Seeing the Big Picture
Segmenting Images to Create Data
Image Segmentation

• Divide up digital images to salient regions/clusters corresponding to individual surfaces, objects, or natural parts of objects

• Clusters should be uniform and homogenous with respect to certain characteristics (color, intensity, texture)

• **Goal:** Useful and analyzable image representation
Wide Applications

- Medical Imaging
  - Locate tissue classes, organs, pathologies and tumors
  - Measure tissue/tumor volume

- Object Detection
  - Detect facial features in photos
  - Detect pedestrians in footages of surveillance videos

- Recognition tasks
  - Fingerprint/Iris recognition
Various Methods

- **Clustering methods**
  - Partition image to clusters based on differences in pixel colors, intensity or texture

- **Edge detection**
  - Based on the detection of discontinuity, such as an abrupt change in the gray level in gray-scale images

- **Region-growing methods**
  - Divides image into regions, then sequentially merges sufficiently similar regions
In this Recitation…

- Review **hierarchical** and **k-means** clustering in R

- Restrict ourselves to gray-scale images
  - Simple example of a flower image (flower.csv)
  - Medical imaging application with examples of transverse MRI images of the brain (healthy.csv and tumor.csv)

- Compare the use, pros and cons of all analytics methods we have seen so far
Grayscale Images

- Image is represented as a matrix of pixel intensity values ranging from 0 (black) to 1 (white)
- For 8 bits/pixel (bpp), 256 color levels
Grayscale Image Segmentation

- Cluster pixels according to their intensity values

Intensity Vector of size \( n = 7 \times 7 \)

Pairwise distances \( n(n-1)/2 \)
Dendrogram Example

Level 5
Level 4
Level 3
Level 2
Level 1

Distance between clusters BCDEF and BC
Dendrogram Example

ABCDEF

A, BC, DE, F
2 clusters
A, BCDEF

A, BC, DEF
3 clusters

A, BCDEF

A, BCDEF

BCDEF

DEF

DE

BC

COARSER

Level 5

Level 4

Level 3

Level 2

Level 1

2 clusters

A, BC, DE, F

4 clusters

A, BCDEF

15.071x – Seeing the Big Picture: Segmenting Images to Create Data
Dendrogram Example

Level 5

ABCDEF

Level 4

BCDEF

Level 3

DEF

Level 2

BC

DE

Level 1

A

B

C

D

E

F

15.071x – Seeing the Big Picture: Segmenting Images to Create Data
Flower Dendrogram

Cluster Dendrogram

2 clusters

3 clusters

4 clusters
**k-Means Clustering**

- The k-means clustering aims at partitioning the data into k clusters in which each data point belongs to the cluster whose mean is the nearest

**k-Means Clustering Algorithm**

1. Specify desired number of clusters $k$
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat 4 and 5 until no improvement is made
$k$-Means Clustering

$k$-Means Clustering Algorithm

1. Specify desired number of clusters $k$

$k = 2$
**k-Means Clustering**

**k-Means Clustering Algorithm**

1. Specify desired number of clusters $k$ 
2. Randomly assign each data point to a cluster
**$k$-Means Clustering**

$k$-Means Clustering Algorithm

1. Specify desired number of clusters $k$
2. Randomly assign each data point to a cluster

$k = 2$
\(k\)-Means Clustering

\(k\)-Means Clustering Algorithm

1. Specify desired number of clusters \(k\)
2. Randomly assign each data point to a cluster

\(k = 2\)
**$k$-Means Clustering**

### $k$-Means Clustering Algorithm

1. Specify desired number of clusters $k$
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid

---

$k = 2$
### k-Means Clustering

#### k-Means Clustering Algorithm

1. Specify desired number of clusters $k$
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid

$k = 2$
**$k$-Means Clustering**

<table>
<thead>
<tr>
<th>$k$-Means Clustering Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Specify desired number of clusters $k$</td>
</tr>
<tr>
<td>2. Randomly assign each data point to a cluster</td>
</tr>
<tr>
<td>3. Compute cluster centroids</td>
</tr>
<tr>
<td>4. Re-assign each point to the closest cluster centroid</td>
</tr>
<tr>
<td>5. Re-compute cluster centroids</td>
</tr>
<tr>
<td>6. Repeat 4 and 5 until no improvement is made</td>
</tr>
</tbody>
</table>

$k = 2$
First Taste of a Fascinating Field

- MRI image segmentation is subject of ongoing research

- \(k\)-means is a good starting point, but not enough
  - Advanced clustering techniques such as the modified fuzzy \(k\)-means (MFCM) clustering technique
  - Packages in R specialized for medical image analysis
    [http://cran.r-project.org/web/views/MedicalImaging.html](http://cran.r-project.org/web/views/MedicalImaging.html)
3D Reconstruction

- 3D reconstruction from 2D cross sectional MRI images

- Important in the medical field for diagnosis, surgical planning and biological research

MRI image removed due to copyright restrictions.
## Comparison of Methods

<table>
<thead>
<tr>
<th>Used For</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
</table>
| Predicting a continuous outcome (salary, price, number of votes, etc.) | • Simple, well recognized  
• Works on small and large datasets | • Assumes a linear relationship  
\[ Y = a \log(X) + b \] |
| Predicting a categorical outcome (Yes/No, Sell/Buy, Accept/Reject, etc.) | • Computes probabilities that can be used to assess confidence of the prediction | • Assumes a linear relationship |

15.071x – Seeing the Big Picture: Segmenting Images to Create Data
## Comparison of Methods

<table>
<thead>
<tr>
<th>Used For</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CART</strong></td>
<td>Predicting a categorical outcome (quality rating 1--5, Buy/Sell/Hold) or a continuous outcome (salary, price, etc.)</td>
<td>• Can handle datasets without a linear relationship • Easy to explain and interpret</td>
</tr>
<tr>
<td><strong>Random Forests</strong></td>
<td>Same as CART</td>
<td>• Can improve accuracy over CART</td>
</tr>
</tbody>
</table>
## Comparison of Methods

<table>
<thead>
<tr>
<th>Used For</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical Clustering • Finding similar groups</td>
<td>• No need to select number of clusters a priori • Visualize with a dendrogram</td>
<td>• Hard to use with large datasets</td>
</tr>
<tr>
<td>• Clustering into smaller groups and applying predictive methods on groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k-means Clustering Same as Hierarchical Clustering</td>
<td>• Works with any dataset size</td>
<td>• Need to select number of clusters before algorithm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>