CSCI568

Lecture 7: Similarity, Dissimilarity
September 9, 2009

Hello. I am a computer.

And I have no idea what love, happiness or similarity mean.

Defining Similarity (to a computer)

Similarity between two objects is a numerical measure of the degree to which the two objects are alike.

Dis/Similarity Values

Usually, use ranges `[−1, 1]` or `[0, 1]`.

(But not everyone does, so you may need to transform the similarity score.)

DM 66, 67

Dissimilarity of Single Attribtues

- nominal: it is or it isn’t
- ordinal
  - $d = |x - y| / (n-1)$
  - $s = 1 - d$
- continuous:
  - $d = |x - y|$
  - $s = 1/1+d$ (more, DM69)

Dissimilarity Between Data Objects

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

Proximity Calculation Issues

- attributes w/ different scales
  - (eg, age vs. income)
- heterogeneous attributes
  - (eg, nominal and interval attributes)
- attributes w/ different importance

Example: Movie Recommendations