

DS 440 Data Mining

Lecture 6: Data Characteristics

1. Some terminology

Features/attributes/covariates						Target
Training samples	Per capita crime rate	Number of rooms	Age	Distance from employment centers	Accessibility to highways	Median Value of house
	0.00632	4	20	6	1	20
	0.00434	5	22	10	5	21
	0.053	8	5	1	2	4
	0.00653	3	13	4	4	32
	0.0134	5	2	20	5	11
Testing samples	Per capita crime rate	Number of rooms	Age	Distance from employment centers	Accessibility to highways	Median Value of house
	0.32	2	12	8	11	?
	0.05	3	02	2	5	?
	0.11	7	11	1	12	?

2. Types of datasets

2.1 Record data

Data matrix

Per capita crime rate	Number of rooms	Age	Distance from employment centers	Accessibility to highways
0.00632	4	20	6	1
0.00434	5	22	10	5
0.053	8	5	1	2
0.00653	3	13	4	4
0.0134	5	2	20	5

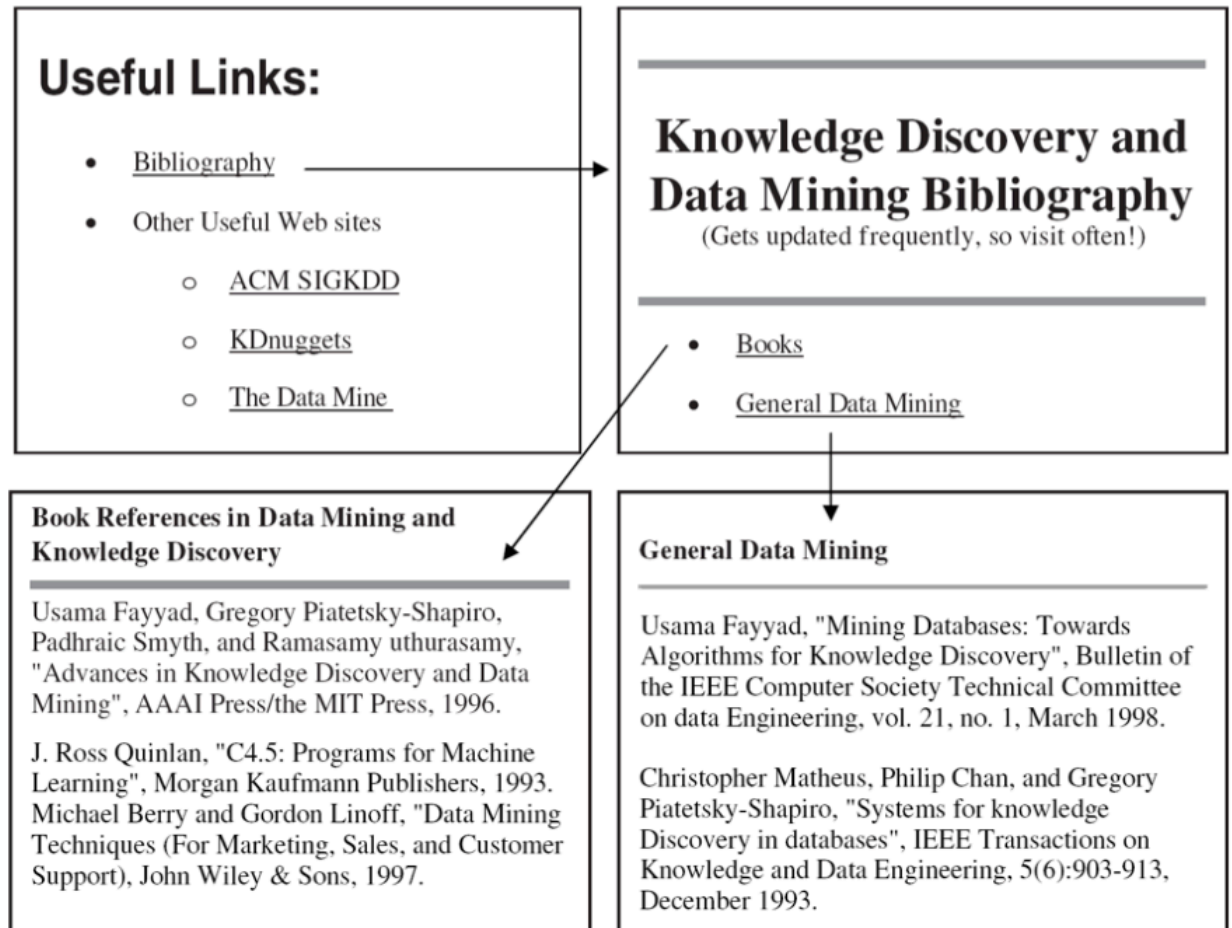
Document data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction data

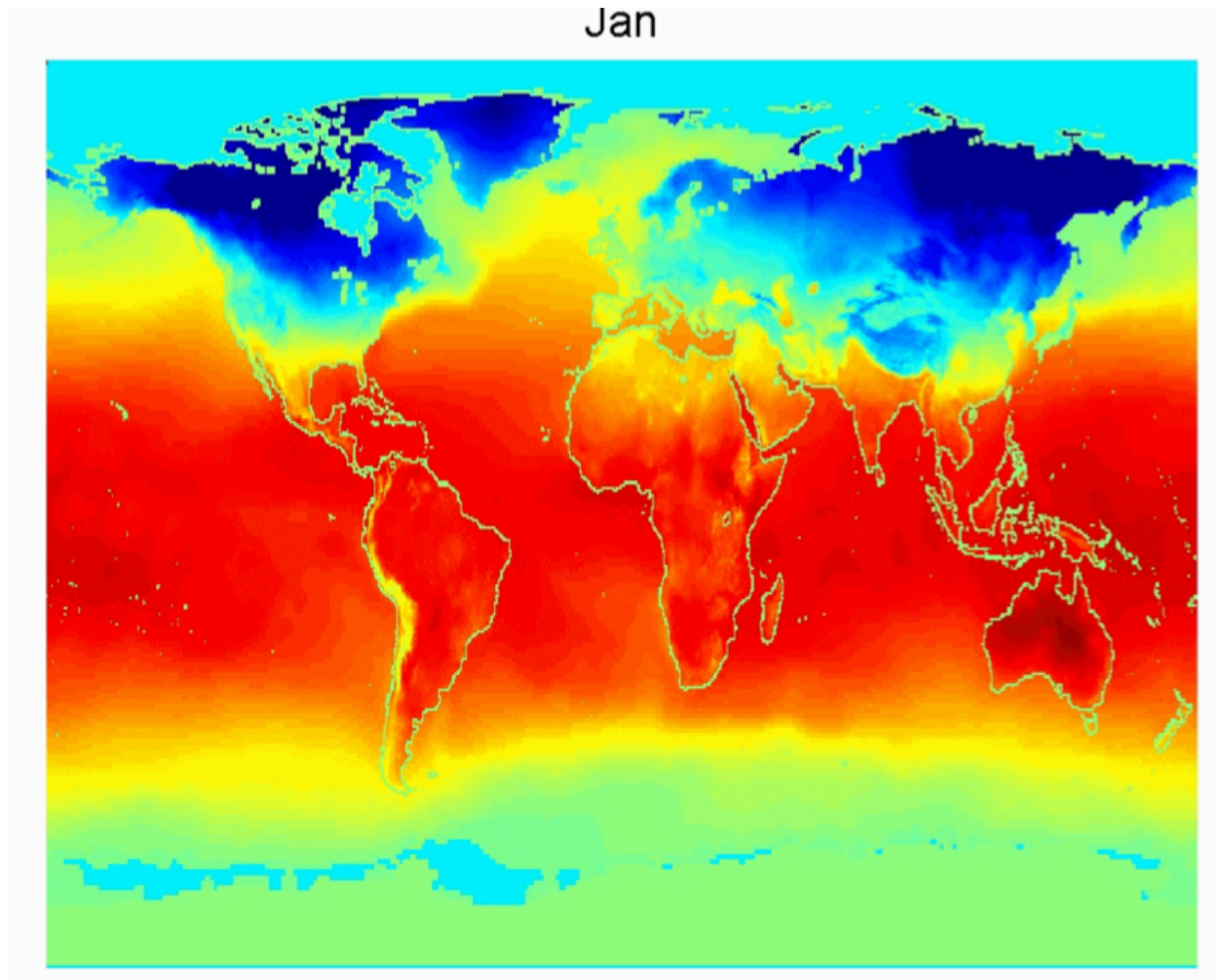
<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

2.2 Graph data



2.3 Ordered data

Average monthly temperature of land and ocean



3. Distance (Similarity) Measures

Data set may include samples that are duplicates, or almost duplicates of one another. This issue becomes very dominant when recording data from multiple sources. Duplicate samples in the dataset are bad because it leads to:

1. Extra computational time
2. Unstable data mining algorithms

Hence we need distance/similarity measures to understand the redundancy in datasets.

3.1 Dissimilarity/Distance measure (d)

1. Numerical measure of how different two data objects are
2. Lower when objects are more alike
3. Minimum dissimilarity is often 0
4. Upper limit varies

3.2 Similarity measure (s)

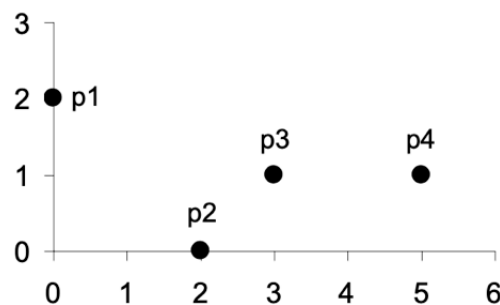
1. Numerical measure of how alike two data objects are
2. Is higher when objects are more alike.
3. Often falls in the range [0,1]

3.3 Some distance/similarity measure

1. Euclidean distance

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects x and y .



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Similarity measure can be some inverse functions of d

2. Minkowski Distance

Minkowski Distance is a generalization of Euclidean Distance

$$d(x, y) = \left(\sum_{k=1}^n |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects x and y .

Similarity measure can be some inverse functions of d

3. Similarity between binary vectors

- f_{01} = Number of attributes where x was 0 and y was 1
- f_{10} = Number of attributes where x was 1 and y was 0
- f_{00} = Number of attributes where x was 0 and y was 0
- f_{11} = Number of attributes where x was 1 and y was 1

For example:

$$\begin{aligned} x &= [1, 0, 0, 0, 0, 0, 0, 0, 0, 0] \\ y &= [0, 0, 0, 0, 0, 0, 1, 0, 0, 1] \end{aligned}$$

- Simple matching (SM)

$$SMC = (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) = (0 + 7) / (2 + 1 + 0 + 7) = 0.7$$

- Jaccard Coefficient (JC)

$$J = f_{11} / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$$

4 Cosine Similarity

If x and y are two document vectors, then

$$\cos(x, y) = \frac{x^T y}{||x|| \cdot ||y||}$$

where $x^T y$ is the vector dot product of x and y and $||x||$ is the length of vector x

5. Correlation

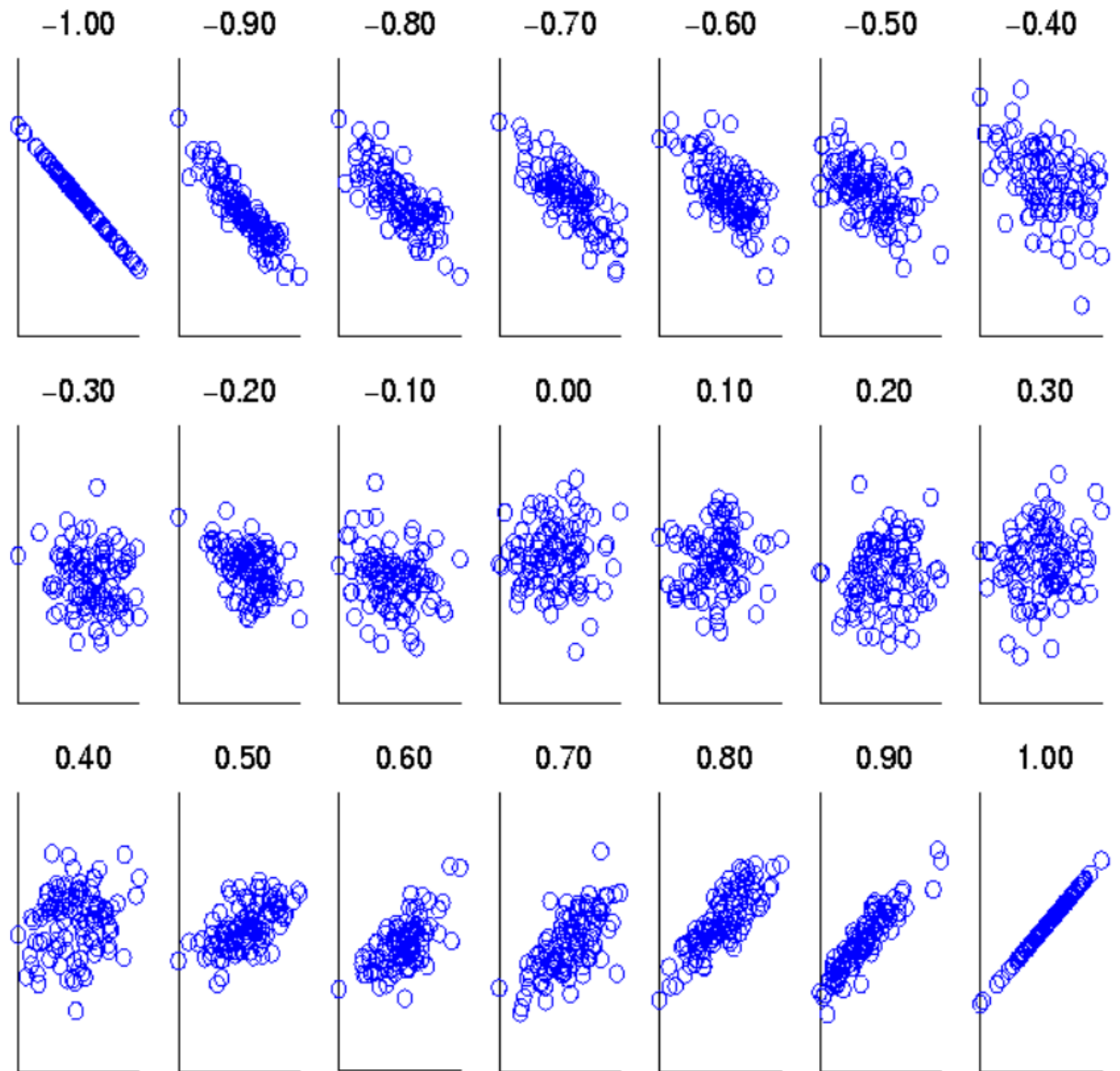
$$\text{corr}(x, y) = \frac{\text{covariance}(x, y)}{\text{standard_deviation}(x) \cdot \text{standard_deviation}(y)} = \frac{s_{xy}}{s_x \cdot s_y}$$

where

- $\text{covariance}(x, y) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \hat{x})(y_k - \hat{y})$
- $\text{standard_deviation}(x) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \hat{x})^2}$
- $\text{standard_deviation}(y) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \hat{y})^2}$
- $\hat{x} = \frac{1}{n} \sum_{k=1}^n x_k$, : mean of x
- $\hat{y} = \frac{1}{n} \sum_{k=1}^n y_k$, : mean of y

Why correlation is useful

Scatter plots showing similarity from -1 to 1

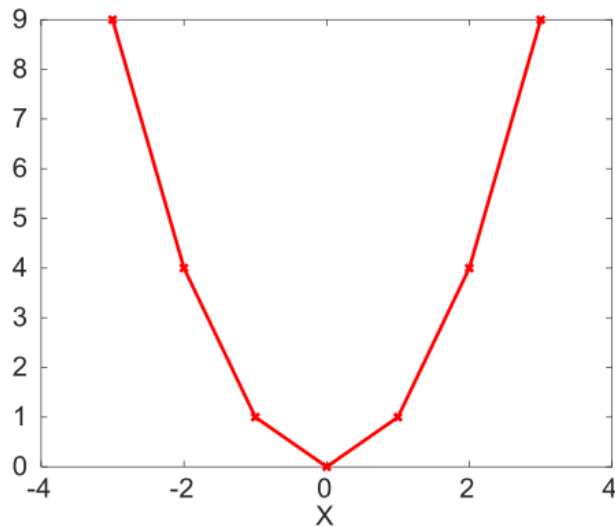


Why correlation can be misleading

Sampling data from $y = x^2$:

$$x = [-3, -2, -1, 0, 1, 2, 3]$$

$$x = [9, 4, 1, 0, 1, 4, 9]$$



Here:

- $\hat{x} = 0$
- $\hat{y} = 4$
- $s_x = 2.16$
- $s_y = 3.74$

$$\text{corr}(x, y) = \frac{-3 * 5 + -2 * 0 + -1 * -3 + 0 * -4 + 1 * -3 + 2 * 0 + 3 * 5}{6 * 2.16 * 3.74} = 0$$

Exercise :

$$d_1 = [3, 2, 0, 5, 0, 0, 0, 2, 0, 0]$$

$$d_2 = [1, 0, 0, 0, 0, 0, 0, 1, 0, 2]$$

1. For the above vectors, compute
 - Euclidean distance
 - Cosine Similarity
 - Correlation
2. Implement all the distances as functions in python

In []:

References

1. <https://www-users.cse.umn.edu/~kumar001/dmbook/index.php> (<https://www-users.cse.umn.edu/~kumar001/dmbook/index.php>)

In []: