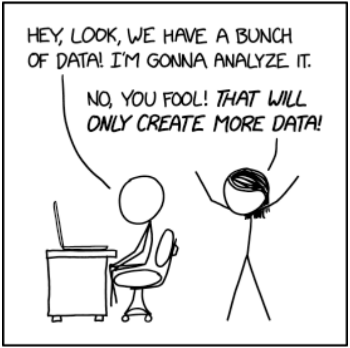


Image. xkcd comics - # 2582

Data Mining: Data

Similarity, Dissimilarity, and Distance

CS 4821 - CS 5831 - s24

Some slides adapted from P. Smyth; A. Moore, D. Klein Han, Kamber, Pei; Tan, Steinbach, Kumar; L. Kaebling; R. Tibshirani; T. Taylor; and L. Hannah

Data Similarity, Dissimilarity and Distance

Similarity, Dissimilarity, and Distance

- For many data mining tasks, we want to be able to measure how alike or unalike two data points are in comparison to one another

Similarity

- numerical measure of how alike are two data objects
- higher when objects are more alike
- often falls in range $[0, 1]$[1]

Dissimilarity / Distances

- numerical measure of how different are two data objects
- lower when objects are more alike
- minimum dissimilarity is often 0, upper limit may vary
- upper limit varies

Defining Distance Measures

What properties should a distance measure have?

A distance (or a metric) on a set S is a function D:
$S \times S \rightarrow [0, +\inf)$ should satisfy

$D(A,B) = D(B,A)$	Symmetry
$D(A,A) = 0$	Constancy, Self-Similarity
$D(A,B) = 0,\ iff\ A = B$	Positivity (Separation)
$D(A,B) \leq D(A,C) + D(B,C)$	Triangle Inequality

Similarity/Dissimilarity for Simple Attributes

p and q are the attribute values for two data objects

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Distance Measure for Numeric Data

- Most common measure for quantitative data is Euclidean distance

$$
\begin{aligned}
d(\mathbf{x}_i, \mathbf{x}_j) &= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \\
&= \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \\
&= \|\mathbf{x}_i - \mathbf{x}_j\|_2
\end{aligned}
$$

- measurements should be commensurate; standardize measurements

Distance Measure for Numeric Data

Minkowski distance - generalization of Euclidean distance calculates the ℓ_λ norm for $\lambda \geq q$:

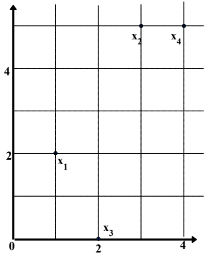
$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_\lambda = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^\lambda \right)^{\frac{1}{\lambda}}$$

Common values of λ

- $\lambda = 2$, Euclidean distance ℓ_2
- $\lambda = 1$, Manhattan distance, ℓ_1
- $\lambda = \infty$, $d(\mathbf{x}_i, \mathbf{x}_j) = \max\limits_{1 \leq k \leq p} |x_{ik} - x_{jk}|$, the sup, or supremum, norm, ℓ_∞

Example of Minkowski Distances

	v1	v2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

	x1	x2	x3	x4
x1	0	5	3	6
x2	5	0	6	1
x3	3	6	0	7
x4	6	1	7	0

Euclidean (L_2)

	x1	x2	x3	x4
x1	0.00	3.61	2.24	4.24
x2	3.61	0.00	5.10	1.00
x3	2.24	5.10	0.00	5.39
x4	4.24	1.00	5.39	0.00

Supremum (L_∞)

	x1	x2	x3	x4
x1	0	3	2	3
x2	3	0	5	1
x3	2	5	0	5
x4	3	1	5	0

Distance Measure for Numeric Data

- **Linear** dependence between variables can be measured by covariance and correlation
- Covariance, Σ, between two variables A, B is

$$Cov(A, B) = \frac{1}{n}\sum_{i=1}^{n}(x_{iA} - \bar{x_A})(x_{iB} - \bar{x_B})$$

- Correlation coefficient

$$\rho(A, B) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{iA} - \bar{x_A})(x_{iB} - \bar{x_B})}{\left(\sum_{i=1}^{n}(x_{iA} - \bar{x_A})^2 \sum_{i=1}^{n}(x_{iB} - \bar{x_B})^2\right)^{\frac{1}{2}}}$$

Other Distance Measures

Cosine Similarity

- Cosine similarity is a commonly used distance measure when dealing with text data
- A document can be represented by thousands of attributes each detailing the frequency of a particular word
- Cosine similarity finds the similarity between documents (vectors), if d_1 and d_2 are vectors then

$$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$$

- Example: find the similarity between two documents:
  $d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
  $d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$ $\qquad cos(d_1, d_2) = 0.94$

Other Distance Measures

- Distance for numeric data
  - Mahalanobis distance
- Distance between binary data
  - Jaccard coefficient
- Distance between strings
  - edit distance
- Distance between images and waveforms
  - shift-invariant, scale-invariant
- Distance between time-series data
  - Euclidean distance, dynamic time-warping
- Other methods
  - Kernel methods