Lecture 21:
Clustering
K-Means
Last time… Boosting

- **Idea:** given a weak learner, run it multiple times on (rewighted) training data, then let the learned classifiers vote

- **On each iteration** $t$:
  - weight each training example by how incorrectly it was classified
  - Learn a hypothesis – $h_t$
  - A strength for this hypothesis – $a_t$

- **Final classifier:**
  - A linear combination of the votes of the different classifiers weighted by their strength $H(X) = \text{sign} \left( \sum \alpha_t h_t(X) \right)$

- **Practically useful**
- **Theoretically interesting**
Last time.. The AdaBoost Algorithm

0) Set $\tilde{W}_i^{(0)} = 1/n$ for $i = 1, \ldots, n$

1) At the $m^{th}$ iteration we find (any) classifier $h(x; \hat{\theta}_m)$ for which the weighted classification error $\epsilon_m$

$$
\epsilon_m = 0.5 - \frac{1}{2} \left( \sum_{i=1}^{n} \tilde{W}_i^{(m-1)} y_i h(x_i; \hat{\theta}_m) \right)
$$

is better than chance.

2) The new component is assigned votes based on its error:

$$
\hat{\alpha}_m = 0.5 \log( (1 - \epsilon_m)/\epsilon_m )
$$

3) The weights are updated according to ($Z_m$ is chosen so that the new weights $\tilde{W}_i^{(m)}$ sum to one):

$$
\tilde{W}_i^{(m)} = \frac{1}{Z_m} \cdot \tilde{W}_i^{(m-1)} \cdot \exp \{ -y_i \hat{\alpha}_m h(x_i; \hat{\theta}_m) \} 
$$
Today

- What is clustering?
- K-means algorithm
What is clustering
Clustering

• Grouping data according to similarity
Clustering

- Grouping data according to similarity
Clustering

- Grouping **data** according to similarity
Clustering

• Grouping **data** according to similarity

  e.g. archaeological dig
Clustering

- Grouping data according to similarity
Clustering

- Grouping data according to similarity

E.g. archaeological dig

Artifac location

Distance North

Distance East

slide by Tamara Broderick
Clustering

- Grouping data according to similarity

*Example: archaeological dig*
Clustering

- Grouping data according to similarity

E.g. archaeological dig
Clustering

• Grouping data according to similarity

E.g. archaeological dig
Clustering

- **Grouping data according to similarity**

E.g. archaeological dig
Clustering

- Grouping data according to similarity

E.g. archaeological dig
Clustering vs. Classification

- Grouping data according to similarity

  Predicting new labels from old labels
Clustering vs. Classification

- **Grouping** data according to similarity

  Predicting new labels from old labels
Clustering vs. Classification

- Grouping data according to similarity

Predicting new labels from old labels
Why use clustering... 
...instead of classification

- Exploratory data analysis
Why use clustering... ...instead of classification

- Exploratory data analysis
Why use clustering...  
...instead of classification

- Exploratory data analysis

Datum: person

Similarity: the number of common interests of two people
Why use clustering…

…instead of classification

• Exploratory data analysis

Datum: a binary vector specifying whether a person has each interest

Similarity: the number of common interests of two people
Why use clustering... 
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)
Why use clustering... 
...instead of classification

• Exploratory data analysis
• Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)
Why use clustering... 
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

<table>
<thead>
<tr>
<th>NEW</th>
<th>MILLION</th>
<th>CHILDREN</th>
<th>SCHOOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

**Topic Analysis**

[Blei 2003]
Why use clustering... 
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

![Topic Analysis](slide by Tamara Broderick)

---

**“Arts”** | **“Budgets”** | **“Children”** | **“Education”**
--- | --- | --- | ---
NEW | MILLION | CHILDREN | SCHOOL
FILM | TAX | WOMEN | STUDENTS
SHOW | PROGRAM | PEOPLE | SCHOOLS
MUSIC | BUDGET | CHILD | EDUCATION
MOVIE | BILLION | YEARS | TEACHERS
PLAY | FEDERAL | FAMILIES | HIGH
MUSICAL | YEAR | WORK | PUBLIC
BEST | SPENDING | PARENTS | TEACHER
ACTOR | NEW | SAYS | BENNETT
FIRST | STATE | FAMILY | MANIGAT
YORK | PLAN | WELFARE | NAMPHY
OPERA | MONEY | MEN | STATE
THEATER | PROGRAMS | PERCENT | PRESIDENT
ACTRESS | GOVERNMENT | CARE | ELEMENTARY
LOVE | CONGRESS | LIFE | HAITI

---

[Blei 2003]

---

Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Philharmonic and Juilliard School. “Our board felt that we had a mark on the future of the performing arts with these grants an act our traditional areas of support in health, medical research, education. Hearst Foundation President Randolph A. Hearst said Monday in Lincoln Center’s share will be $200,000 for its new building, which and provide new public facilities. The Metropolitan Opera Co. and will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Why use clustering... 
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

Datum: word

Similarity: how many documents exist where two words co-occur

[Blei 2003]
Why use clustering...  
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

---

**Topic Analysis**

**Datum:** binary vector indicating document occurrence

**Similarity:** how many documents exist where two words co-occur
Why use clustering...  
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)
Why use clustering...  
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)
Why use clustering...  
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

*Document clustering*

**Datum:** document

**Dissimilarity:** distance between topic distributions of two documents

---

[Carpineto et al. 2009]
Why use clustering...  
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

---

**Document clustering**

**Datum:** vector of topic occurrences

**Dissimilarity:** distance between topic distributions of two documents
Why use clustering... 
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)
Why use clustering... 
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Image segmentation

[Fei-Fei 2011]
Why use clustering... instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

**Image segmentation**

**Datum:** pixel

**Dissimilarity:** difference in color + difference in location

[Fei-Fei 2011]
Why use clustering... 
...instead of classification

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Image segmentation

**Datum:** pixel RGB values and pixel horizontal and vertical locations

**Dissimilarity:** difference in color + difference in location
Clustering algorithms

- **Partitioning algorithms**
  - Construct various partitions and then evaluate them by some criterion
    - K-means
    - Mixture of Gaussians
    - Spectral Clustering

- **Hierarchical algorithms**
  - Create a hierarchical decomposition of the set of objects using some criterion
    - Bottom-up – agglomerative
    - Top-down – divisive
Desirable Properties of a Clustering Algorithm

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Ability to deal with noisy data
- Interpretability and usability

- Optional
  - Incorporation of user-specified constraints
K-Means Clustering
K-Means Clustering

Benefits

• Fast
• Conceptually straightforward
• Popular

 KD-trees, triangle inequality, online version

• Only finds a local optimum

[Ramasubramanian, Paliwal 1990; Moore 2000; Kanungo et al 2002]
K-Means: Preliminaries
K-Means: Preliminaries

**Datum:** Vector of continuous values
K-Means: Preliminaries

**Datum:** Vector of continuous values
Datum: Vector of continuous values
K-Means: Preliminaries

Datum: Vector of continuous values

\[ x_3 = (1.5, 6.2) \]
K-Means: Preliminaries

Datum: Vector of continuous values

\[ x_3 = (1.5, 6.2) \]
K-Means: Preliminaries

**Datum:** Vector of continuous values

\[ x_3 = (x_{3,1}, x_{3,2}) \]

Feature 1

Feature 2

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_{1,1} )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( x_{2,1} )</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( x_{3,1} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( x_N )</td>
<td>( x_{N,1} )</td>
</tr>
</tbody>
</table>
K-Means: Preliminaries

Datum: Vector of \(D\) continuous values

\[ x_3 = (x_{3,1}, x_{3,2}) \]
**K-Means: Preliminaries**

**Datum**: Vector of D continuous values

Datum: Vector of D continuous values
K-Means: Preliminaries

**Dissimilarity:** Distance as the crow flies
K-Means: Preliminaries

Dissimilarity: Distance as the crow flies
K-Means: Preliminaries

**Dissimilarity**: Distance as the crow flies
K-Means: Preliminaries

**Dissimilarity:** Euclidean distance
K-Means: Preliminaries

**Dissimilarity**: Squared Euclidean distance

\[ dis(x_3, x_{17}) = (x_{3,1} - x_{17,1})^2 \]
\[ + (x_{3,2} - x_{17,2})^2 \]
**K-Means: Preliminaries**

**Dissimilarity:** Squared Euclidean distance

\[
dis(x_3, x_{17}) = \sum_{d=1}^{D} (x_{3,d} - x_{17,d})^2
\]

For each feature
K-Means: Preliminaries

Cluster summary

$K = \text{number of clusters}$
K-Means: Preliminaries

Cluster summary

• K cluster centers
K-Means: Preliminaries

Cluster summary

• K cluster centers

Datum: Vector of continuous values

Datum: Vector of $D$ continuous values
K-Means: Preliminaries

Cluster summary

- K cluster centers
K-Means: Preliminaries

Cluster summary

- K cluster centers

Feature 1

Feature 2
K-Means: Preliminaries

Cluster summary

• K cluster centers

Feature 1

Feature 2

\[ \mu_1 = (\mu_{1,1}, \mu_{1,2}) \]
K-Means: Preliminaries

Cluster summary

- K cluster centers

\[ \mu_1, \mu_2, \ldots, \mu_K \]

\[ \mu_1 = (\mu_{1,1}, \mu_{1,2}) \]
K-Means: Preliminaries

Cluster summary

- K cluster centers $\mu_1, \mu_2, \ldots, \mu_K$
- Data assignments to clusters
K-Means: Preliminaries

Cluster summary

- K cluster centers $\mu_1, \mu_2, \ldots, \mu_K$
- Data assignments to clusters
K-Means: Preliminaries

Cluster summary

- K cluster centers
  \( \mu_1, \mu_2, \ldots, \mu_K \)
- Data assignments to clusters

\( S_k \) = set of points in cluster k

Datum: Vector of continuous values

\[ (x_1, y_1, x_2, y_2, \ldots, x_n, y_n) \]
K-Means: Preliminaries

Cluster summary

• K cluster centers
  \[ \mu_1, \mu_2, \ldots, \mu_K \]

• Data assignments to clusters
  \[ S_1, S_2, \ldots, S_K \]

\( S_k = \) set of points in cluster \( k \)
K-Means: Preliminaries

Cluster summary

• K cluster centers
  \( \mu_1, \mu_2, \ldots, \mu_K \)

• Data assignments to clusters
  \( S_1, S_2, \ldots, S_K \)

\( S_k = \) set of points in cluster \( k \)
K-Means: Preliminaries

Dissimilarity

Feature 1

Feature 2

Feature 1

Feature 2
K-Means: Preliminaries

Dissimilarity (global)

\[
dist_{global} = \sum_{k=1}^{K} \sum_{n:x_n \in S_k} \sum_{d=1}^{D} (x_{n,d} - \mu_{k,d})^2
\]
K-Means: Preliminaries

Dissimilarity (global)

$$dis_{global} = \sum_{k=1}^{K} \sum_{n:x_n \in S_k} \sum_{d=1}^{D} (x_{n,d} - \mu_{k,d})^2$$

For each cluster
K-Means: Preliminaries

Dissimilarity (global)

\[ d_{\text{global}} = \sum_{k=1}^{K} \sum_{n:x_n \in S_k} \sum_{d=1}^{D} (x_{n,d} - \mu_{k,d})^2 \]

For each cluster

For each data point in the kth cluster
K-Means: Preliminaries

Dissimilarity (global)

\[ dis_{global} = \sum_{k=1}^{K} \sum_{n:x_n \in S_k} \sum_{d=1}^{D} (x_{n,d} - \mu_{k,d})^2 \]

For each cluster

For each data point in the kth cluster

For each feature
K-Means: Preliminaries

Dissimilarity (global)

\[ \text{dis}_{\text{global}} = \sum_{k=1}^{K} \sum_{n: x_n \in S_k} \sum_{d=1}^{D} (x_{n,d} - \mu_{k,d})^2 \]
K-Means Algorithm

- Initialize K cluster centers
- Repeat until convergence:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points

![Diagram of K-Means Algorithm](image)
K-Means Algorithm

- Initialize K cluster centers
- Repeat until convergence:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points
**K-Means Algorithm**

- For $k = 1, \ldots, K$
  - Randomly draw $n$ from $1, \ldots, N$ without replacement
  - $\mu_k \leftarrow x_n$

- Repeat until convergence:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points.
K-Means Algorithm

- For $k = 1,\ldots, K$
  - Randomly draw $n$ from $1,\ldots,N$ without replacement
  - $\mu_k \leftarrow x_n$

- Repeat until convergence:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

- For $k = 1, \ldots, K$
  - Randomly draw $n$ from $1, \ldots, N$ without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until convergence:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

- For $k = 1, \ldots, K$
  - Randomly draw $n$ from $1, \ldots, N$ without replacement
  - $\mu_k \leftarrow x_n$

- Repeat until convergence:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points.
K-Means Algorithm

- For \( k = 1, \ldots, K \)
  - Randomly draw \( n \) from \( 1, \ldots, N \) without replacement
  - \( \mu_k \leftarrow x_n \)

- Repeat until \( S_1, \ldots, S_k \) don’t change:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points.
K-Means Algorithm

- For $k = 1, \ldots, K$
  - Randomly draw $n$ from $1, \ldots, N$ without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until convergence:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster’s data points.

Or no change in $\text{dis}_{global}$
K-Means Algorithm

• For $k = 1, \ldots, K$
  ✦ Randomly draw $n$ from $1, \ldots, N$ without replacement
  ✦ $\mu_k \leftarrow x_n$

• Repeat until $S_1, \ldots, S_k$ don’t change:
  ✦ Assign each data point to the cluster with the closest center.
  ✦ Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

- For \( k = 1, \ldots, K \)
  - Randomly draw \( n \) from \( 1, \ldots, N \) without replacement
  - \( \mu_k \leftarrow x_n \)
- Repeat until \( S_1, \ldots, S_k \) don’t change:
  - For \( n = 1, \ldots, N \)
    - Find \( k \) with smallest \( \text{dis}(x_n, \mu_k) \)
    - Put \( x_n \in S_k \) (and no other \( S_j \))
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

- For \( k = 1, \ldots, K \)
  - Randomly draw \( n \) from \( 1, \ldots, N \) without replacement
  - \( \mu_k \leftarrow x_n \)
- Repeat until \( S_1, \ldots, S_k \) don’t change:
  - For \( n = 1, \ldots N \)
    - Find \( k \) with smallest \( dis(x_n, \mu_k) \)
    - Put \( x_n \in S_k \) (and no other \( S_j \))
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

- For $k = 1,\ldots, K$
  - Randomly draw $n$ from $1,\ldots,N$ without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1,\ldots,S_k$ don’t change:
  - For $n = 1,\ldots,N$
    - Find $k$ with smallest $\text{dis}(x_n, \mu_k)$
    - Put $x_n \in S_k$ (and no other $S_j$)
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

• For $k = 1, \ldots, K$
  ✦ Randomly draw $n$ from $1, \ldots, N$ without replacement
  ✦ $\mu_k \leftarrow x_n$
• Repeat until $S_1, \ldots, S_k$ don’t change:
  ✦ For $n = 1, \ldots, N$
    ✤ Find $k$ with smallest $\text{dis}(x_n, \mu_k)$
    ✤ Put $x_n \in S_k$ (and no other $S_j$)
  ✦ Assign each cluster center to be the mean of its cluster’s data points
• For $k = 1, \ldots, K$
  ✦ Randomly draw $n$ from $1, \ldots, N$ without replacement
  ✦ $\mu_k \leftarrow x_n$
• Repeat until $S_1, \ldots, S_K$ don’t change:
  ✦ For $n = 1, \ldots, N$
    ✷ Find $k$ with smallest $\text{dis}(x_n, \mu_k)$
    ✷ Put $x_n \in S_k$ (and no other $S_j$)
  ✦ For $k = 1, \ldots, K$
    ✷ $\mu_k \leftarrow \left| S_k \right|^{-1} \sum_{n : n \in S_k} x_n$
K-Means Algorithm

- For $k = 1,\ldots, K$
  - Randomly draw $n$ from $1,\ldots,N$ without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1,\ldots,S_K$ don’t change:
  - For $n = 1,\ldots,N$
    - Find $k$ with smallest $dis(x_n, \mu_k)$
    - Put $x_n \in S_k$ (and no other $S_j$)
  - For $k = 1,\ldots,K$
    - $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$
K-Means Algorithm

- For \( k = 1, \ldots, K \)
  - Randomly draw \( n \) from 1, \ldots, \( N \) without replacement
  - \( \mu_k \leftarrow x_n \)
- Repeat until \( S_1, \ldots, S_K \) don’t change:
  - For \( n = 1, \ldots, N \)
    - Find \( k \) with smallest \( dis(x_n, \mu_k) \)
    - Put \( x_n \in S_k \) (and no other \( S_j \))
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

- For \( k = 1, \ldots, K \)
  - Randomly draw \( n \) from \( 1, \ldots, N \) without replacement
  - \( \mu_k \leftarrow x_n \)
- Repeat until \( S_1, \ldots, S_k \) don’t change:
  - For \( n = 1, \ldots N \)
    - Find \( k \) with smallest \( \text{dis}(x_n, \mu_k) \)
    - Put \( x_n \in S_k \) (and no other \( S_j \))
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

- For $k = 1, \ldots, K$
  - Randomly draw $n$ from $1, \ldots, N$ without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1, \ldots, S_k$ don’t change:
  - For $n = 1, \ldots, N$
    - Find $k$ with smallest $\text{dis}(x_n, \mu_k)$
    - Put $x_n \in S_k$ (and no other $S_j$)
  - Assign each cluster center to be the mean of its cluster’s data points
K-Means Algorithm

• For $k = 1, \ldots, K$
  ♦ Randomly draw $n$ from $1, \ldots, N$ without replacement
  ♦ $\mu_k \leftarrow x_n$
• Repeat until $S_1, \ldots, S_k$ don’t change:
  ♦ For $n = 1, \ldots, N$
    ♦ Find $k$ with smallest $\text{dis}(x_n, \mu_k)$
    ♦ Put $x_n \in S_k$ (and no other $S_j$)
  ♦ For $k = 1, \ldots, K$
    ♦ $\mu_k \leftarrow |S_k|^{-1} \sum_{n : n \in S_k} x_n$
K-Means Algorithm

- For $k = 1,\ldots, K$
  - Randomly draw $n$ from $1,\ldots,N$ without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1,\ldots,S_k$ don’t change:
  - For $n = 1,\ldots,N$
    - Find $k$ with smallest $\text{dis}(x_n, \mu_k)$
    - Put $x_n \in S_k$ (and no other $S_j$)
  - For $k = 1,\ldots,K$
    - $\mu_k \leftarrow |S_k|^{-1} \sum_{n : n \in S_k} x_n$
K-Means Algorithm

- For $k = 1, \ldots, K$
  - Randomly draw $n$ from $1, \ldots, N$ without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1, \ldots, S_K$ don’t change:
  - For $n = 1, \ldots N$
    - Find $k$ with smallest $dis(x_n, \mu_k)$
    - Put $x_n \in S_k$ (and no other $S_j$)
  - For $k = 1, \ldots, K$
    - $\mu_k \leftarrow \frac{1}{|S_k|} \sum_{n : n \in S_k} x_n$
K-Means Algorithm

• For \( k = 1, \ldots, K \)
  - Randomly draw \( n \) from \( 1, \ldots, N \) without replacement
  - \( \mu_k \leftarrow x_n \)

• Repeat until \( S_1, \ldots, S_K \) don’t change:
  - For \( n = 1, \ldots N \)
    - Find \( k \) with smallest \( \text{dis}(x_n, \mu_k) \)
    - Put \( x_n \in S_k \) (and no other \( S_j \))
  - For \( k = 1, \ldots, K \)
    - \( \mu_k \leftarrow \left| S_k \right|^{-1} \sum_{n:n \in S_k} x_n \)
K-Means Algorithm

- For $k = 1, \ldots, K$
  - Randomly draw $n$ from $1, \ldots, N$ without replacement
  - $\mu_k \leftarrow x_n$

- Repeat until $S_1, \ldots, S_k$ don’t change:
  - For $n = 1, \ldots, N$
    - Find $k$ with smallest $\text{dis}(x_n, \mu_k)$
    - Put $x_n \in S_k$ (and no other $S_j$)
  - For $k = 1, \ldots, K$
    - $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$
K-Means: Evaluation

- For $k = 1, \ldots, K$:
  - Randomly draw $n$ from $1, \ldots, N$ without replacement.
  - Repeat until convergence:
    - Assign each data point to the cluster with the closest center.
    - Assign each cluster center to be the mean of its cluster's data points.

\[ d_k(x_n) = \| \mu_k - x_n \| \]

- For $k = 1, \ldots, K$:
  - For $n = 1, \ldots, N$:
    - Find $k$ with smallest $d_k(x_n)$ and no other $S_j$.
  - For $k = 1, \ldots, K$:
    - Update $\mu_k$ using the formula:
    \[ \mu_k \leftarrow \frac{\sum_{x_n \in S_k} x_n}{|S_k|} \]
K-Means: Evaluation

• Will it terminate?
  Yes. Always.
K-Means: Evaluation

- Will it terminate?
  Yes. Always.
- Is the clustering any good?
  Global dissimilarity only useful for comparing clusterings.
K-Means: Evaluation

- Guaranteed to converge in a finite number of iterations

- Running time per iteration:
  1. Assign data points to closest cluster center 
     \( O(KN) \) time
  2. Change the cluster center to the average of its assigned points 
     \( O(N) \) time
K-Means: Evaluation

Objective

\[ \min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 \]

1. Fix \( \mu \), optimize \( C \):

\[ \min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 = \min_{C} \sum_{i} |x_i - \mu_{x_i}|^2 \]

2. Fix \( C \), optimize \( \mu \):

\[ \min_{\mu} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 \]

- Take partial derivative of \( \mu_i \) and set to zero, we have

\[ \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \]

Step 1 of kmeans

Step 2 of kmeans

K-Means takes an alternating optimization approach, each step is guaranteed to decrease the objective – thus guaranteed to converge.
Demo time…
K-Means Algorithm: Some Issues

- How to set $k$?
- Sensitive to initial centers
  - Multiple initializations
- Sensitive to outliers
- Detects spherical clusters
- Assuming means can be computed
  - It requires continuous, numerical features
Next Lecture:
K-Means Applications,
Spectral clustering,
Hierarchical clustering and
What is a good clustering?