

The background of the slide is a complex, abstract composition. It features a network of thin, intersecting lines in shades of brown and orange, creating a web-like structure. Scattered throughout this network are numerous small, colored dots in green, blue, and orange. On the left side, there is a vertical strip with a grid of small, light-colored squares. In the upper left corner, there is a small, semi-transparent rectangular area containing a cluster of orange and red dots. The overall aesthetic is technical and data-oriented.

Distance between Categorical Attributes, Ordinal Attributes, and Mixed Types

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes

- Example: Color (red, yellow, blue, green), profession, etc.

- Method 1: Simple matching

- m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p} = \frac{\text{mismatches}}{\text{total}}$$

- Method 2: Use a large number of binary attributes

categorical \rightarrow set of binary

- Creating a new binary attribute for each of the M nominal states

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
 - Replace *an ordinal variable value* by its rank: $r_{if} \in \{1, \dots, M_f\}$
 - Map the range of each variable onto $[0, 1]$ by replacing i -th object in the f -th variable by
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
 - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
 - Then distance: $d(\text{freshman}, \text{senior}) = 1$, $d(\text{junior}, \text{senior}) = 1/3$
 - Compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A dataset may contain all attribute types
 - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

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

- If f is numeric: Use the normalized distance
- If f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise
- If f is ordinal
 - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
 - Treat z_{if} as interval-scaled