Clustering with Categorical/Mixed Variables

Cluster Analysis

Similarity Measures and Clustering

Typically, when the data are all categorical, similarity measures are used to determine the proximity between objects. Similarity measures are

- Chosen so that larger numbers indicate close or related objects;
- Often employed for mixed data sets as well, i.e., data sets in which both categorical and continuous data are present;
- Usually scaled to be between 0 and 1 (or 0% and 100%).

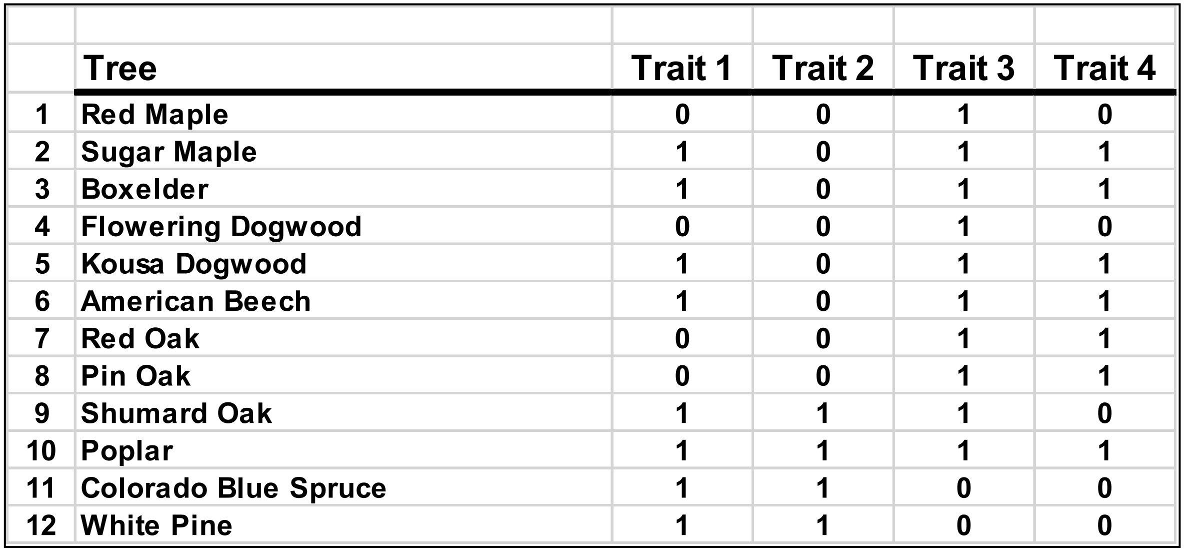
We often denote the similarity coefficient between objects *i* and *j* by s_{ij}, that is

$$s_{ij} = s(O_i, O_j)$$

Example of binary data matrix

We first turn our attention to the case where each of the measured traits is a binary variable. Consider the following data collected from 12 common trees.

Data Matrix:



	Tree	Trait 1	Trait 2	Trait 3	Trait 4
1	Red Maple	0	0	1	0
2	Sugar Maple	1	0	1	1
3	Boxelder	1	0	1	1
4	Flowering Dogwood	0	0	1	0
5	Kousa Dogwood	1	0	1	1
6	American Beech	1	0	1	1
7	Red Oak	0	0	1	1
8	Pin Oak	0	0	1	1
9	Shumard Oak	1	1	1	0
10	Poplar	1	1	1	1
11	Colorado Blue Spruce	1	1	0	0
12	White Pine	1	1	0	0

As an example, these traits could be:

Trait 1: Is the tree shade tolerant (1 = yes, 0 = no)

Trait 2: Does the tree produce edible nuts (1 = yes, 0 = no)

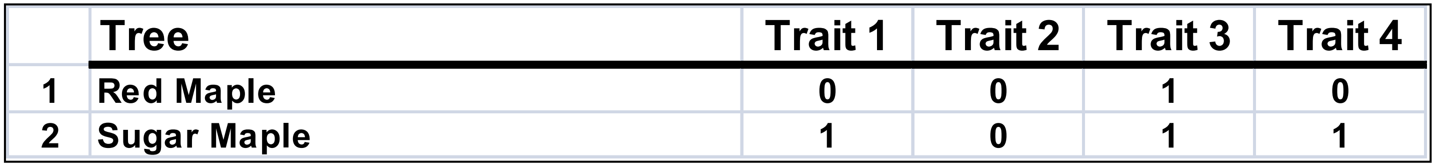
Trait 3: Is the tree susceptible to verticillium wilt (1 = yes, 0 = no)

Trait 4: Is the tree sensitive to leaf scorch in city plantings (1 = yes, 0 = no)

What is the best way to group the trees according to the data collected?

Similarity Measures for Binary Data

Several measures have been suggested for measuring the similarity between binary (yes/no, 0 or 1) data.



	Tree	Trait 1	Trait 2	Trait 3	Trait 4
1	Red Maple	0	0	1	0
2	Sugar Maple	1	0	1	1

Consider the table

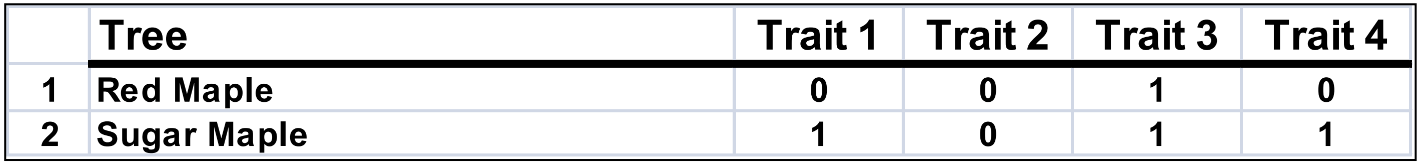
		Individual *i*		
	Outcome	**1**	**0**	**Total**
Individual *j*	**1**	a	b	$a + b$
	0	c	d	$c + d$
	Total	$a + c$	$b + d$	$p = a + b + c + d$

So, a is the number of 1-1 matches between O_i and O_j,
b is the number of 0-1 mismatches between O_i and O_j,
c is the number of 1-0 mismatches between O_i and O_j,
d is the number of 0-0 matches between O_i and O_j,

The total number of binary variables measured for the objects is p and since exactly one of the above cases must hold for each, we have $p = a + b + c + d$.

Similarity Measures for Binary Data

Compute the values for *a*, *b*, *c*, and *d* for the two tree objects shown below.



	Tree	Trait 1	Trait 2	Trait 3	Trait 4
1	Red Maple	0	0	1	0
2	Sugar Maple	1	0	1	1

		Individual *i*		
	Outcome	**1**	**0**	**Total**
Individual *j*	**1**	***a***	***b***	***a + b***
	0	***c***	***d***	***c + d***
	Total	***a + c***	***b + d***	***p = a + b + c + d***

We will let $O_1 = (0, 0, 1, 0)$ and $O_2 = (1, 0, 1, 1)$. I.e Object *i* is the Red Maple and Object *j* is the Sugar Maple. Then

$$a = 1 \qquad b = 2 \qquad c = 0 \qquad d = 1$$

What is the best way to construct a similarity measure between the Red Maple and the Sugar Maple based on the numbers *a*, *b*, *c*, and *d* ? What is the simplest way?

Similarity Measures for Binary Data

The simplest similarity coefficient is the matching coefficient given by:

S1: Matching coefficient $s_{ij} = (a+d)/(a+b+c+d)$

What is the matching coefficient?

The matching coefficient is the proportion of the responses on which the two objects agree.

Compute s_{ij} for the two trees using the matching coefficient.

$$s_{12} = \frac{a+d}{a+b+c+d} = \frac{1+1}{1+2+0+1} = \frac{2}{4} = 0.5$$

What are the potential pitfalls using the matching coefficient?

Many 0-0 matches would suggest related objects, but co-absences may not indicate two objects are similar the same way 1-1 matches would.

Similarity Measures for Binary Data

In addition to the matching coefficient, many different similarity measures for binary data have been proposed. A few of the more popular ones include:

S2: Jaccard's coefficient (1908) $s_{ij} = a/(a+b+c)$

S3: Sokal and Sneath (1963) $s_{ij} = a/[a+2(b+c)]$

S4: Gower and Legendre (1986) $s_{ij} = a/[a+\tfrac{1}{2}(b+c)]$

Measures S2, S3,and S4 were created to deal with the possibility that a zero-zero (co-absence) match does not contain useful information. For a larger list of suggested measures, see Gower, J.C. and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3, 5–48. Gower and Legendre (1986).

Compute s_{ij} for the two trees using each of the similarity measures S2-S4. Compare with the matching coefficient.

S2: (Jaccard's) $$s_{12} = \frac{a}{a+b+c} = \frac{1}{1+2+0} = \frac{1}{3} = 0.\overline{3}$$

S3: $$s_{12} = \frac{a}{a+2(b+c)} = \frac{1}{1+2(2+0)} = \frac{1}{5} = 0.2$$

S4: $$s_{12} = \frac{a}{a+\frac{1}{2}(b+c)} = \frac{1}{1+\frac{1}{2}(2+0)} = \frac{1}{2} = 0.5$$

Note: there are no universally accepted rules for how to handle zero-zero matches. It is important that the researcher collaborate with the data scientist to decide which of the variables carry information in zero-zero matches.

- Give an example where a zero-zero match carries information about the similarity between the two objects.

  A voter is classified as 0 = urban and 1 = rural. Then a 0-0 match carries as much information as a 1-1 match.

- Give another example where a zero-zero match likely does not carry information about the similarity between the two objects.

  Two customers did not buy a particular toaster.

  Two cancers lack a mutation at a particular spot in the genome.

Examples with the `Trees` Data

```
> Trees
                     V2 V3 V4 V5
RedMaple              0  0  1  0
SugarMaple            1  0  1  1
Boxelder              1  0  1  1
FloweringDogwood      0  0  1  0
KousaDogwood          1  0  1  1
AmericanBeech         1  0  1  1
RedOak                0  0  1  1
PinOak                0  0  1  1
ShumardOak            1  1  1  0
Poplar                1  1  1  1
ColoradoBlueSpruce    1  1  0  0
WhitePine             1  1  0  0
```

The Trees data appears to the left. How can we find the matching and Jaccard coefficients?

Using 1 – `dist` with the manhattan method after dividing by the number of traits gives the matching coefficient.

```
> Trees_Match<-1-dist(Trees,method="manhattan")/4
```

as a `dist` object

```
> Trees_Match_M<-1-as.matrix(dist(Trees,method="manhattan")/4)
```

as a similarity matrix

Using 1 – `dist` function with the `binary` method gives the Jaccard coefficient.

```
> Trees_Jaccard<-1-dist(Trees,method="binary")

> Trees_Jaccard_M<-1-as.matrix(dist(Trees,method="binary"))
```

Examples with the `Trees` Data

We can use the `abbreviate` command to produce a shorter version of the tree names.

```
> Trees_Jaccard_M<-1-as.matrix(dist(Trees,method="binary"))
> Trees_Jaccard_M<-round(Trees_Jaccard_M,3)

> rownames(Trees_Jaccard_M)
 [1] "RedMaple"           "SugarMaple"         "Boxelder"           "FloweringDogwood"
 [5] "KousaDogwood"       "AmericanBeech"      "RedOak"             "PinOak"
 [9] "ShumardOak"         "Poplar"             "ColoradoBlueSpruce" "WhitePine"

> rownames(Trees_Jaccard_M)<-abbreviate(rownames(Trees_Jaccard_M))
> colnames(Trees_Jaccard_M)<-abbreviate(colnames(Trees_Jaccard_M))
> Trees_Jaccard_M
      RdMp  SgrM  Bxld  FlwD  KsDg  AmrB  RdOk  PnOk  ShmO Pplr  ClBS  WhtP
RdMp 1.000 0.333 0.333 1.000 0.333 0.333 0.500 0.500 0.333 0.25 0.000 0.000
SgrM 0.333 1.000 1.000 0.333 1.000 1.000 0.667 0.667 0.500 0.75 0.250 0.250
Bxld 0.333 1.000 1.000 0.333 1.000 1.000 0.667 0.667 0.500 0.75 0.250 0.250
FlwD 1.000 0.333 0.333 1.000 0.333 0.333 0.500 0.500 0.333 0.25 0.000 0.000
KsDg 0.333 1.000 1.000 0.333 1.000 1.000 0.667 0.667 0.500 0.75 0.250 0.250
AmrB 0.333 1.000 1.000 0.333 1.000 1.000 0.667 0.667 0.500 0.75 0.250 0.250
RdOk 0.500 0.667 0.667 0.500 0.667 0.667 1.000 1.000 0.250 0.50 0.000 0.000
PnOk 0.500 0.667 0.667 0.500 0.667 0.667 1.000 1.000 0.250 0.50 0.000 0.000
ShmO 0.333 0.500 0.500 0.333 0.500 0.500 0.250 0.250 1.000 0.75 0.667 0.667
Pplr 0.250 0.750 0.750 0.250 0.750 0.750 0.500 0.500 0.750 1.00 0.500 0.500
ClBS 0.000 0.250 0.250 0.000 0.250 0.250 0.000 0.000 0.667 0.50 1.000 1.000
WhtP 0.000 0.250 0.250 0.000 0.250 0.250 0.000 0.000 0.667 0.50 1.000 1.000
```

Converting Similarity to Dissimilarity

To perform Hierarchal clustering, a similarity matrix **S** is often converted to a dissimilarity matrix **D** using a transformation of **S**. The two most often used are

$$d_{ij} = 1 - s_{ij} \qquad \text{for } 1 \le i, j \le n$$

and

$$d_{ij} = \sqrt{1 - s_{ij}} \qquad \text{for } 1 \le i, j \le n$$

where the similarities are assumed to have been scaled between 0 and 1.

As we have seen, the `hclust` function accepts a `dist` object as the starting point to perform the hierarchal clustering algorithm.

Exercise

Revisit the `vertebrates` dataset from the homework.

Create a dendrogram where 0-0 matches contain no useful information.

Use single linkage clustering.

Other Categorical Data

Many categorical variables have more than one level. Examples include

- Blood Type
- Eye Color
- Political Party Affiliation

How should we deal with data of this kind?

One possibility is to create a new binary variable for each level in the category. For example, ignoring the Rh factor, there are 4 different blood types, A, B, AB, and O. Following the suggestion above, we would create four binary variables:

BT1: Does the individual have type A blood (1 = yes, 0 = no);

BT2: Does the individual have type B blood (1 = yes, 0 = no);

BT3: Does the individual have type AB blood (1 = yes, 0 = no);

BT4: Does the individual have type O blood (1 = yes, 0 = no).

What issues would this type of solution generate?

This would create many 0-0 matches which we have seen can be difficult to interpret.

Other Categorical Data

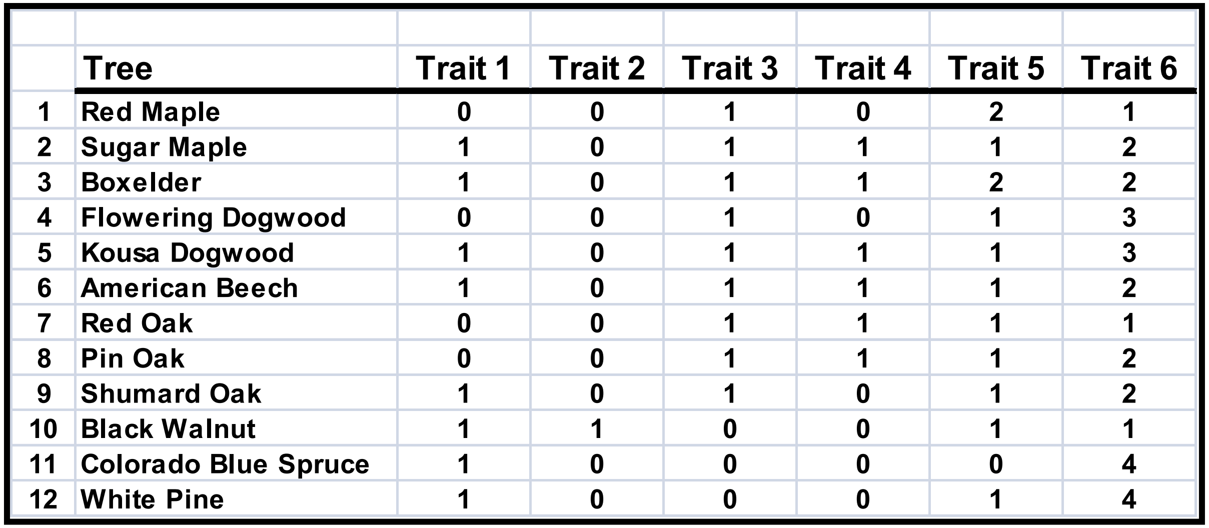
Typically, the creation of multiple binary variables to analyze categorical data involving more than one level is considered to be an inferior solution to the problem because of the large number of 0-0 matches that are necessarily created.

In general, this situation is approached by creating a single binary variable for each category which records whether or not the two individuals agree on the category in question. That is, for each categorical variable *k*, we create a binary variable, s_{ijk}, which is 1 if object *i* and object *j* agree on trait *k* and zero otherwise.

$$s_{ijk} = s_k\left(O_i, O_j\right) = \begin{cases} 1, & \text{if } O_i \text{ and } O_j \text{ agree on trait } k \\ 0, & \text{otherwise.} \end{cases}$$

Clustering with Categorical Data

Consider the following expansion of the data matrix from the tree example:



	Tree	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6
1	Red Maple	0	0	1	0	2	1
2	Sugar Maple	1	0	1	1	1	2
3	Boxelder	1	0	1	1	2	2
4	Flowering Dogwood	0	0	1	0	1	3
5	Kousa Dogwood	1	0	1	1	1	3
6	American Beech	1	0	1	1	1	2
7	Red Oak	0	0	1	1	1	1
8	Pin Oak	0	0	1	1	1	2
9	Shumard Oak	1	0	1	0	1	2
10	Black Walnut	1	1	0	0	1	1
11	Colorado Blue Spruce	1	0	0	0	0	4
12	White Pine	1	0	0	0	1	4

Trait 1: Is the tree shade tolerant (1 = yes, 0 = no)

Trait 2: Does the tree produce edible nuts (1 = yes, 0 = no)

Trait 3: Is the tree susceptible to verticillium wilt (1 = yes, 0 = no)

Trait 4: Is the tree sensitive to leaf scorch in city plantings (1 = yes, 0 = no)

Trait 5: Type of soil in which the tree thrives (0 = dry, 1 = moist, 2 = wet)

Trait 6: Crown Shape (round = 1, oval = 2, vase = 3, pyramidal = 4)

We observe that traits 5 and 6 are categorical with more than two levels.

Other Categorical Data

The resulting similarity coefficient for an object with *p* categorical traits (and no continuous ones) then becomes:

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk} = \frac{1}{p}\sum_{k=1}^{p} s_k(O_i, O_j).$$

Note that all binary variables are a special cases of categorical variables in which the number of categories is two and as such the above similarity coefficient can be used with data sets containing both binary and multiple level categorical data.

Note that, as written, s_{ij} treats 1-1 and 0-0 matches the same. More on this shortly.

Calculate $s_{9\,10}$, the similarity between the Shumard Oak and Black Walnut trees, using the similarity measure above

We have $O_9 = (1, 0, 1, 0, 1, 2)$ and $O_{10} = (1, 1, 0, 0, 1, 1)$. Thus,

$$s_{9\,10} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk} = \frac{1}{6}\sum_{k=1}^{6} s_k(O_9, O_{10}) = \frac{1}{6}[1+0+0+1+1+0] = \frac{3}{6} = 0.5$$

Generalized Similarity Measure

A generalized similarity measure proposed by Gower (1971) is given by:

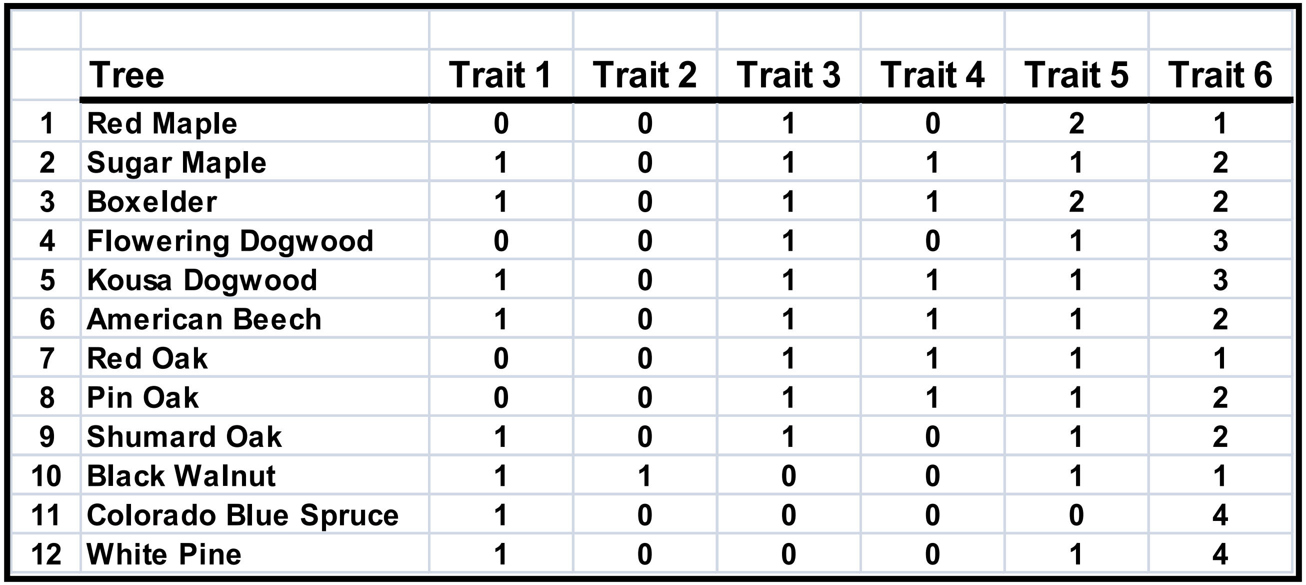
$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} \Big/ \sum_{k=1}^{p} w_{ijk}$$

For each k, $1 \leq k \leq p$, The variable w_{ijk} is an <u>indicator variable</u>, meaning the value of w_{ijk} is 0 or 1 for any (i, j, k) triple. The value of w_{ijk} assigned depending on whether or not the comparison is considered valid. For categorical variables components are assigned a value of one when the two individuals have the same value and zero otherwise.

$$w_{ijk} = w_k(O_i, O_j) = \begin{cases} 1, & \text{comparison between } O_i \text{ and } O_j \\ & \text{in variable } k \text{ is valid} \\ 0, & \text{otherwise.} \end{cases}$$

Here, the w_{ijk} variables are typically used to remove 0-0 matches when they are deemed to contain no useful information for clustering the objects.

Generalized Similarity Measure



	Tree	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6
1	Red Maple	0	0	1	0	2	1
2	Sugar Maple	1	0	1	1	1	2
3	Boxelder	1	0	1	1	2	2
4	Flowering Dogwood	0	0	1	0	1	3
5	Kousa Dogwood	1	0	1	1	1	3
6	American Beech	1	0	1	1	1	2
7	Red Oak	0	0	1	1	1	1
8	Pin Oak	0	0	1	1	1	2
9	Shumard Oak	1	0	1	0	1	2
10	Black Walnut	1	1	0	0	1	1
11	Colorado Blue Spruce	1	0	0	0	0	4
12	White Pine	1	0	0	0	1	4

Calculate $s_{9\,10}$, the similarity between the Shumard Oak and Black Walnut trees, using the generalized similarity measure under the assumptions that:

1. 0-0 matches for Trait 2 contain no useful information
2. 0-0 matches for Trait 4 contain no useful information

Given:

1. 0-0 matches for Trait 2 contain no useful information
2. 0-0 matches for Trait 4 contain no useful information

We have $O_9 = (1, 0, 1, 0, 1, 2)$ and $O_{10} = (1, 1, 0, 0, 1, 1)$. Thus,

$$s_{9\,10} = \sum_{k=1}^{6} w_{9\,10k} s_{9\,10k} \Big/ \sum_{k=1}^{6} w_{9\,10k}$$

$$= \frac{w_{9\,10\,1}s_{9\,10\,1} + w_{9\,10\,2}s_{9\,10\,2} + w_{9\,10\,3}s_{9\,10\,3} + w_{9\,10\,4}s_{9\,10\,4} + w_{9\,10\,5}s_{9\,10\,5} + w_{9\,10\,6}s_{9\,10\,6}}{w_{9\,10\,1} + w_{9\,10\,2} + w_{9\,10\,3} + w_{9\,10\,4} + w_{9\,10\,5} + w_{9\,10\,6}}$$

$$= \frac{w_{9\,10\,1}(1) + w_{9\,10\,2}(0) + w_{9\,10\,3}(0) + w_{9\,10\,4}(1) + w_{9\,10\,5}(1) + w_{9\,10\,6}(0)}{w_{9\,10\,1} + w_{9\,10\,2} + w_{9\,10\,3} + w_{9\,10\,4} + w_{9\,10\,5} + w_{9\,10\,6}}$$

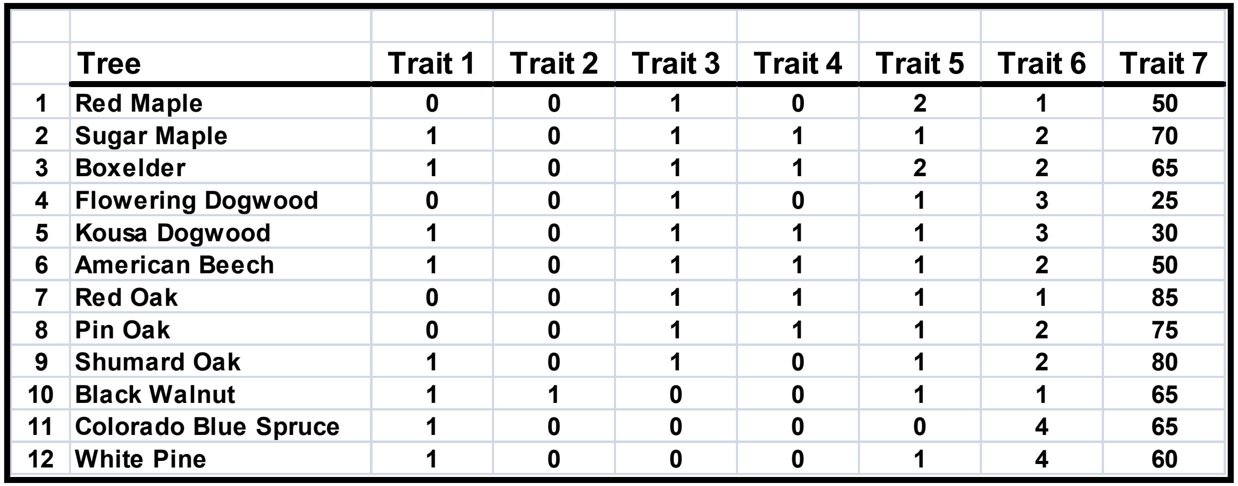
$$= \frac{(1)(1) + (1)(0) + (1)(0) + (0)(1) + (1)(1) + (1)(0)}{1 + 1 + 1 + 0 + 1 + 1}$$

$$= \frac{2}{5} = 0.4$$

Clustering with Mixed Data Types

Similarity measures are also often employed when clustering objects according to mixed (categorical and continuous) data sets.

Consider the following expansion of the data matrix from the tree example:



	Tree	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6	Trait 7
1	Red Maple	0	0	1	0	2	1	50
2	Sugar Maple	1	0	1	1	1	2	70
3	Boxelder	1	0	1	1	2	2	65
4	Flowering Dogwood	0	0	1	0	1	3	25
5	Kousa Dogwood	1	0	1	1	1	3	30
6	American Beech	1	0	1	1	1	2	50
7	Red Oak	0	0	1	1	1	1	85
8	Pin Oak	0	0	1	1	1	2	75
9	Shumard Oak	1	0	1	0	1	2	80
10	Black Walnut	1	1	0	0	1	1	65
11	Colorado Blue Spruce	1	0	0	0	0	4	65
12	White Pine	1	0	0	0	1	4	60

Trait 1: Is the tree shade tolerant (1 = yes, 0 = no)

Trait 2: Does the tree produce edible nuts (1 = yes, 0 = no)

Trait 3: Is the tree susceptible to verticillium wilt (1 = yes, 0 = no)

Trait 4: Is the tree sensitive to leaf scorch in city plantings (1 = yes, 0 = no)

Trait 5: Type of soil in which the tree thrives (0 = dry, 1 = moist, 2 = wet)

Trait 6: Crown Shape (round = 1, oval = 2, vase = 3, pyramidal = 4)

Trait 7: Average height of tree (ft.)

What is the best way to group the trees according to the data collected?

Similarity Measures for Mixed-Mode Data

When the data matrix for the objects contains both continuous and categorical data, we expand the similarity measure proposed by Gower (1971)

$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} \Big/ \sum_{k=1}^{p} w_{ijk}$$

to include provisions for the continuous data. Again, the variable w_{ijk} is an indicator variable assigned a 0 or 1 depending on whether or not the comparison is considered valid. As before, all categorical variables, s_{ijk}, are assigned a value of 1 when the two objects agree on trait k and 0 otherwise. The continuous portion of the analysis is scaled with:

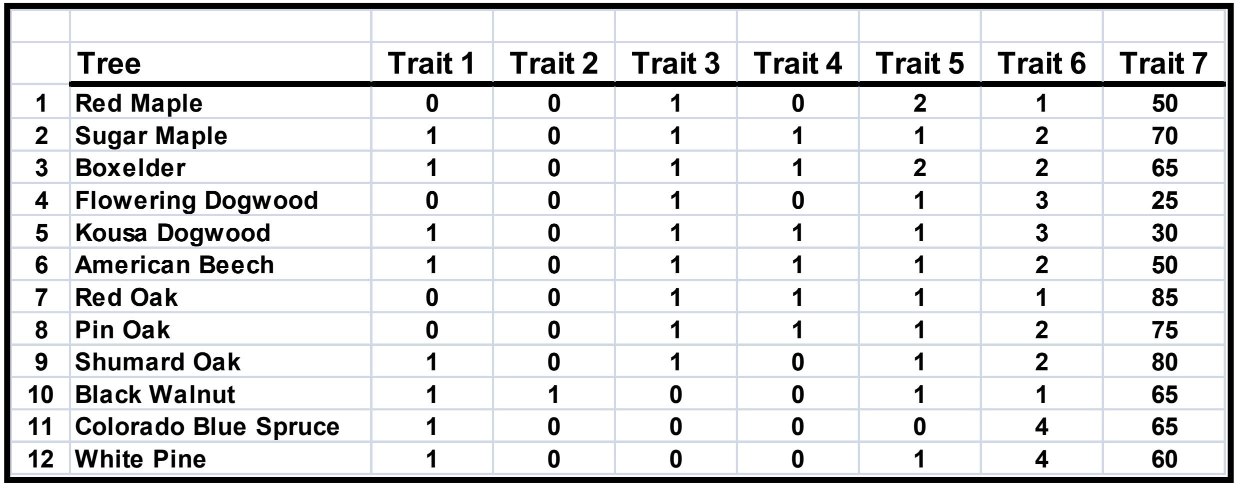
$$s_{ijk} = s_k(O_i, O_j) = 1 - \left| x_{ik} - x_{jk} \right| / R_k$$

where R_k is the range of observations for the k^{th} variable.

Why are we using the range here instead some function based on the standard deviation?

We want the variables to have approximately the same influence on the similarity measure and this restricts the value between 0 and 1.

Similarity Measures for Mixed-Mode Data



	Tree	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6	Trait 7
1	Red Maple	0	0	1	0	2	1	50
2	Sugar Maple	1	0	1	1	1	2	70
3	Boxelder	1	0	1	1	2	2	65
4	Flowering Dogwood	0	0	1	0	1	3	25
5	Kousa Dogwood	1	0	1	1	1	3	30
6	American Beech	1	0	1	1	1	2	50
7	Red Oak	0	0	1	1	1	1	85
8	Pin Oak	0	0	1	1	1	2	75
9	Shumard Oak	1	0	1	0	1	2	80
10	Black Walnut	1	1	0	0	1	1	65
11	Colorado Blue Spruce	1	0	0	0	0	4	65
12	White Pine	1	0	0	0	1	4	60

Calculate $s_{9\,10}$, the similarity between the Shumard Oak and Black Walnut trees, using the generalized similarity measure for mixed mode data, again under the assumptions that

1. 0-0 matches for Trait 2 contain no useful information;
2. 0-0 matches for Trait 4 contain no useful information.

Trait 7 is continuous and hence

$$s_{ij7} = s_7(O_i, O_j) = 1 - \left| x_{i7} - x_{j7} \right| / R_7 \qquad \text{where} \qquad R_7 = 85 - 25 = 60$$

$$= 1 - \left| x_{i7} - x_{j7} \right| / 60$$

Similarity Measures for Mixed-Mode Data

Given:

1. 0-0 matches for Trait 2 contain no useful information
2. 0-0 matches for Trait 4 contain no useful information

We have O_9 = (1, 0, 1, 0, 1, 2, 80) and O_{10} = (1, 1, 0, 0, 1, 1, 65). Thus,

$$s_{9\,10\,7} = s_7(O_9, O_{10}) = 1 - \left| x_{97} - x_{107} \right| / 60$$

$$= 1 - \left| 80 - 65 \right| / 60$$

$$= 0.75$$

$$s_{9\,10} = \sum_{k=1}^{7} w_{9\,10k} s_{9\,10k} \Bigg/ \sum_{k=1}^{7} w_{9\,10k}$$

$$= \frac{w_{9\,10\,1}s_{9\,10\,1} + w_{9\,10\,2}s_{9\,10\,2} + w_{9\,10\,3}s_{9\,10\,3} + w_{9\,10\,4}s_{9\,10\,4} + w_{9\,10\,5}s_{9\,10\,5} + w_{9\,10\,6}s_{9\,10\,6} + w_{9\,10\,7}s_{9\,10\,7}}{w_{9\,10\,1} + w_{9\,10\,2} + w_{9\,10\,3} + w_{9\,10\,4} + w_{9\,10\,5} + w_{9\,10\,6} + w_{9\,10\,7}}$$

$$= \frac{w_{9\,10\,1}(1) + w_{9\,10\,2}(0) + w_{9\,10\,3}(0) + w_{9\,10\,4}(1) + w_{9\,10\,5}(1) + w_{9\,10\,6}(0) + w_{9\,10\,7}(0.75)}{w_{9\,10\,1} + w_{9\,10\,2} + w_{9\,10\,3} + w_{9\,10\,4} + w_{9\,10\,5} + w_{9\,10\,6} + w_{9\,10\,7}}$$

$$= \frac{(1)(1) + (1)(0) + (1)(0) + (0)(1) + (1)(1) + (1)(0) + (1)(0.75)}{1 + 1 + 1 + 0 + 1 + 1 + 1} = \frac{2.75}{6} = 0.458$$

Implementing Gower’s Measure for Mixed-Mode Data

Consider the following example of the data frame `HeartData` containing variables on patients in a cardiac clinic

Sex:	0 = Female, 1 = Male
HBP:	High Blood Pressure: 0 = No, 1 = Yes
BType:	1 = A, 2 = B, 3 = AB, 4 = O
Smoking:	1 = Nonsmoker, 2 = Moderate Smoker, 3 = Heavy Smoker
MVP:	Mitral Valve Prolapse: 0 = No, 1 = Yes
MVS:	Mitral Valve Stenosis: 0 = No, 1 = Yes
LDL:	“bad cholesterol” continuous variable
HDL:	“good cholesterol” continuous variable

```
> head(HeartData)
  sex HBP BType smoking MVP MVS   LDL  HDL
1   0   0     4       1   0   0 172.5 45.5
2   1   0     4       1   0   0 145.0 52.3
3   0   0     1       2   0   1 181.5 56.3
4   1   0     4       1   1   0 187.7 47.7
5   1   0     1       2   0   0 177.1 53.6
6   0   1     4       2   0   0 104.3 40.4
```

Function `daisy` in the `cluster` Package

The function `daisy` is part of the `cluster` package in R. The `cluster` package is part of base R and can be accessed using

```
> library(cluster)
```

The `daisy` function allows for dissimilarity calculation for mixed-mode data using Gower's formulation. Variables that are to be treated in a particular way need to have their `type` specified (see below); type `?daisy` for more information.

The function `daisy` can calculate the dissimilarity based on Gower's function (Actually, it does 1 – (this function) since it is computing dissimilarities.) If a data frame containing mixed-mode data is submitted, the variables that are stored as factors are treated as categorical variables as described in the Class Notes. Numerical data is scaled using Gower's formulation as well. The treatment of binary variables can be set using the optional argument `type` to indicate how 0-0 matches are to be handled.

When setting the types of variables, use the following

`asymm` asymmetric binary variable: this designation discards any 0-0 matches

`symm` symmetric binary variable: this designation keeps (retains) 0-0 matches

Variables in the `HeartData` Data Frame

Let’s look at the variables in the `HeartData` data frame:

Sex:	0 = Female, 1 = Male	**Binary variables where both 1-1 and 0-0 matches indicate related objects** symmetric binary variables
HBP:	High Blood Pressure: 0 = No, 1 = Yes	
BType:	1 = A, 2 = B, 3 = AB, 4 = O	**Categorical (Nominal)**
Smoking:	1 = None, 2 = Moderate, 3 = Heavy	**Categorical (Ordinal)**
MVP:	Mitral Valve Prolapse: 0 = No, 1 = Yes	**Binary variables where and 0-0 do not indicate related objects** asymmetric binary variables
MVS:	Mitral Valve Stenosis: 0 = No, 1 = Yes	
LDL:	“bad cholesterol” continuous variable	**Continuous variables**
HDL:	“good cholesterol” continuous variable	

Variables in the `HeartData` Data Frame

```
> summary(HeartData)
      sex             HBP            BType          smoking           MVP             MVS       
 Min.   :0.00   Min.   :0.0   Min.   :1.0   Min.   :1.0   Min.   :0.0   Min.   :0.0  
 1st Qu.:0.00   1st Qu.:0.0   1st Qu.:1.0   1st Qu.:1.0   1st Qu.:0.0   1st Qu.:0.0  
 Median :0.00   Median :0.0   Median :3.0   Median :1.0   Median :0.0   Median :0.0  
 Mean   :0.30   Mean   :0.1   Mean   :2.6   Mean   :1.4   Mean   :0.1   Mean   :0.1  
 3rd Qu.:0.75   3rd Qu.:0.0   3rd Qu.:4.0   3rd Qu.:2.0   3rd Qu.:0.0   3rd Qu.:0.0  
 Max.   :1.00   Max.   :1.0   Max.   :4.0   Max.   :2.0   Max.   :1.0   Max.   :1.0  
      LDL              HDL       
 Min.   :104.3   Min.   :40.40  
 1st Qu.:151.1   1st Qu.:45.58  
 Median :174.8   Median :49.55  
 Mean   :165.4   Mean   :50.29  
 3rd Qu.:185.4   3rd Qu.:53.27  
 Max.   :203.3   Max.   :64.40  
```

Not treating these as categorical variables (i.e. factors)

```
> HeartData$BType<-factor(HeartData$BType)
> summary(HeartData)
      sex             HBP        BType     smoking           MVP             MVS             LDL       
 Min.   :0.00   Min.   :0.0   1:4   Min.   :1.0   Min.   :0.0   Min.   :0.0   Min.   :104.3  
 1st Qu.:0.00   1st Qu.:0.0   2:1   1st Qu.:1.0   1st Qu.:0.0   1st Qu.:0.0   1st Qu.:151.1  
 Median :0.00   Median :0.0   4:5   Median :1.0   Median :0.0   Median :0.0   Median :174.8  
 Mean   :0.30   Mean   :0.1         Mean   :1.4   Mean   :0.1   Mean   :0.1   Mean   :165.4  
 3rd Qu.:0.75   3rd Qu.:0.0         3rd Qu.:2.0   3rd Qu.:0.0   3rd Qu.:0.0   3rd Qu.:185.4  
 Max.   :1.00   Max.   :1.0         Max.   :2.0   Max.   :1.0   Max.   :1.0   Max.   :203.3  
      HDL       
 Min.   :40.40  
 1st Qu.:45.58  
 Median :49.55  
 Mean   :50.29  
 3rd Qu.:53.27  
 Max.   :64.40  
```

`BType` is now being treated as a factor

We will leave `smoking` as a numeric variable, since this captures the relationship better than a nominal categorical variable

Using the daisy Function

```
> Heart_Dist<-daisy(HeartData,type=list(symm=c(1,2),asymm=c(5,6)))
> round(Heart_Dist,3)
Dissimilarities :
       1     2     3     4     5     6     7     8     9
2  0.260
3  0.434 0.576
4  0.321 0.232 0.615
5  0.481 0.313 0.308 0.408
6  0.400 0.568 0.635 0.664 0.714
7  0.213 0.380 0.261 0.477 0.278 0.602
8  0.238 0.404 0.341 0.458 0.424 0.472 0.370
9  0.221 0.477 0.308 0.462 0.348 0.621 0.096 0.407
10 0.410 0.578 0.500 0.672 0.557 0.490 0.446 0.482 0.464

Metric :  mixed ;  Types = S, S, N, I, A, A, I, I
Number of objects : 10
```

S: symmetric binary
A: asymmetric binary
N: Nominal (categorical)
I: Interval scaled (continuous)

We could have also used

```
> Heart_Dist<-daisy(HeartData,type=list(symm=c("sex","HBP"),asymm=c("MVP","MVS")))
```

Obtaining a Similarity Matrix

We could produce a similarity matrix using

```
> Heart_Sim<-1-round(as.matrix(Heart_Dist),3)
> Heart_Sim
       1     2     3     4     5     6     7     8     9    10
1  1.000 0.740 0.566 0.679 0.519 0.600 0.787 0.762 0.779 0.590
2  0.740 1.000 0.424 0.768 0.687 0.432 0.620 0.596 0.523 0.422
3  0.566 0.424 1.000 0.385 0.692 0.365 0.739 0.659 0.692 0.500
4  0.679 0.768 0.385 1.000 0.592 0.336 0.523 0.542 0.538 0.328
5  0.519 0.687 0.692 0.592 1.000 0.286 0.722 0.576 0.652 0.443
6  0.600 0.432 0.365 0.336 0.286 1.000 0.398 0.528 0.379 0.510
7  0.787 0.620 0.739 0.523 0.722 0.398 1.000 0.630 0.904 0.554
8  0.762 0.596 0.659 0.542 0.576 0.528 0.630 1.000 0.593 0.518
9  0.779 0.523 0.692 0.538 0.652 0.379 0.904 0.593 1.000 0.536
10 0.590 0.422 0.500 0.328 0.443 0.510 0.554 0.518 0.536 1.000
```

```
> head(HeartData)
  sex HBP BType smoking MVP MVS   LDL  HDL
1   0   0     4       1   0   0 172.5 45.5
2   1   0     4       1   0   0 145.0 52.3
3   0   0     1       2   0   1 181.5 56.3
4   1   0     4       1   1   0 187.7 47.7
5   1   0     1       2   0   0 177.1 53.6
6   0   1     4       2   0   0 104.3 40.4
```

Another Look at the `smoking` Variable

The ordinal variable smoking was coded as

Smoking: 1 = Nonsmoker, 2 = Moderate Smoker, 3 = Heavy Smoker

We let R maintain the numerical coding of this variable for the `daisy` function. So, the contribution for the similarity here is

$$s_{ijk} = s_k(O_i, O_j) = 1 - \left| x_{ik} - x_{jk} \right| / R_k$$

$$s_4(\text{Non, Moderate}) = 1 - \frac{|1-2|}{3-1} = 1 - \frac{1}{2} = \frac{1}{2} \qquad s_4(\text{Moderate, Heavy}) = 1 - \frac{|2-3|}{3-1} = 1 - \frac{1}{2} = \frac{1}{2}$$

$$s_4(\text{Non, Heavy}) = 1 - \frac{|1-3|}{3-1} = 1 - \frac{2}{2} = 0$$

Another Look at the `smoking` Variable

We could (after talking with the cardiologist) assign the variables as

Smoking: 1 = Nonsmoker, 4 = Moderate Smoker, 6 = Heavy Smoker

So, the contribution to the similarity would be

$$s_4(\text{Non, Moderate}) = 1 - \frac{|1-4|}{6-1} = 1 - \frac{3}{5} = \frac{2}{5} \qquad s_4(\text{Moderate, Heavy}) = 1 - \frac{|4-6|}{6-1} = 1 - \frac{2}{5} = \frac{3}{5}$$

$$s_4(\text{Non, Heavy}) = 1 - \frac{|1-6|}{6-1} = 1 - \frac{5}{5} = 0$$

Perhaps this better describes the relationship between smoking and the similarity between the patients. Remember, there are no hard rules for this kind of analysis and collaboration with subject matter experts is crucial!