k \)
  - A certain required compression in VQ
  - Clustering customers for \( k \) salespeople in a business
- Usually, we seek the “natural” clustering, but what does this mean?
- It is not well-defined.
What happens as $k$ increases?
What happens as $k$ increases?
What happens as $k$ increases?
What happens as $k$ increases?
What happens as $k$ increases?
What happens as $k$ increases?
What happens as $k$ increases?
What happens as $k$ increases?
Heuristic: A kink in the objective

- Notice the “kink” in the objective between 3 and 5.
Heuristic: A kink in the objective

- Notice the “kink” in the objective between 3 and 5.
- This suggests that 4 is the right number of clusters.
Heuristic: A kink in the objective

- Notice the “kink” in the objective between 3 and 5.
- This suggests that 4 is the right number of clusters.
- Tibshirani (2001) presents a method for finding this kink.
Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
Archeology

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel
Archeology

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel
- Make inferences about political history based on the clusters
Archeology

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel
- Make inferences about political history based on the clusters
- Choose $k$ very carefully, with a complicated computational technique.
• Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate (Gracey et al., 2004)
• Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate (Gracey et al., 2004)
• Exposed carp to different levels of cold
• Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate (Gracey et al., 2004)
• Exposed carp to different levels of cold
• Clustered genes based on their response in different tissues
• Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate (Gracey et al., 2004)
• Exposed carp to different levels of cold
• Clustered genes based on their response in different tissues
• (No mention of how $k = 23$ was chosen.)
Education

- Teachers as Sources of Middle School Students’ Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
• Teachers as Sources of Middle School Students’ Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)

• Clustered survey results of 206 students
• Teachers as Sources of Middle School Students’ Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)

• Clustered survey results of 206 students

• Used the clusters to identify groups to buttress an analysis of what affects motivation.
• Teachers as Sources of Middle School Students’ Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)

• Clustered survey results of 206 students

• Used the clusters to identify groups to buttress an analysis of what affects motivation.

• I.e., the levels of encouragement are corrected for
• Teachers as Sources of Middle School Students’ Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)

• Clustered survey results of 206 students

• Used the clusters to identify groups to buttress an analysis of what affects motivation.

• I.e., the levels of encouragement are corrected for

• Chose the number of clusters to get nice results
### Table 3. Five-Cluster Solution: Z scores on Each Clustering Variable

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher caring</td>
<td>-.5</td>
<td>-.5 to .5</td>
<td>-.5 to .5</td>
<td>-.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Peers’ academic support</td>
<td>1.0</td>
<td>-.5</td>
<td>1.0</td>
<td>-.5</td>
<td>-.5 to .5</td>
</tr>
<tr>
<td>Parents’ academic support</td>
<td>.5</td>
<td>-1.0</td>
<td>-.5 to .5</td>
<td>-.5 to .5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Table 4. Means and Standard Deviations for Each Cluster on Grade 8 Motivational Variables

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Academic Self-Efficacy</th>
<th>Intrinsic Valuing of Education</th>
<th>Teacher-Rated Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>1. All positive</td>
<td>3.59</td>
<td>.48a</td>
<td>2.99</td>
</tr>
<tr>
<td>2. Peer negative, parents very negative</td>
<td>2.44</td>
<td>.66b</td>
<td>2.16</td>
</tr>
<tr>
<td>3. Peer positive</td>
<td>3.01</td>
<td>.73c</td>
<td>2.43</td>
</tr>
<tr>
<td>4. Negative teacher and peer</td>
<td>2.47</td>
<td>.63b</td>
<td>2.24</td>
</tr>
<tr>
<td>5. Positive teacher and parents</td>
<td>3.19</td>
<td>.65c</td>
<td>2.89</td>
</tr>
</tbody>
</table>
• Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
• Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)

• Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity
• Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)

• Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity

• $K$ chosen for interpretability and “stability,” which means that they could interpret multiple $k$-means runs on different data in the same way.
• Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)

• Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity

• $K$ chosen for interpretability and “stability,” which means that they could interpret multiple $k$-means runs on different data in the same way.

• Draw the conclusion that patterns exist. What’s wrong with this?
• Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)

• Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity

• $K$ chosen for interpretability and “stability,” which means that they could interpret multiple $k$-means runs on different data in the same way.

• Draw the conclusion that patterns exist. What’s wrong with this?

• $k$-means will find patterns everywhere!
<table>
<thead>
<tr>
<th>Cluster type and behavioral patterns</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Light substance dabblers</strong>—infrequent or no current use of substances†</td>
<td>24.4</td>
</tr>
<tr>
<td>None have had sex</td>
<td></td>
</tr>
<tr>
<td><strong>Abstainers</strong>—none have ever used substances† or had sex</td>
<td>22.7</td>
</tr>
<tr>
<td><strong>Sex dabblers</strong>—all have had sex</td>
<td>14.5</td>
</tr>
<tr>
<td>Median no. of partners=1</td>
<td></td>
</tr>
<tr>
<td>60% used a condom at last sex</td>
<td></td>
</tr>
<tr>
<td>Infrequent use of substances†</td>
<td></td>
</tr>
<tr>
<td><strong>Drinkers</strong>—all consumed alcohol in past 12 mos.</td>
<td>7.4</td>
</tr>
<tr>
<td>49% report binge drinking</td>
<td></td>
</tr>
<tr>
<td>Infrequent or no illicit drug use</td>
<td></td>
</tr>
<tr>
<td>None have had sex</td>
<td></td>
</tr>
<tr>
<td><strong>Smokers</strong>—all smoke cigarettes daily</td>
<td>7.3</td>
</tr>
<tr>
<td>Infrequent use of alcohol/illicit drugs</td>
<td></td>
</tr>
<tr>
<td>62% have had sex</td>
<td></td>
</tr>
<tr>
<td><strong>Alcohol-and-sex dabblers</strong>—all drink occasionally; all have had sex</td>
<td>5.4</td>
</tr>
<tr>
<td>Infrequent tobacco/illicit drug use</td>
<td></td>
</tr>
<tr>
<td><strong>Binge drinkers</strong>—all binge frequently</td>
<td>4.4</td>
</tr>
<tr>
<td>Infrequent cigarette, marijuana and other drug use</td>
<td></td>
</tr>
<tr>
<td>60% binge ≥1 time/wk</td>
<td></td>
</tr>
<tr>
<td>45% have had sex</td>
<td></td>
</tr>
<tr>
<td><strong>Heavy dabblers</strong>—all smoke, drink and binge drink with moderate frequency</td>
<td>3.6</td>
</tr>
<tr>
<td>45% use marijuana; few use other illicit drugs</td>
<td></td>
</tr>
<tr>
<td>91% have had sex</td>
<td></td>
</tr>
<tr>
<td><strong>Combination sex and drug use</strong>—all have had sex; all used alcohol/illicit drug at last sex</td>
<td>3.4</td>
</tr>
<tr>
<td><strong>Marijuana users</strong>—all use marijuana frequently; few have used other illicit drugs</td>
<td>1.7</td>
</tr>
<tr>
<td>94% use alcohol</td>
<td></td>
</tr>
<tr>
<td>79% smoke cigarettes</td>
<td></td>
</tr>
<tr>
<td>74% have had sex</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple partners</strong>—all report ≥14 sexual partners</td>
<td>1.3</td>
</tr>
<tr>
<td>75% report low or moderate use of substances†</td>
<td></td>
</tr>
<tr>
<td><strong>Sex for drugs or money</strong>—all have had sex for drugs or money</td>
<td>1.2</td>
</tr>
<tr>
<td>50% report low or moderate use of substances†</td>
<td></td>
</tr>
<tr>
<td>Median no. of partners=3</td>
<td></td>
</tr>
<tr>
<td><strong>High marijuana use and sex</strong>—all use marijuana frequently; all have had sex</td>
<td>1.1</td>
</tr>
<tr>
<td>All used alcohol/other drug at last sex</td>
<td></td>
</tr>
<tr>
<td>82% have had &gt;1 partner (median=6)</td>
<td></td>
</tr>
<tr>
<td><strong>Marijuana and other drug users</strong>—95% report heavy marijuana use; all use other illicit drugs</td>
<td>0.6</td>
</tr>
<tr>
<td>68% have had sex</td>
<td></td>
</tr>
<tr>
<td>28% used alcohol/other drug at last sex</td>
<td></td>
</tr>
<tr>
<td><strong>Injection-drug users</strong>—all have injected drugs</td>
<td>0.6</td>
</tr>
<tr>
<td>82% have had sex</td>
<td></td>
</tr>
<tr>
<td>Median no. of partners=4</td>
<td></td>
</tr>
<tr>
<td><strong>Males who have sex with males</strong>—all are males who have had sex with another male</td>
<td>0.3</td>
</tr>
<tr>
<td>78% have had multiple partners (median=5)</td>
<td></td>
</tr>
<tr>
<td>40% used marijuana in past 30 days</td>
<td></td>
</tr>
<tr>
<td>50% use alcohol ≥1 time/mo.</td>
<td></td>
</tr>
<tr>
<td>17% have had sex for drugs or money</td>
<td></td>
</tr>
</tbody>
</table>