Types of Data
How to Calculate Distance?

Dr. Ryan Benton and Vijay Raghavan

February 4, 2020

Book Information

- Data Mining, Concepts and Techniques
  - Han et al.: Chapter 2, section 4 (2nd Edition, Chapter 7, Section 2, *Types of Data in Cluster Analysis).*
- Advances in Instance-Based Learning Algorithms
  - Dissertation by D. Randall Wilson, August 1997.
  - Chapters 4 and 5.
- Prototype Styles of Generalization
  - Thesis by D. Randall Wilson, August 1994.
  - Chapters 3.

Data

- Each instance (point, record, example, entity, sample, observation)
  - Composed of one or more features.
- Feature (attribute, variable, field, dimension, characteristic)
  - Composed of a data type
  - Data type has a range of values.

Data Types

- Quantitative
  - Interval-Scaled
    - Real
    - Integer
  - Ratio-Scaled
    - Discrete
    - Continuous

Data Types

- Qualitative
  - Binary
    - Symmetric
    - Asymmetric
  - Ordinal
    - Discrete
    - Continuous
- Others
  - Vectors
  - Shape
  - Etc.

Comparing Instances

- How does one compare instances?
  - Clustering
  - Classification
    - Instance-Base Classifiers
    - Artificial Neural Networks
    - Support Vector Machines
- Distance Functions (Measures)

Distance Measures

- Properties
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$

Interval-Scaled Variables

- Many Different Distance Measures
  - Euclidean
  - Manhattan (City Block)
  - Minkowski
- For purpose of discussion, assume all features in data point are Interval-Scaled.

Euclidean

- Also called the L_2 *norm*
- Assumes a straight-line from two points
- $$d(i,j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{in} - x_{jn}\right)^2}$$
- Where
  - i, j are two different instances
  - n is the number of interval-features
  - x_{iz} is the value at z^{th} feature value for X_i.

Manhattan

- Also classed the *L_1 norm*
- Non-Linear.

- $$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}|$$
- Where
  - i, j are two different instances
  - n is the number of interval-features
  - x_{iz} is the value at z^{th} feature value for X_i.

Minkowski

- Euclidean and Manhattan
  - Special Cases
- $$d(i,j) = \left(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{in} - x_{jn}|^p \right)^{1/p}$$
- Where ***p*** is a positive integer
- Also called the L_p norm fuction

Minkowski

- Not all features are equal.
  - Some are irrelevant
  - Some are should be highly influential

- $d(i,j) = \left(w_1 |x_{i1} - x_{j1}|^p + w_2 |x_{i2} - x_{j2}|^p + \ldots + w_n |x_{in} - x_{jn}|^p \right)^{1/p}$
- Where, w_z is the 'weight' of zth feature, where $w_z >= 0$.

Example

- $X_i = (1,2)$, $X_j = (3,5)$

- Euclidean: $d(i,j) = \sqrt{(1-3)^2 + (2-5)^2} = 3.61$
- Manhattan: $d(i,j) = |1-3| + |2-5| = 5$
- Minkowski (p = 3):
  - $d(i,j) = \left(|1-3|^3 + |2-5|^3\right)^{1/3} = (8+27)^{1/3} = 3.27$

Other Distance Measures

- Camberra
- Chebychev
- Quadratic
- Mahalanobis
- Correlation
- Chi-Squared
- Kendall’s Rank Correlation
- And so forth.

Problem

- Feature value ranges may distort results.
- Example:
  - Feature 1: [0, 2]
  - Feature 2: [-2, 2]
- Changes in feature 2, in the distance functions, has greater impact.

Scaling

- Scale each feature to a range
  - [0,1]
  - [-1, 1]
- Possible Issue
  - Say feature range is [0, 2].
  - 99% of the data >= 1.5
    - Outliers have large impact on distance
    - Normal values have almost none.

Normalize

- Modify each feature so
  - Mean (m_f) = 0
  - Standard Deviation (σ_f) = 1
- $$y_{if} = \frac{x_{if} - m_f}{\sigma_f}, \quad \sigma_f = \frac{1}{N}\sqrt{\left|x_{1f} - m_f\right|^2 + \left|x_{2f} - m_f\right|^2 + \ldots + \left|x_{Nf} - m_f\right|^2}$$
- where
  - y_{if} is the new feature value
  - N is the number of data points.

Z-Score

- $z_{if} = \frac{x_{if} - m_f}{s_f}$
- $s_f = \frac{1}{N}\left(\left|x_{1f} - m_f\right| + \left|x_{2f} - m_f\right| + \ldots + \left|x_{Nf} - m_f\right|\right)$
- where
  - z_f is the z-score
  - s_f is the mean absolute deviation
    - More robust to outliers, compared to standard deviation.

Symmetric Binary

- Assume, for now, all features are symmetric binary.
- How to compare?
  - Can use Euclidean, Manhattan, or Minkowski functions.
  - Symmetric binary similarity

Symmetric Binary

Object j / Object i	1	0	sum
1	q	r	q + r
0	s	t	s + t
sum	q + s	r + t	p

- q, r, s and t are counts.

Symmetric Binary

- $$d(i,j) = \frac{r+s}{p}$$
- Properties
  - Range is [0, 1]
  - 0 indicates perfect match
  - 1 indicates no matches

Asymmetric Binary

- Assume, for now, all features are asymmetric binary.

- Like Symmetric Binary
  - Can use Euclidean, Manhattan, or Minkowski functions.
- Alternatively, can use
  - Asymmetric binary similarity

Asymmetric Binary

Object i \ Object j	1	0	sum
1	q	r	q + r
0	s	t	s + t
sum	q + s	r + t	p

- q, r, s and t are counts.

Asymmetric Binary

- $$d(i,j) = \frac{r+s}{q+r+s}$$
- Properties
  - Range is [0, 1]
  - 0 indicates perfect match
  - 1 indicates no matches
  - Note, as (0==0) is considered unimportant, it is not factored in.

Examples

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	Y	N	P	N	N	N
Mary	Y	N	P	N	P	N

- Set
  - Y and P to 1
  - N to 0
- Symmetric
  - d(Jack, Mary) = (0 + 1) / 6 = 0.167
- Asymmetric
  - d(Jack, Mary) = (0 + 1) / (2 + 0 + 1) = 0.33

Categorical

- $$d(i,j) = \frac{p - m}{p}$$
- Where
  - p = number of variable
  - m = number of matches

Example

Student	Test-1 (categorical)	Test-2 (ordinal)	Test-3 (ratio)
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code-C	Good	164
4	Code-A	Excellent	1,210

- $d(2, 1) = (1 – 0) / 1 = 1$
- $d(1, 4) = (1 – 1) / 1 = 0$

Categorical

- Weighting
  - Can add weights to
    - Increase effect of m
    - Increase importance of variables with more states
      - Can do this for Binary as well.
  - Convention
    - Some of weights should be equal to 1.

Categorical – Other measures

- Value Difference Metric
  - For Classification problems (not Clustering).
  - Estimates conditional probabilities for each feature value for each class.
  - Distance is based on difference in conditional probabilities.
  - Includes a weighting scheme.
- Modified Value Difference Metric
  - Handles weight estimation differently.

Value Difference Metric (VDM)

- $$d(i,j) = \sum_{f=1}^{n} \sum_{g=1}^{C} \left(P(x_{if}, g) - P(x_{jf}, g) \right)^{q}$$
- Where
  - $P(x_{if},g)$ = conditional probability of the class g occuring, given the value x_i for feature f.
  - C is the number of classes
  - n is the number of features
  - q is either 1 or 2.
- Note, for simplification, weights are not included.

Ordinal

- Assume all Features are Ordinal.
- Feature f has M_f ordered states, representing ranking 1, 2, …, M_f.
- For each instance i
  - For each feature f
    - Replace value x_{if} by corresponding rank r_{if}
    - $r_{if} \in [1,\ldots,M_f]$
- To calculate d(i,j)
  - Use Interval-Scaled Distance Functions.

Ordinal

- Like Interval-Scaled
  - Different Ordinal features may have different number of states.
  - This leads to different features having different implicit weights.
  - Hence, scaling necessary.
- $$y_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Example

Student	Test-1 (categorical)	Test-2 (ordinal)	Test-3 (ratio)
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code-C	Good	164
4	Code-A	Excellent	1,210

- Mappings
  - Fair = 1, Good = 2, Excellent = 3
- Normalized Values
  - Fair = 0.0, Good = 0.5, Excellent = 1.0

Example

Student	Test-1 (categorical)	Test-2 (ordinal)	Test-3 (ratio)
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code-C	Good	164
4	Code-A	Excellent	1,210

- Euclidean: $d(2,3) = \sqrt{(0-0.5)^2} = 0.5$

Ordinal – Other Measures

- Hamming Distance
- Absolute Difference
- Normalized Absolute Difference
- Normalized Hamming Distance

Ratio-Scaled

- Can’t treat directly as Interval-Scaled
  - The scale for Ratio-Scaled would lead to distortion of results.
- Apply
  - a logarithmic transformation first.
    - $y_{if} = \log(x_{if})$
  - Other type of transformation.
- Treat result as continuous Ordinal Data.

Example

Student	Test-1 (categorical)	Test-2 (ordinal)	Test-3 (ratio)	Test-3 (logarithmic)
1	Code-A	Excellent	445	2.68
2	Code-B	Fair	22	1.34
3	Code-C	Good	164	2.21
4	Code-A	Excellent	1,210	3.08

- Euclidean: $d(4,3) = \sqrt{(3.08 - 2.21)^2} = 0.87$

Mixed Types

- The above approaches assumed that all features are the same type!
- This is rarely the case.
- Need a distance function that handles all types.

Mixed Distance

- $$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{f} d_{ij}^{f}}{\sum_{f=1}^{p} \delta_{ij}^{f}}$$
- Where
  - δ_{ij}, for feature f is
    - 0
      - If either x_{if} or x_{jf} is missing
      - (x_{if} == x_{jf} == 0) and f is asymmetric binary
    - Else 1

Mixed Distance

- Where
  - If feature f is
    - Interval-scaled, use this formula
      - $$d_{ij}^{f} = \frac{\left| x_{if} - x_{jf} \right|}{\max_h x_{hf} - \min_h x_{hf}}$$
      - Where h runs over non-missing values for feature f.
    - Ensures distance returned is in range [0,1].

Mixed Distance

- Where
  - If feature f is
    - Binary or categorical
      - If $x_{if} == x_{jf}$, $d_{ij} = 0$
      - Else, $d_{ij} = 1$
    - Ordinal
      - Compute ranks and apply the ordinal scaling
      - Then use the interval-scaled distance measure.

Mixed Distance

- Where
  - If feature f is
    - Ratio-Scaled
      - Do logarithmic (or similar) transform and then apply interval-scaled distance.
      - Or, treat as ordinal data.

Mixed Distance

- $$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{f} d_{ij}^{f}}{\sum_{f=1}^{p} \delta_{ij}^{f}}$$
- Distance calculation for each feature will be 0 to 1.
- Final distance calculation will be [0.0, 1.0]

Example

Student	Test-1 (categorical)	Test-2 (ordinal)	Test-3 (ratio)	Test-3 (logarithmic)
1	Code-A	Excellent	445	2.68
2	Code-B	Fair	22	1.34
3	Code-C	Good	164	2.21
4	Code-A	Excellent	1,210	3.08

$$d(2,1) = \frac{1(1) + 1\left(\frac{|0-1|}{1-0}\right) + \left(\frac{|1.34-2.68|}{3.08-1.34}\right)}{3} = 0.92$$

Mixed Distance

- Problems
  - Doesn’t permit use, for interval-scaled, more advanced distance functions.
  - Binary and categorical values have more potential impact than other types of features.

Mixed Distance

- Minkowski
- Heterogeneous Overlap-Euclidean Metric
- Heterogeneous Value Difference Metric
- Interpolated Value Difference Metric
- Windowed Value Difference Metric
- K*
  - Violates some of the conditions for distance measure.

- Not a complete list.

Questions?