CSCI568

Discussion: Similarity Metrics

Dis/similarity Between Two Attributes

Type	Dissimilarity	Similarity
Nominal		
Ordinal		
Interval/Ratio		

Dis/similarities Between Data Objects

- Euclidean distance
- Pearson Correlation Coefficient
- Simple Matching Coefficient (SMC)
- Jaccard / Tanimoto
- Cosine Similarity
- Bregman Divergence

Minkowski Distance Metric

- General distance calculation
- r=1 “City Block”
- r=2 “Euclidean”
- r=(infinity) “Supremum” (think lim(r->inf.))

DM 70

Euclidean Distance

Simple! Linear distance between two points.

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

x_k and y_k are values of k^{th} attribute of objects x and y

DM 69 - 71

Simple Matching Coefficient

Linear distance is good for many things, but not necessarily binary data!

Simple Matching Coefficient

Simple!

$$\text{SMC} = \frac{\text{\# of matching attributes}}{\text{\# of attributes}}$$

$$\text{SMC} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

Jaccard Coefficient

Simple!

$$\text{Jaccard} = \frac{\text{\# of matching } \textit{present} \text{ attributes}}{\text{\# of attributes w/ values } \textit{present}}$$

$$\text{Jaccard} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

SMC vs. Jaccard

Like SMC, but for asymmetric binary attributes.
(we only care about presence)

Think: market basket data (sparse dataset, asymmetric/ binary attributes)

SMC --> most transactions are alike (everyone doesn’t purchase most items

Jaccard --> only compares attributes w/ existing values

SMC / Jaccard Example

$$x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$
$$y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$$

$f_{01} = 2$
$f_{10} = 1$
$f_{00} = 7$
$f_{11} = 0$

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = \frac{0+7}{2+1+0+7}$$

$$Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2+1}$$

Cosine Similarity

Often used for document word-frequency.

$$\text{cos\_sim}(x,y) = \frac{x \cdot y}{||x||\ ||y||}$$

Cosine Similarity Example

	cow	pig	dog	cat	log	bug	fox	ape	man	car
x	3	2	0	5	0	0	0	2	0	0
y	1	2	0	5	0	0	0	1	0	2

$$x = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$$
$$y = (1, 2, 0, 0, 0, 0, 0, 1, 0, 2)$$

$$x \cdot y = 3*1 \ldots 2*0 + 0*0 + 5*0 + 0*0 \ldots 2*1 \ldots 0*2 = 5$$

$$||x|| = \mathrm{sqrt}(3*3+2*2\ldots) = 6.48$$
$$||y|| = \mathrm{sqrt}(1*1+0*0\ldots) = 2.24$$

Extended Jaccard aka Tanimoto Coefficient

(reduces to Jaccard for binary attributes)

jaccard() --> compute similarities of binary attributes

tanimoto() --> compute similarities of continuous attributes

Extended Jaccard (Tanimoto Coefficient)

$$EJ(x,y) = \frac{x \cdot y}{||x||^2 + ||y||^2 - x \cdot y}$$

Pearson Correlation

Think: Like Euclidean, but corrects for “grade inflation.”

eg: Movie ratings. Some users consistently give more stars than others. Euclidean is ok, Pearson is better.

Pearson Correlation

For binary/continuous attributes.

Always [-1, 1]

DM 77
CI 11

Example: Movie Recommendations

CI chapter 2