Multivariate Statistics
Cluster Methods Lecture 3.
What is clustering?

- A way of grouping together data samples that are similar in some way - according to some criteria that you pick
- A form of unsupervised learning – you generally don’t have examples demonstrating how the data should be grouped together
- So, it’s a method of data exploration – a way of looking for patterns or structure in the data that are of interest
Clustering is arbitrary: from Borges’ Other Inquisitions, discussing an encyclopedia entitled Celestial Emporium of Benevolent Knowledge

“On these remote pages it is written that animals are divided into: a) those that belong to the Emperor; b) embalmed ones; c) those that are trained; d) suckling pigs; e) mermaids; f) fabulous ones; g) stray dogs; h) those that are included in this classification; i) those that tremble as if they were mad; j) innumerable ones; k) those drawn with a very fine camel brush; l) others; m) those that have just broken a flower vase; n) those that resemble flies at a distance.”
Group objects according to their similarity

**Cluster:**
a set of objects that are similar to each other and separated from the other objects.

Example: green/red data points were generated from two different normal distributions.
Clustering algorithms come in two basic flavors:

- **Partitioning**
- **Hierarchical**
Hierarchical Clustering (cont.)

This produces a binary tree or *dendrogram*.

The final cluster is the root and each data item is a leaf.

The height of the bars indicate how close the items are.
Clustering data

- Experiments/samples are given as the row and column vectors of an expression data matrix.
- Clustering may be applied either to objects experiments (regarded as vectors in $\mathbb{R}^o$ or $\mathbb{R}^n$).
Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960 given in this data frame. The variables seem to have been re-scaled to convenient numbers. 16 Variables /Features

- M' percentage of males aged 14-24
- 'So' indicator variable for a southern state
- 'Ed' mean years of schooling
- 'Po1' police expenditure in 1960
- 'Po2' police expenditure in 1959
- 'LF' labour force participation rate
- 'M.F' number of males per 1000 females
- 'Pop' state population
- 'NW' number of nonwhites per 1000 people
- 'U1' unemployment rate of urban males 14-24
- 'U2' unemployment rate of urban males 35-39
- 'GDP' gross domestic product per head
- 'Ineq' income inequality
- 'Prob' probability of imprisonment
- 'Time' average time served in state prison
- 'y' rate of crimes in a particular category per head of population
### Clustering Examples

<table>
<thead>
<tr>
<th>M</th>
<th>So</th>
<th>Ed</th>
<th>Po1</th>
<th>Po2</th>
<th>LF</th>
<th>M.F</th>
<th>Pop</th>
<th>NW</th>
<th>U1</th>
<th>U2</th>
<th>GDP</th>
<th>Ineq</th>
<th>Prob</th>
<th>Time</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>151</td>
<td>1</td>
<td>91</td>
<td>58</td>
<td>56</td>
<td>510</td>
<td>950</td>
<td>33</td>
<td>301</td>
<td>108</td>
<td>41</td>
<td>394</td>
<td>261</td>
<td>0.084602</td>
<td>26.2011</td>
</tr>
<tr>
<td>2</td>
<td>143</td>
<td>0</td>
<td>113</td>
<td>103</td>
<td>95</td>
<td>583</td>
<td>1012</td>
<td>13</td>
<td>102</td>
<td>96</td>
<td>36</td>
<td>557</td>
<td>194</td>
<td>0.029599</td>
<td>25.2999</td>
</tr>
<tr>
<td>3</td>
<td>142</td>
<td>1</td>
<td>89</td>
<td>45</td>
<td>44</td>
<td>533</td>
<td>969</td>
<td>18</td>
<td>219</td>
<td>94</td>
<td>33</td>
<td>318</td>
<td>250</td>
<td>0.083401</td>
<td>24.3006</td>
</tr>
<tr>
<td>4</td>
<td>136</td>
<td>0</td>
<td>121</td>
<td>149</td>
<td>141</td>
<td>577</td>
<td>994</td>
<td>157</td>
<td>80</td>
<td>102</td>
<td>39</td>
<td>673</td>
<td>167</td>
<td>0.015801</td>
<td>29.9012</td>
</tr>
<tr>
<td>5</td>
<td>141</td>
<td>0</td>
<td>121</td>
<td>109</td>
<td>101</td>
<td>591</td>
<td>985</td>
<td>18</td>
<td>30</td>
<td>91</td>
<td>20</td>
<td>578</td>
<td>174</td>
<td>0.041399</td>
<td>21.2998</td>
</tr>
</tbody>
</table>
Clustering Examples

'state.x77': matrix with 50 rows and 8 columns giving the following statistics in the respective columns.

'Population': population estimate as of July 1, 1975

'Income': per capita income (1974)

'Illiteracy': illiteracy (1970, percent of population)

'Life Exp': life expectancy in years (1969-71)

'Murder': murder and non-negligent manslaughter rate per 100,000 population (1976)

'HS Grad': percent high-school graduates (1970)

'Frost': mean number of days with minimum temperature below freezing (1931-1960) in capital or large city

'Area': land area in square miles
## Clustering Example

<table>
<thead>
<tr>
<th>State</th>
<th>Population</th>
<th>Income</th>
<th>Illiteracy</th>
<th>LifeExp</th>
<th>Murder</th>
<th>HSGrad</th>
<th>Frost</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>3615</td>
<td>3624</td>
<td>2.1</td>
<td>69.05</td>
<td>15.1</td>
<td>41.3</td>
<td>20</td>
<td>50708</td>
</tr>
<tr>
<td>Alaska</td>
<td>365</td>
<td>6315</td>
<td>1.5</td>
<td>69.31</td>
<td>11.3</td>
<td>66.7</td>
<td>152</td>
<td>566432</td>
</tr>
<tr>
<td>Arizona</td>
<td>2212</td>
<td>4530</td>
<td>1.8</td>
<td>70.55</td>
<td>7.8</td>
<td>58.1</td>
<td>15</td>
<td>113417</td>
</tr>
<tr>
<td>Arkansas</td>
<td>2110</td>
<td>3378</td>
<td>1.9</td>
<td>70.66</td>
<td>10.1</td>
<td>39.9</td>
<td>65</td>
<td>51945</td>
</tr>
<tr>
<td>California</td>
<td>21198</td>
<td>5114</td>
<td>1.1</td>
<td>71.71</td>
<td>10.3</td>
<td>62.6</td>
<td>20</td>
<td>156361</td>
</tr>
</tbody>
</table>
Clustering Examples: Cereal

The 'UScereal' data frame has 65 rows and 11 columns. The data come from the 1993 ASA Statistical Graphics Exposition, and are taken from the mandatory F&DA food label. The data have been normalized here to a portion of one American cup.

'mfr' Manufacturer, represented by its first initial: G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.

'calories' number of calories in one portion
'protein' grams of protein in one portion
'fat' grams of fat in one portion
'sodium' milligrams of sodium in one portion
'fibre' grams of dietary fibre in one portion
'carbo' grams of complex carbohydrates in one portion
'sugars' grams of sugars in one portion
'shelf' display shelf (1, 2, or 3, counting from the floor)
'potassium' grams of potassium
'vitamins' vitamins and minerals (none, enriched, or 100%)
## Clustering Example

<table>
<thead>
<tr>
<th></th>
<th>mfr calories</th>
<th>protein</th>
<th>fat</th>
<th>sodium</th>
<th>fibre</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Bran</td>
<td>N 212.1212</td>
<td>12.1212</td>
<td>3.0303</td>
<td>393.94</td>
<td>30.303</td>
</tr>
<tr>
<td>All-Bran</td>
<td>K 212.1212</td>
<td>12.1212</td>
<td>3.0303</td>
<td>787.88</td>
<td>27.272</td>
</tr>
<tr>
<td>All-Bran Ex Fiber</td>
<td>K 100.0000</td>
<td>8.0000</td>
<td>0.0000</td>
<td>280.00</td>
<td>28.000</td>
</tr>
<tr>
<td>App Cin Cheerios</td>
<td>G 146.6667</td>
<td>2.6667</td>
<td>2.6667</td>
<td>240.00</td>
<td>2.0000</td>
</tr>
<tr>
<td>Apple Jacks</td>
<td>K 110.0000</td>
<td>2.0000</td>
<td>0.0000</td>
<td>125.00</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>carbo</th>
<th>sugars</th>
<th>shelf</th>
<th>potassium</th>
<th>vitamin</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.15152</td>
<td>3</td>
<td>18.18182</td>
<td>3</td>
<td>848.48485</td>
<td>enriched</td>
</tr>
<tr>
<td>21.21212</td>
<td>3</td>
<td>15.15151</td>
<td>3</td>
<td>969.69697</td>
<td>enriched</td>
</tr>
<tr>
<td>16.00000</td>
<td>0</td>
<td>0.00000</td>
<td>3</td>
<td>660.00000</td>
<td>enriched</td>
</tr>
<tr>
<td>14.00000</td>
<td>1</td>
<td>13.33333</td>
<td>1</td>
<td>93.33333</td>
<td>enriched</td>
</tr>
<tr>
<td>11.00000</td>
<td>2</td>
<td>14.00000</td>
<td>2</td>
<td>30.00000</td>
<td>enriched</td>
</tr>
</tbody>
</table>
How do we define “similarity”? 

- Recall that the goal is to group together “similar” data – but what does this mean?
- No single answer – it depends on what we want to find or emphasize in the data; this is one reason why clustering is an “art”
- The similarity measure is often more important than the clustering algorithm used – don’t overlook this choice!
Covariance

- Variance – measure of the deviation from the mean for points in one dimension e.g. heights
- Covariance as a measure of how much each of the dimensions vary from the mean with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & marks obtained
- The covariance between one dimension and itself is the variance
Representing Covariance between dimensions as a matrix e.g. for 3 dimensions:

\[
C = \begin{pmatrix}
\text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\
\text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\
\text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z)
\end{pmatrix}
\]

Diagonal is the variances of x, y and z

cov(x,y) = cov(y,x) hence matrix is symmetrical about the diagonal

N-dimensional data will result in nxn covariance matrix
Covariance

- Exact value is not as important as it’s sign.

- A positive value of covariance indicates both dimensions increase or decrease together e.g. as the number of hours studied increases, the marks in that subject increase.

- A negative value indicates while one increases the other decreases, or vice-versa e.g. active social life at RIT vs performance in CS dept.

- If covariance is zero: the two dimensions are independent of each other e.g. heights of students vs the marks obtained in a subject
Pattern (Obs) matrix \( \rightarrow \) Proximity matrix

- **Pattern matrix (nxp)**
  - \( p = \) attributes
  - \( n = \) # of observations

\[
\begin{bmatrix}
x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
x_{n1} & \cdots & x_{nf} & \cdots & x_{np}
\end{bmatrix}
\]

- **Proximity matrix (nxn)**
  - \( d(i,j) = \) difference/dissimilarity between \( i \) and \( j \)

\[
\begin{bmatrix}
0 \\
d(2,1) & 0 \\
d(3,1) & d(3,2) & 0 \\
\vdots & \vdots & \vdots \\
d(n,1) & d(n,2) & \cdots & \cdots & 0
\end{bmatrix}
\]
(Dis)similarity measures

- Instead of talking about similarity measures, we often equivalently refer to dissimilarity measures.
- Jagota defines a dissimilarity measure as a function \( f(x,y) \) such that \( f(x,y) > f(w,z) \) if and only if \( x \) is less similar to \( y \) than \( w \) is to \( z \).
- This is always a pair-wise measure.
- Think of \( x \), \( y \), \( w \), and \( z \) as Variable (feature or attributes) profiles/patterns (rows or columns).
Pearson Linear Correlation

- Pearson linear correlation (PLC) is a measure that is invariant to scaling and shifting (vertically) of the expression values.
- Always between –1 and +1 (perfectly anti-correlated and perfectly correlated).
- This is a similarity measure, but we can easily make it into a dissimilarity measure:

\[ d_p = \frac{1 - \rho(x, y)}{2} \]
Different proximity measures

- (Euclidean distance)
  \[ [4^2 + 2^2]^{1/2} = 4.472 \]

- (Manhattan distance)
  \[ 4 + 2 = 6 \]

- ("sup" distance)
  \[ \max\{4,2\} = 4 \]
Normalizing Values

We can alleviate this by normalizing the values of the variables.

For example, we can divide by the mean absolute deviation

\[
\text{Change each } x_i \text{ to } z_i, \text{ where}
\]

\[
z_i = \frac{x_i - m}{S}
\]

\[
m = \frac{1}{n} \{x_0 + x_1 + \ldots + x_n\}
\]

\[
S = \frac{1}{n} \{ |x_0-m| + |x_1-m| + \ldots + |x_n-m| \}
\]
MINKOWSKI R METRICS

\[ d(X, Y) = \sqrt[\gamma]{\sum_{i=1}^{p} |x_i - y_i|^\gamma} \]

- **r = 2** (Euclidean distance)
  \[ [4^2 + 2^2]^{1/2} = 4.472 \]

- **r = 1** (Manhattan distance)
  \[ 4 + 2 = 6 \]

- **r \to \infty** ("sup" distance)
  \[ \text{max}\{4,2\} = 4 \]
Hierarchical Clustering

- Start with every data point in a separate cluster
- Keep merging the most similar pairs of data points/clusters until we have one big cluster left

  This is called a bottom-up or agglomerative method
Hierarchical Clustering (cont.)

- This produces a binary tree or dendrogram.
- The final cluster is the root and each data item is a leaf.
- The height of the bars indicate how close the items are.
Levels of Clustering

a) Six Clusters

b) Four Clusters

c) Three Clusters

d) Two Clusters

e) One Cluster
Hierarchical Clustering Demo
Linkage in Hierarchical Clustering

- We already know about distance measures between data items, but what about between a data item and a cluster or between two clusters?
- We just treat a data point as a cluster with a single item, so our only problem is to define a linkage method between clusters.
- As usual, there are lots of choices...
Hierarchical Algorithms

The key step in hierarchical algorithms is the decision of which sets to join (or split).

One of the drawbacks is that once you’ve made this decision, you can’t go back and split (or re-join) clusters.

There are lots of ways to make these decisions.
Single Linkage

One of the oldest and simplest methods.

aka “Nearest-Neighbor”

Choose the clusters with the shortest distance between the closest object in one cluster and the closest object in the other cluster.
Complete Linkage
aka “Furthest-Neighbor”

Choose the clusters with the shortest distance between the furthest object in one cluster and the furthest object in the other cluster.
Centroid Linkage

Choose the clusters with the shortest distance between the centroid of one cluster and the centroid of the other cluster.
Distances between clusters (summary)

- Calculation of the distance between two clusters is based on the pairwise distances between members of the clusters.
  - **Complete linkage**: largest distance between points
  - **Average linkage**: average distance between points
  - **Single linkage**: smallest distance between points
  - **Centroid**: distance between centroids

Complete linkage gives preference to compact/spherical clusters. Single linkage can produce long stretched clusters.
EXAMPLE

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**a) Single Link**

```
A   B   C   D   E
```

---

**b) Complete Link**

```
A   B   E   C   D
```

---

**b) Average Link**

```
A   B   C   D   E
```
Some Comparisons

Single linkage
Makes long, drawn-out clusters

Complete linkage
Makes compact clusters

Centroid linkage
Somewhere in between
Clustering Example States
dist(state.x77, "euc")

hclust (*, "ward")
dist(scale(state.x77), "euc")

hclust (", "single")
Number of Clusters (Skree plot)
Clustering Examples: Crimes
```r
dist((UScrime), "euc")
hclust(\*, "ward")
```
dist((scale(UScrime)), "euc")
hclust (*, "ward")
```
dist(scale(UScrime), "euc")
hclust(*, "single")
```
Clustering Example: Cereals
Number of Clusters?

8-10
Missing Values

Here are a few options:

- Remove the objects with missing values from the overall clustering
- Replace the missing values of the variable with the average value of the variable over all the objects.
- Use prior knowledge of the variable’s distribution and replace the missing value with the most likely value.
More on Hierarchical Clustering Methods

- **Major advantage**
  - Conceptually very simple
  - Easy to implement → most commonly used technique

- **Major weakness of agglomerative clustering methods**
  - do not scale well: time complexity of at least $O(n^2)$, where $n$ is the number of total objects
  - can never undo what was done previously → high likelihood of getting stuck in local minima
Hierarchical Clustering Issues

- Distinct clusters are not produced – sometimes this can be good, if the data has a hierarchical structure w/o clear boundaries
- There are methods for producing distinct clusters, but these usually involve specifying somewhat arbitrary cutoff values
- What if data doesn’t have a hierarchical structure? Is HC appropriate?
K-means Clustering

- Choose a number of clusters \( k \)
- Initialize cluster centers \( \mu_1, \ldots, \mu_k \)
  - Could pick \( k \) data points and set cluster centers to these points
  - Or could randomly assign points to clusters and take means of clusters
- For each data point, compute the cluster center it is closest to (using some distance measure) and assign the data point to this cluster
- Re-compute cluster centers (mean of data points in cluster)
- Stop when there are no new re-assignments
K-means Clustering (cont.)

How many clusters do you think there are in this data? How might it have been generated?
The K-Means Clustering Results

Example

1. Arbitrarily choose K object as initial cluster center
2. Assign each object to most similar center
3. Update the cluster means
4. Reassign
5. Update the cluster means
6. Reassign

K=2
The K-Means Clustering Method (cntd)
K-means Clustering Demo

\[ k = 2 \]
K-means Clustering Issues

- Random initialization means that you may get different clusters each time
- Data points are assigned to only one cluster (hard assignment)
- Implicit assumptions about the “shapes” of clusters (more about this in project #3)
- You have to pick the number of clusters...