Intro to Data Mining Lecture 2a

Measures of Distance

Created by Jon Witkowski on 12/29/2023
using some slides from Dr. Breitzman

Types of Data

- Last week we went over several different types of data, including
  - Nominal
    - Symmetric Binary
    - Asymmetric Binary
  - Ordinal
  - Numeric
- We also very briefly discussed term/frequency data. We'll go over this too today.

Term Data

- Remember this table from last week? This is term data.
- Each row is a vector, so we're going to use linear algebra to compute the distance
- Remember what vectors are?
- Remember dot products and norms?
- Those will be used to calculate the cosine of the angle between vectors, which is how we're going to determine the similarity

	asparagus	beans	broccoli	corn	peppers	squash	tomatoes
1	0	1	1	1	1	0	1
2	0	0	1	1	1	0	0
3	1	0	0	1	0	1	0
4	0	1	0	1	0	1	1
5	0	1	0	1	1	0	1
6	1	1	1	0	0	0	0
7	1	1	0	0	0	1	1
8	0	0	0	1	0	0	1
9	0	0	1	0	1	0	1
10	1	1	0	0	0	1	0
11	0	1	0	1	0	0	0
12	0	1	1	0	1	1	0
13	1	1	0	0	0	1	0
14	1	1	0	1	0	1	0
Sum	6	10	5	8	5	7	6

Term Data

- Let’s think of the first 2 rows as

  <0,1,1,1,1,0,1> and <0,0,1,1,1,0,0>

- The dot product is just the sum of the products of each component

  $$\sum_{i=0}^{n} x_i * y_i$$ for vectors x and y

- The norm is going to be

  $$\sqrt{\sum_{i=0}^{n} x_i^2}$$ for vector x

	asparagus	beans	broccoli	corn	peppers	squash	tomatoes
1	0	1	1	1	1	0	1
2	0	0	1	1	1	0	0
3	1	0	0	1	0	1	0
4	0	1	0	1	0	1	1
5	0	1	0	1	1	0	1
6	1	1	1	0	0	0	0
7	1	1	0	0	0	1	1
8	0	0	0	1	0	0	1
9	0	0	1	0	1	0	1
10	1	1	0	0	0	1	0
11	0	1	0	1	0	0	0
12	0	1	1	0	1	1	0
13	1	1	0	0	0	1	0
14	1	1	0	1	0	1	0

Term Data

- Let’s think of the first 2 rows as

  <0,1,1,1,1,0,1> and <0,0,1,1,1,0,0>

- Given our formulas on the last slide, the dot product of rows 1 and 2 will be 3 and the norms will respectively be 5 and 3.
- Now, those will be used to calculate the cosine similarity, which can be found as the dot product divided by the product of the norms. In this case, it will be 0.2

	asparagus	beans	broccoli	corn	peppers	squash	tomatoes
1	0	1	1	1	1	0	1
2	0	0	1	1	1	0	0
3	1	0	0	1	0	1	0
4	0	1	0	1	0	1	1
5	0	1	0	1	1	0	1
6	1	1	1	0	0	0	0
7	1	1	0	0	0	1	1
8	0	0	0	1	0	0	1
9	0	0	1	0	1	0	1
10	1	1	0	0	0	1	0
11	0	1	0	1	0	0	0
12	0	1	1	0	1	1	0
13	1	1	0	0	0	1	0
14	1	1	0	1	0	1	0

Exercise

- Suppose we have 2 vectors <1, 0, 0, 2, 1> and <-1, 0, 0, -2, -1>
- What do you think the cosine similarity will be? Why?

- What about for <1, 0, 0, 1, 1> and <0, 1, 1, 0, 0>?
- What about for <1, 3, 2, 0, 1> and <2, 6, 4, 0, 2>?

- Let’s do the math and see

Distances for Other Types of Data

- **Term data, Frequency Data, Vector Data**
  - E.g. Movies, Terms in a document, Products bought at amazon, etc.
  - Use Cosine Similarity
- **Asymmetric Binary Data** – Data with 2 values, but positive value is of high importance, negative is ignored
  - E.g. Has fever, has Diabetes, tested positive for something
  - Use Jaccard
- **Symmetric Binary Data** – Data with 2 values both of equal importance
  - E.g M/F, Smoker/non-Smoker, Like/Didn't Like
- **Categorical (Nominal) Data** – Data that can take on multiple states but values have no meaningful order
  - E.g. Red, Yellow, Blue, Green; Code X, Code Y, Code Z;
  - Distance: match=1, non-match=0
- **Ordinal Data** – Attributes that have a meaningful order
  - Poor, Fair, Good, Excellent; Freezing, Cold, Mild, Warm, Hot
  - Normalize then Euclidean distance
- **Numeric or Continuous Data** – Regular data like 10, 20, 30, etc.
  - Normalize then Euclidean

Asymmetric Binary Data Example

	Cough	Fever	Rapid Heartbeat
Susan	Y	Y	N
Joe	Y	N	N
Jim	Y	Y	N
Jane	N	N	Y

- Data with 2 values, but positive value is of high importance, negative is ignored
- D(Susan,Joe)=1/2; Sim(Susan,Joe)=1/2
- D(Susan,Jim)=0; Sim(Susan,Jim)=1
- D(Susan,Jane)=(2+1)/3=1 Sim(Susan,Jane)=0
- This distance measure is widely known as the jaccard distance
- It can be used for term frequencies as well, but cosine is a bit easier to implement weighting factors like IDF (inverse document frequency)

Symmetric Binary Data

	Cough	Fever	Rapid Heartbeat
Susan	Y	Y	N
Joe	Y	N	N
Jim	Y	Y	N
Jane	N	N	Y

- Symmetric Binary Distance:
  d(A,B)= Number of Differences/Number of variables

- Note Sim(A,B)=1-d(A,B), also called Simple Matching Coefficient

- D(Susan,Joe)=1/3; Sim(Susan,Joe)=2/3
- D(Susan,Jim)=0; Sim(Susan,Joe)=1
- D(Susan,Jane)=3/3=1 Sim(Susan,Jane)=0

Categorical Data Example

	Color	Code	vehicle
Susan	Blue	X	Sedan
Joe	Blue	Y	Sedan
Jim	Green	X	Truck
Jane	Yellow	Z	Coupe

- Similarity=Matches/Variables, SMC
- Sim(Susan,Susan)=3/3=1; D(Susan,Susan)=0
- Sim(Susan,Joe)=2/3; D(Susan,Joe)=1/3
- Sim(Susan,Jim)=1/3; D(Susan,Jim)=2/3
- Sim(Susan,Jane)=0; D(Susan,Jane)=1

Ordinal and Numerical Data Example

	Grades	Weather Preference	Family	Age
Susan	A's	Cold	Small	20
Joe	B's	Freezing	Large	30
Jim	C's	Hot	Small	50
Jane	F's	Warm	Large	34

- Convert Ordinal Data to [0,1] by (Rank-1)/(Max-1)
- For example Max Rank=A=5,Rank(F)=1
  F=0, D=.25, C=.5, B=.75, A=1
- Freezing=0, Cold=.33, Warm=.66, Hot=1
- Small=0, Large=1
- For age use min-max normalization (x-min)/(max-min)
  20→0, 30→(30-20)/(50-20)=1/3, 50→1, 34→14/30=7/15

Transformed Table

	Grades	Weather Preference	Family	Age
Susan	1	0.33	0	0
Joe	0.75	0	0.5	0.33
Jim	0.5	1	0	1
Jane	0	0.66	0.5	0.47

- Normalized Weighted Euclidean Distance

$$d(i,j) = \frac{\sqrt{w_1(x_{1i}-x_{1j})^2 + w_2(x_{2i}-x_{2j})^2 + \ldots + w_n(x_{ni}-x_{nj})^2}}{\sqrt{w_1 + \ldots + w_n}}$$

- Regular Euclidean Distance is the same except we let each w=1
- It should be clear that distance=0 for identical items and 1 for complete opposites (provided the data is transformed to be in [0,1])
- Note weights can be added to the other distance measures as well

What do we do with mixed data?

Name (Identifier)	Gender (Symmetric Binary)	Favorite Color (Nominal)	Blood Type (Nominal)	General Health (ordinal)	Test1 (numeric)	Cough (asymmetric binary)	High Blood Pressure (asymmetric binary)
Susan	F	Blue	O-	excellent	75	N	N
Jim	M	Red	O+	good	65	N	N
Joe	M	Red	AB-	fair	64	N	Y
Jane	F	Green	A+	poor	83	Y	Y
Sam	M	Blue	A-	good	71	N	N
Michelle	F	Blue	O-	good	90	N	N

- Normalize the data and Compute Similarity Matrices for each type (5 types in this case)
- Each Matrix is the Same Dimension (6x6 in this case)
- Add Each Matrix together (first make them all similarity or dis-similarity matrices if they are not all the same type)
- Divide each entry by 6 so that final similarities (or dis-similarities are between 0 and 1)
- Actually, if we want to weight each variable equally multiply the Nominal and Asymmetric matrices by 2 before adding and then divide by 7
- Matrix should be symmetric with diagonals = 1 for similarity or 0 for dis-similarity

Exercise

- Let’s compute the distance matrix for the table on the previous slide. I think it’s good practice.

Name (Identifier)	Gender (Symmetric Binary)	Favorite Color (Nominal)	Blood Type (Nominal)	General Health (ordinal)	Test1 (numeric)	Cough (asymmetric binary)	High Blood Pressure (asymmetric binary)
Susan	F	Blue	O-	excellent	75	N	N
Jim	M	Red	O+	good	65	N	N
Joe	M	Red	AB-	fair	64	N	Y
Jane	F	Green	A+	poor	83	Y	Y
Sam	M	Blue	A-	good	71	N	N
Michelle	F	Blue	O-	good	90	N	N

Correlation

- We'll talk about this here because I don't know where else to put it
- 2 variables are correlated if there is a linear relationship
    - For example if I have one variable that is measured in degrees Fahrenheit and another that is measured in degrees Celsius, the 2 variables will be correlated

- Here's some scary looking definitions

- Correlation is frequently used to measure the linear relationship between two variables that are observed together

$$\mathrm{corr}(\mathbf{x},\mathbf{y}) = \frac{\mathrm{covariance}(\mathbf{x},\mathbf{y})}{\mathrm{standard\_deviation}(\mathbf{x}) * \mathrm{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x\, s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\mathrm{covariance}(\mathbf{x},\mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}) \quad (2.12$$

$$\mathrm{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x})^2}$$

$$\mathrm{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\overline{y})^2}$$

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n}x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n}\sum_{k=1}^{n}y_k \text{ is the mean of } \mathbf{y}$$