Similarity:

Beyond numerical Data

Mining Massive Datasets

Prof. Carlos Castillo — https://chato.cl/teach

Universitat Pompeu Fabra Barcelona

Main Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapter 3) + slides by Lijun Zhang
- Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al. (Section 2.4)
- Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)
- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al. (Chapter 3)

Categorical and mixed data

Simple similarity for categorical data

- Given $\overline{X} = (x_1, \ldots, x_d); \overline{Y} = (y_1, \ldots, y_d)$
- Compute similarity as

$$\text{sim}(\overline{X}, \overline{Y}) = \sum_{i=1}^{d} S(x_i, y_i)$$

- Simple coordinate wise similarity

$$S(x_i, y_i) = \begin{cases} 1, & \text{if } x_i = y_i \\ 0, & \text{otherwise} \end{cases}$$

Weighing feature values by how rare they are

- Compute similarity as $$\text{sim}(\overline{X}, \overline{Y}) = \sum_{i=1}^{d} S(x_i, y_i)$$

- Inverse occurrence frequency

$p_i(z)$ is the probability that feature i takes value z

$$S(x_i, y_i) = \begin{cases} 1/p_i(x_i)^2, & \text{if } x_i = y_i \\ 0, & \text{otherwise} \end{cases}$$

$$S(x_i, y_i) = \begin{cases} 1 - p_i(x_i), & \text{if } x_i = y_i \\ 0, & \text{otherwise} \end{cases}$$

Goodall measure

Mixture of quantitative and categorical data

- Given $\overline{X} = (\overline{X_c}, \overline{X_n}); \overline{Y} = (\overline{Y_c}, \overline{Y_n});$
- Where c denotes the subset of categorical data and n the subset of numerical data

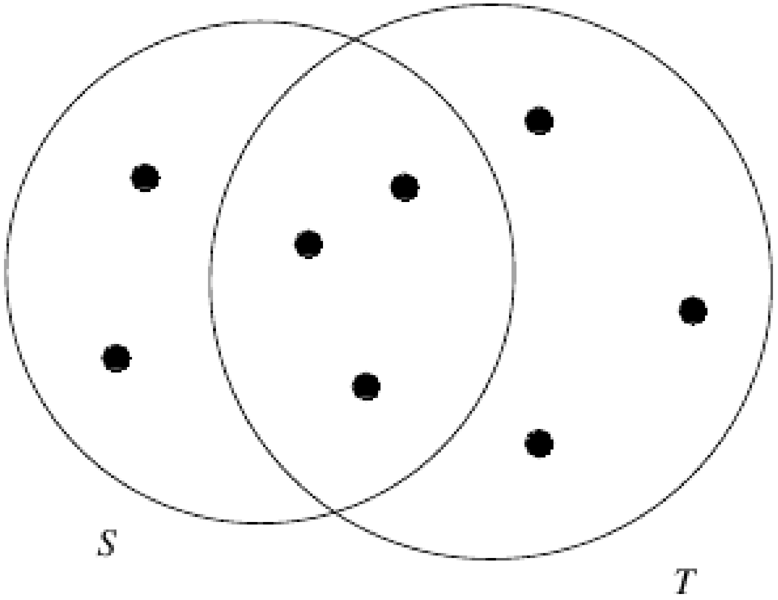
$$\mathrm{sim}(\overline{X}, \overline{Y}) = \lambda\,\mathrm{CatSim}(\overline{X_c}, \overline{Y_c}) + (1-\lambda)\,\mathrm{NumSim}(\overline{X_n}, \overline{Y_n})$$

- In general λ is difficult to set, and additionally we should have variables with similar variances or normalize by variance

Binary and set data

Jaccard coefficient

Example: J(S,T) = 3/8



$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

Binary variables can be set as set inclusion variables

- If $\overline{X}=(x_1, \ldots, x_d)$ is such that $x_i = 1$, this can be seen as element $\overline{X}$ belonging to set i
- Alternatively, $\overline{X}$ can be seen as $S_{\overline{X}}$ the set of all variables i such that $x_i = 1$
- Extended Jaccard coefficient (Tanimoto)

$$J(\overline{X},\overline{Y}) = \frac{\sum_{i=1}^{d} x_i \cdot y_i}{\sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d} y_i^2 - \sum_{i=1}^{d} x_i \cdot y_i}$$

Exercise

Answer in
Nearpod open-ended question
Code to be given in class

- Compute Tanimoto and Jaccard* coefficient between:

  $(0, 2, 1, 0, 3)$

  $(1, 2, 0, 0, 0)$

  * For the Jaccard coefficient, binarize the vectors

$$J(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} x_i \cdot y_i}{\sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d} y_i^2 - \sum_{i=1}^{d} x_i \cdot y_i}$$

Similarity != Distance

- When using Jaccard:
  - The similarity between an object and itself is 1.0
  - The distance between an object and itself is 0.0
- Hence:
  - Jaccard **similarity** = Jaccard_coefficient
  - Jaccard **distance** = 1 - Jaccard_coefficient

Text data

Text documents as vectors: L_p norms

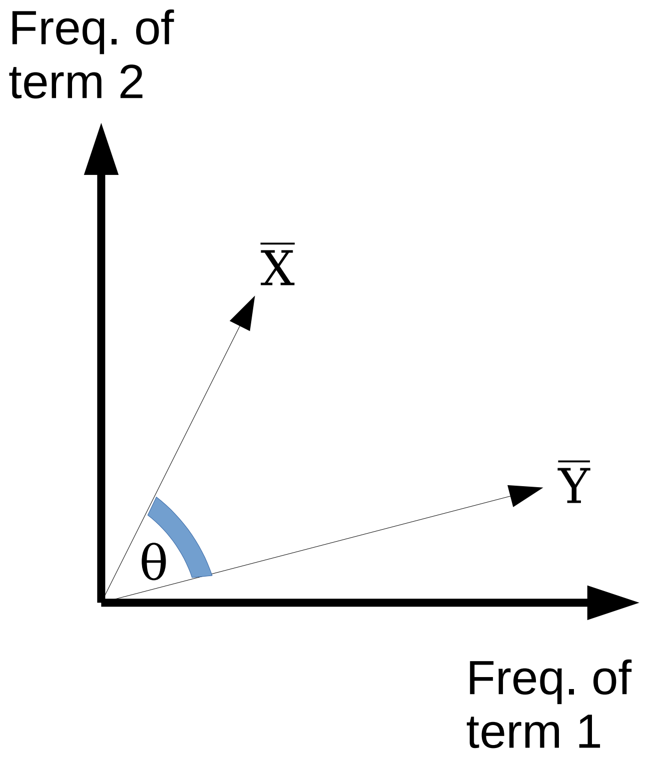
- As Quantitative Multidimensional Data
  - Bag of words model
  - They are very sparse
  - L_p norm does not work well
  - Long documents have long distance
- Dimensionality Reduction (A Possible Solution)
  - Latent Semantic Analysis (equivalent to SVD)
  - L_p norm in the new space

Text documents as vectors: angles

- What we care about is the relative frequency of terms

$$\mathrm{sim}(\overline{X}, \overline{Y}) = \cos\theta$$

$$\mathrm{sim}(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}}$$



However, some terms are very common and others are very rare …

Text documents as vectors: tf-idf weighting (idf)

Skipped in 2021

- $\mathrm{idf}(t) = \log \frac{n}{n_t}$
  - Global inverse document frequency of term t
  - Where n_t is the number of documents where term t appears, n is the total number of documents
- Typical variation (i.e. Okapi BM25)

$$\mathrm{idf}(t) = \log \frac{n - n_t + 0.5}{n_t + 0.5}$$

Text documents as vectors: tf-idf weighting (tf)

Skipped in 2021

- $tf(x_i)$
  - Frequency in a document of term x_i
  - Log frequency, square root of frequency, or similar to reduce the impact of terms of very high frequency

Text documents as vectors: tf-idf weighting (cont.)

Skipped in 2021

- $h(x_i) = tf(x_i) \times idf(x_i)$

$$\mathrm{sim}(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} h(x_i) \cdot h(y_i)}{\sqrt{\sum_{i=1}^{d} h(x_i)^2} \cdot \sqrt{\sum_{i=1}^{d} h(y_i)^2}}$$

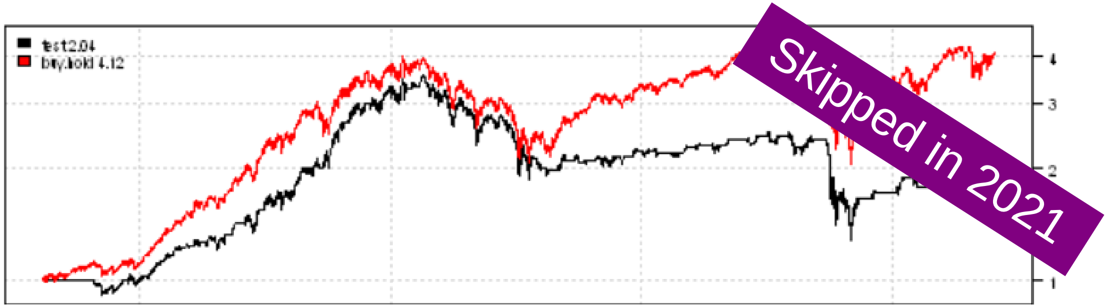
- Or Tanimoto like:

$$J(\overline{X}, Y) = \frac{\sum_{i=1}^{d} h(x_i) \cdot h(y_i)}{\sum_{i=1}^{d} h(x_i)^2 + \sum_{i=1}^{d} h(y_i)^2 - \sum_{i=1}^{d} h(x_i) \cdot h(y_i)}$$

Skipped in 2021

Continuous time series data

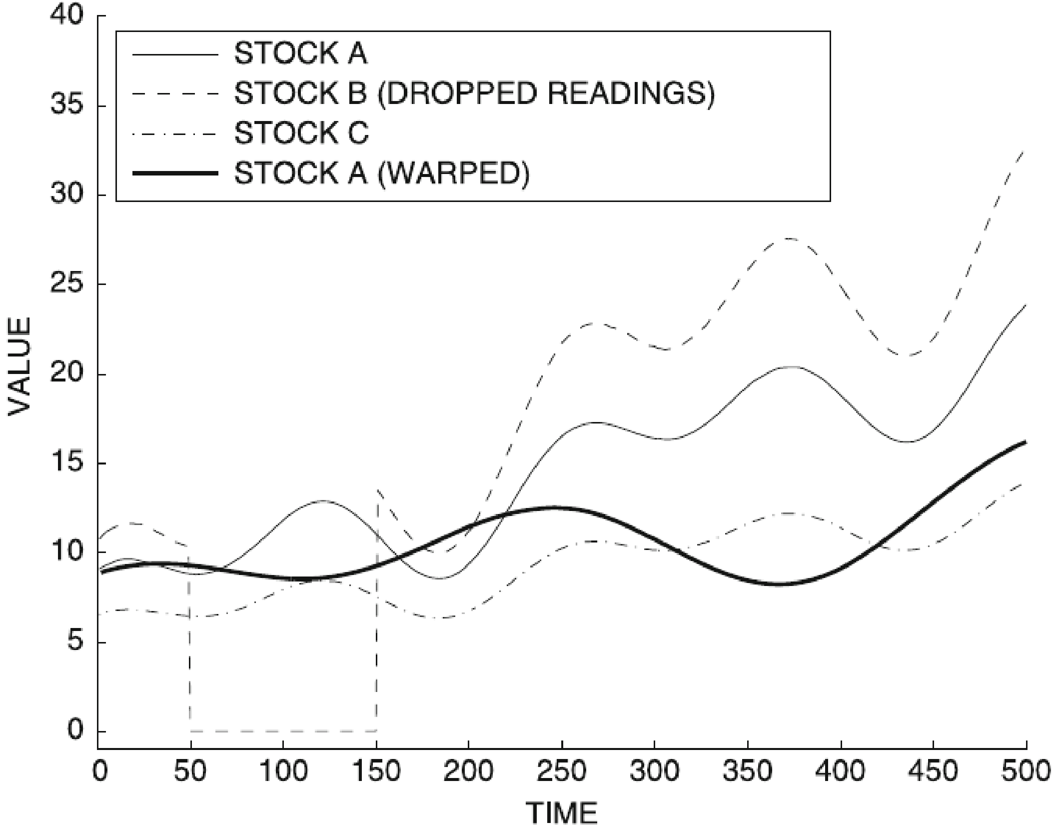
Misalignment between series



- Behavioral attributes
  - Scaling (range is larger or narrower)
  - Translation (series is shifted up or down)
- Contextual attribute (typically, time)
  - Scaling (time is stretched or compressed)
  - Translation or shift (starting time changes)
- Matches might not be contiguous (noisy segments)

Example of scaling, translation, noise

Skipped in 2021



More on this later in the course, in the sequence mining topic

Data Mining, The Textbook (2015) by Charu Aggarwal (Chapter 3).

Warping means stretching or compressing time.

Skipped in 2021

Discrete sequence data

Discrete sequences can be treated as strings

Skipped in 2021

- Compute edit distance
- Compute longest common sub-sequence
- In genetic sequences, use PAM (*Point Accepted Mutation*) matrices
  - Indicate rarity (cost) of replacement

Skipped in 2021

Example PAM matrix

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg	R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn	N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp	D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys	C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln	Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu	E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly	G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His	H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile	I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu	L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys	K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met	M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe	F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro	P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser	S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr	T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp	W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr	Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val	V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

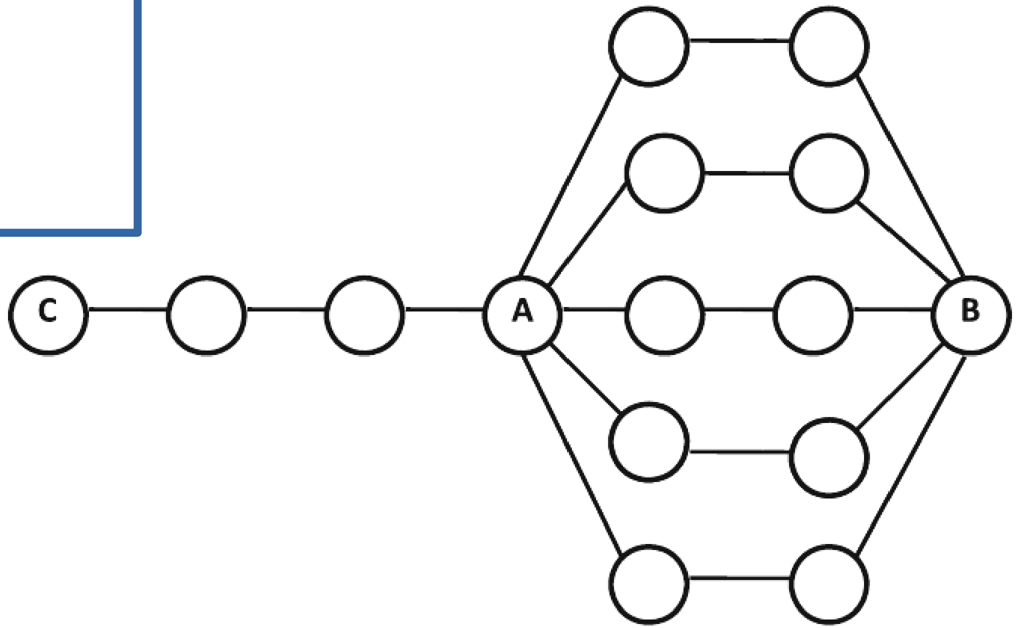
Skipped in 2021

Graph data

Distance/similarity in graph data

Skipped in 2021

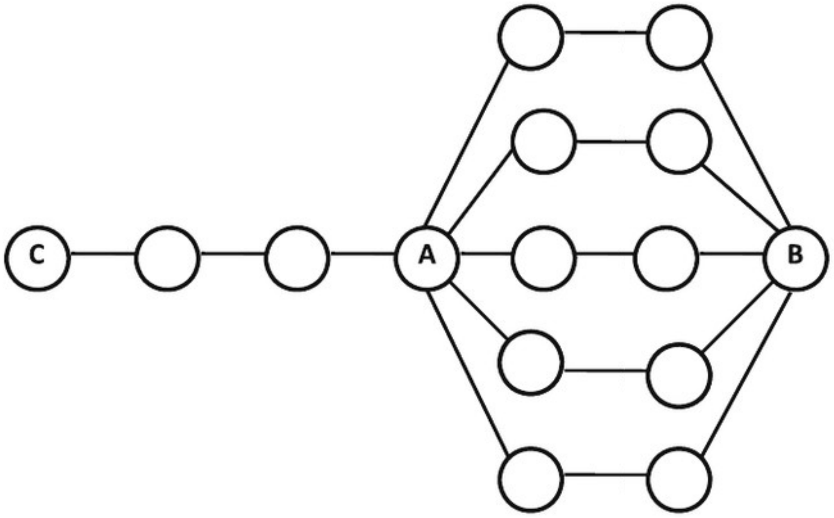
- Comparing A-B and A-C?
  - A-B should be closer
  - A-C should be closer
  - Both should be equal



Distance/similarity in graph data

Skipped in 2021

- Distance-Based Measure
  - Shortest-path on the graph
  - Dijkstra algorithm
- Random Walk-Based Similarity
  - (e.g. personalized PageRank)
  - Accounts for multiplicity in paths during similarity computation



Under random walk similarity, A-B are closer than A-C

Supervised similarity functions

Learning a distance function through supervised ML

- Suppose you have data from experts, annotators, or user feedback:

$$\mathcal{S} = \{O_i, O_j : O_i \text{ is similar to } O_j\}$$

$$\mathcal{D} = \{O_i, O_j : O_i \text{ is dissimilar to } O_j\}$$

$$\min_{\theta} \sum_{(O_i, O_j) \in \mathcal{S}} (f(O_i, O_j, \theta) - 0)^2 + \sum_{(O_i, O_j) \in \mathcal{D}} (f(O_i, O_j, \theta) - 1)^2$$

Summary

Things to remember

- For similarity/distance computation, there are different solutions for different data types

Exercises for this topic

- **Data Mining, The Textbook (2015) by Charu Aggarwal**
  - **Exercises 3.9 on similarity measures**
- Introduction to Data Mining 2nd edition (2019) by Tan et al.
  - Exercises 2.6 $\rightarrow$ 14-28
- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al.
  - Exercises 3.5.7 on distance measures
- Data Mining Concepts and Techniques, 3rd ed. (2011) by Han et al.
  - Exercises 2.6 $\rightarrow$ 2.5-2.8