Types of Data

How to calculate distance?

Satya Katragadda
January 25, 2016

Books

- Data Mining, Concepts and Techniques. Chapter 2, Sections 1,2,4. 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.

What is Data?

- Collection of data objects/instances and their attributes/features.
- Object is known as a record, point, case, sample, instance, or entity
- An attribute is a property or characteristic of an object
  - Attribute is known as variable, field, characteristic, or feature.
  - Attribute is composed of a data type and has a range of values
  - Examples: temperature, price of an item etc.
- In context of database, rows -> data objects; columns -> attributes.

Types of Data

- Nominal
  - ID's, colors etc.
- Binary
  - Gender
- Ordinal
  - grades, rankings
- Numeric: quantitative
  - Interval-scaled
    - calendar dates, body temperatures
  - Ratio-scaled
    - length, time

Properties of Attribute Values

- Type of an attribute depends on which of the following properties it posses
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /
- Nominal: distinctness
- Ordinal: distinctness & order
- Interval-scaled: distinctness, order & addition
- Ratio- scaled: All 4 attributes

Attribute Types

- Nominal: categories, states or “names of things”
  - Marital status = {single, married, divorced}
  - Occupation, zip codes etc.
- Binary
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes are equally important, e.g., gender
  - Asymmetric binary: All outcomes are not equally important
    - Medical test (positive vs. negative), assign 1 to important outcome

Attribute Types

- Ordinal
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}
  - grades, rankings

Numeric Attribute Types

- Quantity (integer or real-m values)
- Interval
  - Measured on scale of equal-sized units
  - Values have order, temperature in centigrade
  - No true zero-point
- Ratio
  - Inherent zero-point
  - Values are presented in a order of magnitude (4 lb. is twice as heavy as 2lb.)
  - e.g., temperature in kelvin, length

Types of Attributes - Summary

	Nominal	Ordinal	Interval	Ratio
Frequency distribution/ counts	YES	YES	YES	YES
Mode/median	NO	YES	YES	YES
Add, Subtract, mean, standard deviation and mean	NO	NO	YES	YES
Ratio/coefficient of variation Has "true zero"	NO	NO	NO	YES

Discrete vs. Continuous Attributes

- Discrete Attribute
  - Has finite set of values, e.g., zip codes, set of words in a corpus
  - Can be represented as a integer
  - Binary attributes as a special case of discrete attributes
- Continuous Attributes
  - Has real numbers as attributes, e.g., temperature, height, weight
  - Practically, real values can only be measured and represented using finite number of digits

Comparing Instances

- How does one compare instances?
  - Clustering
  - Classification
    - Instance based classifiers
    - Artificial neural networks
    - Support vector machines
- Distance Functions

Distance Measures

- Many different distance measures
  - Euclidean
  - Manhattan
  - Minkowski
- Assume all features in data point are interval- scaled

Distance Measures - Euclidean

- Also called L_2 *norm*
- Assumes a straight-line from two points
- $d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$
- 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 i

Distance Measures - Manhattan

- Also called L_1 *norm*
- Non-Linear
- $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$
- 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 i

Distance Measures - Minkowski

- Generalized distance measure
- Also called L_1 *norm*
- $d(i,j) = \left(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p\right)^{1/p}$
- Where P is a positive integer
- Euclidean and Manhattan are special cases where p=1,2

Distance Measures - Minkowski

- Not all features are equal
  - Some are relevant
  - Some are highly influential
- $d(i,j) = \left(w_1|x_{i1} - x_{j1}|^p + w_2|x_{i2} - x_{j2}|^p + \cdots + w_n|x_{in} - x_{jn}|^p\right)^{1/p}$
- Where, W_z is the 'weight'
  - $W_z > 0$

Distance Measures - Example

- x_1= (1,2,3), x_2 = (3,5,7)
- Euclidean: $d(x_1, x_2) = \sqrt{(1-3)^2+(2-5)^2+(3-7)^2}$ = 5.385
- Manhattan: $d(x_1, x_2) = |1-3| + |2-5| + |3-7| = 11$
- Minkowski (p=3):

$$d(x_1, x_2) = (|1-3|^3 + |2-5|^3 + |3-7|^3)^{1/3} = 4.30886$$

Distance Measures

- Camberra
- Chebychev
- Quadratic
- Mahalanobis
- Correlation
- Chi-Sqaured
- Kendall's Rank Correlation
- And so forth

Distance Measure - Problems

- 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

Distance Measures - 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

Distance Measures - Normalize

- Modify each feature such that
  - Mean(m_f) =0, Standard Deviation (σ_f) = 1
- $y_{if} = \frac{x_{if} - m_f}{\sigma_f}$, $\sigma_f = \frac{\sqrt{|x_{1f} - m_f|^2 + |x_{2f} - m_f|^2 + \cdots + |x_{1f} - m_f|^2}}{N}$
- Where
  - y_{if} is the new feature value
  - N is the number of data points
- Z-score, use absolute deviation instead of standard deviation

Distance Measures – Binary Data

- How to compare binary variables?
  - Can we use Euclidean, Manhattan and Minkowski functions
  - Are all symmetric measures same?

Distance Measures – Binary Data

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

- Symmetric binary variables:
  - Both states are equally valuble and carry same weight
  - $d(i,j) = \frac{r+s}{q+r+s+t}$
- Asymmetric binary variables:
  - One state is more important than the other
  - $d(i,j) = \frac{r+s}{q+r+s}$

Dissimilarity – Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- All other attributes are asymmetric
- Set Y , P = 1 and N =0
- D(*Jack, Mary*) = $\frac{0+1}{2+0+1}$ = 0.33

Dissimilarity – Categorical

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

Dissimilarity – Categorical

- Example

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

- $d(2,1) = \frac{1-0}{1} = 1$
- $d(1,4) = \frac{1-1}{1} = 0$

Dissimilarity - Ordinal

- Identify the rank of variables
- Treat variables like interval scaled variables
  - Replace x_{if} by their rank
  - Map the range of each variable onto [0,1] by replacing i^{th} object in the f^{th} variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Compute the dissimilarity using methods for interval scaled variables

Dissimilarity - Ordinal

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

- Example
- Mappings
  - Fair = 1, Good = 2, Excellent =3
- Normalized values
  - Fair = 0.0, Good = 0.5, Excellent = 1.0
- Euclidean: $d(2,3) = \sqrt{(0-0.5)^2} = 0.5$

Dissimilarity – Ratio-Scaled

- Cant treat directly as interval-scaled
  - Scale of ratio-scaled would lead to distortion of results
- Eliminate distortions by applying
  - logarithmic transformations $y_{if} = \log x_{if}$
  - Other type of transformations
- Treat results as continuous ordinal data

Dissimilarity – Ratio-Scaled

- Example
- Convert ratio scaled to

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

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

Dissimilarity – Mixed distance

- All the above examples assume all features are all the same type
- This scenario is rarely true
- Need a distance function that handles all kinds of data
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal

Dissimilarity – Mixed distance

- Use a weighted formula to combine their effects

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

- Where
  - $\delta_{ij}^{(f)}$ 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

Dissimilarity – Mixed distance

- f is numeric: use normalized distance
- f is ordinal:
  - compute rank r_{if}
  - Treat the feature as interval-scaled value

$$d_{i,j}^{f} = \frac{|x_i^f - x_j^f|}{max^f - min^f}$$

Dissimilarity – Mixed distance

- Example

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

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

Questions?

- Email: satya@louisiana.edu