Measuring distances

Applied multivariate statistics – Spring 2012
Overview

- Distances between samples or variables?
- Scaling gives equal weight to all variables
- Dissimilarity is a generalization of Distance
- Dissimilarity for different data types:
  - interval scaled
  - binary (symmetric / asymmetric)
  - nominal
  - ordinal
  - mixed
Different perspective of one thing

- Data context (e.g. biologist, doctor, …) determines distance measure, not statistician
- In practice: Statistician has to offer choices with pros and cons
Between samples or variables?

Rest of this lecture

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>3.4</td>
<td>1.6</td>
</tr>
<tr>
<td>4.3</td>
<td>5.3</td>
<td>5.3</td>
</tr>
<tr>
<td>6.3</td>
<td>9.4</td>
<td>8.9</td>
</tr>
</tbody>
</table>

Use correlation

$$d(X_i, X_j) = \frac{1 - Cor(X_i, X_j)}{2}$$
Properties of distance measures

- D1: \( d(i,j) \geq 0 \)
- D2: \( d(i,i) = 0 \)
- D3: \( d(i,j) = d(j,i) \)
- D4: \( d(i,j) \leq d(i,h) + d(h,j) \) (triangle inequality)
Examples

- **Euclidean distance:**
  \[ d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \]

- **Manhattan distance:**
  \[ d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}| \]

- **Maximum distance:**
  \[ d(i, j) = \left( |x_{i1} - x_{j1}|^\infty + |x_{i2} - x_{j2}|^\infty + \ldots + |x_{ip} - x_{jp}|^\infty \right)^{\frac{1}{\infty}} = \max_{k=1}^{p} |x_{ik} - x_{jk}| \]

- **Special cases of Minkowski distance:**
  \[ d(i, j) = \left( |x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q \right)^{\frac{1}{q}} \]
Intuition for Minkowski Distance

- **p**: Index of Minkowski Distance
- Points on the line have equal Minkowski Distance from center
- **R**: Function “dist”

**Diagram:**
- Manhattan distance
- Euclidean distance
- Maximum distance
Distance metrics in practice

- Euclidean Distance: By far most common
  Our intuitive notion of distance
- Manhattan Distance: Sometimes seen
- Rest: Very rare
To scale or not to scale...
Example 1: cm

- 4 persons

<table>
<thead>
<tr>
<th>Person</th>
<th>Age [years]</th>
<th>Height [cm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35</td>
<td>190</td>
</tr>
<tr>
<td>B</td>
<td>40</td>
<td>190</td>
</tr>
<tr>
<td>C</td>
<td>35</td>
<td>160</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>160</td>
</tr>
</tbody>
</table>
Example 1: feet

- 4 persons

**Table:**

<table>
<thead>
<tr>
<th>Person</th>
<th>Age [years]</th>
<th>Height [feet]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35</td>
<td>6.232</td>
</tr>
<tr>
<td>B</td>
<td>40</td>
<td>6.232</td>
</tr>
<tr>
<td>C</td>
<td>35</td>
<td>5.248</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>5.248</td>
</tr>
</tbody>
</table>
Example 1: scaled

- 4 persons

<table>
<thead>
<tr>
<th>Person</th>
<th>Age [scaled]</th>
<th>Height [scaled]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>B</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>C</td>
<td>-0.87</td>
<td>-0.87</td>
</tr>
<tr>
<td>D</td>
<td>0.87</td>
<td>-0.87</td>
</tr>
</tbody>
</table>

No subgroups anymore
Example 2

- 4 objects

<table>
<thead>
<tr>
<th>Object</th>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>13.3</td>
<td>38.0</td>
</tr>
<tr>
<td>B</td>
<td>12.4</td>
<td>45.4</td>
</tr>
<tr>
<td>C</td>
<td>-122.7</td>
<td>45.6</td>
</tr>
<tr>
<td>D</td>
<td>-122.4</td>
<td>37.7</td>
</tr>
</tbody>
</table>
Need knowledge of context

Example 2

- 4 objects

<table>
<thead>
<tr>
<th>Object</th>
<th>Long.</th>
<th>Lat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palermo</td>
<td>13.3</td>
<td>38.0</td>
</tr>
<tr>
<td>Venice</td>
<td>12.4</td>
<td>45.4</td>
</tr>
<tr>
<td>Portland</td>
<td>-122.7</td>
<td>45.6</td>
</tr>
<tr>
<td>San Francisco</td>
<td>-122.4</td>
<td>37.7</td>
</tr>
</tbody>
</table>
To scale or not to scale…

- If variables are not scaled
  - variable with largest range has most weight
  - distance depends on scale
- Scaling gives every variable equal weight
- Similar alternative is re-weighing:
  \[
  d(i, j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \ldots + w_p(x_{ip} - x_{jp})^2}
  \]
- Scale if,
  - variables measure different units (kg, meter, sec,…)
  - you explicitly want to have equal weight for each variable
- Don’t scale if units are the same for all variables
- Most often: Better to scale.
Dissimilarities

- More flexible than distances
  
  D1: $d(i,j) \geq 0$
  
  D2: $d(i,i) = 0$
  
  D3: $d(i,j) = d(j,i)$

- Example: What do you think, how different are the topics Mathematics, Physics, History on a scale from 0 to 10 (very different)?

- Could also work with “Similarities” (e.g. 1-Dissimilarity)
Dissimilarities for different data types

- **Interval-scaled:**
  - continuous, positive or negative
  - examples: height, weight, temperature, age, cost,...
  Difference of values has a fixed interpretation
  - use metrics we just discussed

- **Ratio-scaled:**
  - continuous, positive
  - example: concentration
  Ratio of values has fixed interpretation
  - use log-transformation, then metrics we just discussed

- **R:**
  - Function “dist” in base distribution (includes Minkowski)
  - Function “daisy” in package “cluster”
Binary symmetric: Simple matching coefficient

- “Symmetric”: No clear asymmetry between group 0 and group 1
- Example: Gender, Right-handed
  Two right-handed people are as similar as two left-handed people
- Counter-example: Having AIDS, being Nobel Laureate
  Two Nobel Laureates are more similar than two non-Nobel-Laureates (e.g. Uni Prof at Harvard without Nobel Prize and baby from Sudan)
Binary symmetric: Simple matching coefficient

<table>
<thead>
<tr>
<th></th>
<th>X=1</th>
<th>X=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>X=0</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

\[ d(i, j) = \frac{b+c}{a+b+c+d} \]

Proportion of variables, in which people disagree
Binary asymmetric: Jaccard distance

Object i

<table>
<thead>
<tr>
<th></th>
<th>X=1</th>
<th>X=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>X=0</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Object j

Simple matching coefficient

\[
d(i, j) = \frac{b+c}{a+b+c}
\]

Uninformative

Proportion of variables, in which people disagree ignoring (0,0)

\[
a+b+c+d = \text{Number of variables}
\]
Nominal: Simple matching coefficient

Simple matching coefficient
- mm: Number of variables in which object i and j mismatch
- p: Number of variables

\[ d(i, j) = \frac{mm}{p} \]

Proportion of variables, in which people disagree
Ordinal: Normalized ranks

- Rank outcome of variable \( f=1,2,\ldots,M: r_{if} \)
- Normalize: \( z_{if} = \frac{r_{if} - 1}{M_f - 1} \)
- Treat \( z_{if} \) as interval-scaled
Mixed: Gower Distance

- Idea: Use distance measure between 0 and 1 for each variable: $d_{ij}^{(f)}$

- Aggregate: $d(i, j) = \frac{1}{p} \sum_{i=1}^{p} d_{ij}^{(f)}$

- Binary (a/s), nominal: Use methods discussed before

- Interval-scaled: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$
  
  $x_{if}$: Value for object $i$ in variable $f$
  $R_f$: Range of variable $f$ for all objects

- Ordinal: Use normalized ranks; then like interval-scaled based on range
Concepts to know

- Effect of scaling / no scaling
- Distance measures for
  - interval scaled
  - binary (s/a)
  - nominal
  - categorical
  - mixed data
R functions to know

- dist, daisy