Date your Data!

Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Attributes

- **Attribute (**or **dimensions, features, variables**)
  - A data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal (e.g., red, blue)
  - Binary (e.g., {true, false})
  - Ordinal (e.g., {freshman, sophomore, junior, senior})
  - Numeric: quantitative (discrete vs continuous)
  - Text
- Q1: Is student ID a nominal, ordinal, or interval-scaled data (measured on a scale of **equal-sized units and the order matters**)?
- Q2: What about eye color? Or color in the color spectrum of physics?

Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity or Dissimilarity
- Summary

Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function**
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike
  - Often falls in the range [0,1]: 0: no similarity; 1: 100% similar
- **Dissimilarity** (or **distance**) measure
  - Numerical measure of how different two data objects are
  - **In some sense, the inverse of similarity**: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range [0, 1] or [0, ∞) , depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with *l* dimensions

$$D = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1l} \\ x_{21} & x_{22} & \ldots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nl} \end{pmatrix}$$

- Dissimilarity (distance) matrix (n by n)
  - n data points, but registers only the distance *d*(*i, j*) (typically metric)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
  - Weights can be associated with different variables based on applications and data semantics

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \ldots & 0 \end{pmatrix}$$

Standardizing Numeric Data

- **Z-score**: $z = \frac{x - \mu}{\sigma}$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, "+" when above
- An alternative way: Calculate the mean absolute deviation

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

  - standardized measure (*z-score*): $z_{if} = \frac{x_{if} - m_f}{s_f}$
- **Using mean absolute deviation is more robust than using standard deviation**

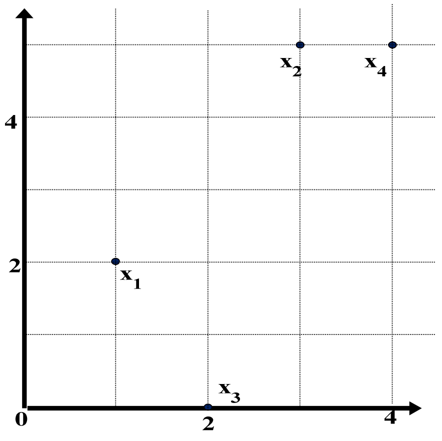
Z-score - An example

- ☐ John gets a mark of 64 in a physics test, where the mean is 50 and the standard deviation is 8.
- ☐ Jane gets a mark of 74 in a chemistry test, where the mean is 58 and the standard deviation is 10.

Who has a better class performance?

- ☐ John's z = (64 - 50) / 8 = 1.75
- ☐ Jane's z = (74 - 58) / 10 = 1.6
- ☐ Although Jane's score is higher, John's score is further above the mean, and it might be concluded that John has achieved greater success.

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix (by Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure

$$d(i,j) = \sqrt[p]{|x_{i1}-x_{j1}|^p + |x_{i2}-x_{j2}|^p + \cdots + |x_{il}-x_{jl}|^p}$$

where *i* = (x_{i1}, x_{i2}, …, x_{il}) and *j* = (x_{j1}, x_{j2}, …, x_{jl}) are two *l*-dimensional data objects, and *p* is the order (the distance so defined is also called L-*p* norm)

- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positivity)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric
- Note: There are nonmetric dissimilarities, e.g., set differences

Special Cases of Minkowski Distance

- *p* = 1: (L_1 norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{il}-x_{jl}|$$

- *p* = 2: (L_2 norm) Euclidean distance

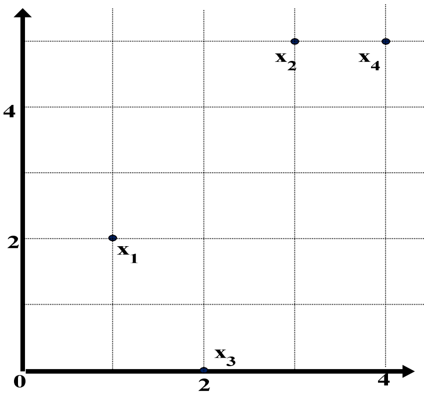
$$d(i,j) = \sqrt{|x_{i1}-x_{j1}|^2 + |x_{i2}-x_{j2}|^2 + \cdots + |x_{il}-x_{jl}|^2}$$

- $p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p\to\infty} \sqrt[p]{|x_{i1}-x_{j1}|^p + |x_{i2}-x_{j2}|^p + \cdots + |x_{il}-x_{jl}|^p} = \max_{f=1}^{l} |x_{if}-x_{jf}|$$

Example: Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Proximity Measure for Binary Attributes

- ❑ A contingency table for binary data

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

- ❑ Distance measure for symmetric binary variables $d(i, j) = \frac{r+s}{q+r+s+t}$
- ❑ Distance measure for asymmetric binary variables: $d(i, j) = \frac{r+s}{q+r+s}$
- ❑ Jaccard coefficient (*similarity/coherence* measure for *asymmetric* binary variables): $sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$
- ❑ Note: Jaccard coefficient is the same as (a concept discussed in Pattern Discovery)

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q+r) + (q+s) - q}$$

Example: Dissimilarity between Asymmetric 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 (**not counted in**)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- Distance: $d(i, j) = \frac{r+s}{q+r+s}$

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Mary

Jack

	1	0	Σ_{row}
1	2	0	2
0	1	3	4
Σ_{col}	3	3	6

Jim

Jack

	1	0	Σ_{row}
1	1	1	2
0	1	3	4
Σ_{col}	2	4	6

Mary

Jim

	1	0	Σ_{row}
1	1	1	2
0	2	2	4
Σ_{col}	3	3	6

Proximity Measure for Categorical Attributes

- Categorical data, also called **nominal** attributes
  - Example: Body Type (pear, banana, apple – 3 nominal states), profession, ~~Color (red, yellow, blue, green)~~, Categories (1, 2, 3), etc.
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of categorical variables

$$d(i, j) = \frac{p-m}{p}$$

- Method 2: Use a large number of asymmetric binary attributes
  - Creating a new binary attribute for each of the *M* nominal states

Proximity Measure for Categorical Attributes

- Method 3: Target encoding

1. Group the data by category
2. Calculate the average of the target variable per each group
3. Assign the average to each observation belonging to that group

Country	Target Variable	Target Encoding
United States	1	0.40
Germany	0	0.50
United States	0	0.40
United States	1	0.40
France	1	0.67
Germany	1	0.50
United States	0	0.40
France	1	0.67
United States	0	0.40
France	0	0.67

Ordinal Variables

- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
  - Replace *an ordinal variable value* by its rank: $r_{if} \in \{1, ..., M_f\}$
  - Normalization: Map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

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

    - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
      - Then L-1 distance: d(freshman, senior) = 1, d(junior, senior) = 1/3
  - Compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A dataset may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

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

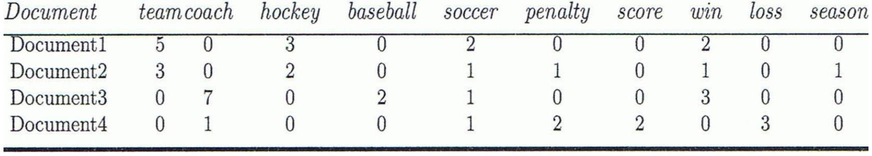
  - If f is numeric: Use the normalized distance
  - If f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise (there are other options …pp. 75-76)
  - If f is ordinal
    - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
    - Treat z_{if} as interval-scaled

Example

Object ID	Test 1 (nominal)	Test 2 (ordinal)	Test-3 (numeric)
1	A	Excellent	45
2	B	Fair	22
3	C	Good	64
4	D	Excellent	28

Cosine Similarity of Two Vectors (commonly used in document comparison)

- A **document** can be represented by a bag of terms/words or a long vector (very sparse), with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document



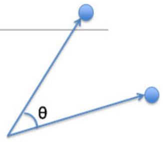
Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine similarity: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

where • indicates vector dot product, $||d||$: the **'length'** (Euclidean Norm) of vector *d*

Example: Calculating Cosine Similarity



- Calculating Cosine Similarity: $$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$ $$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

  where • indicates vector dot product, $||d||$: the length of vector *d*
- Ex: Find the **similarity** between documents 1 and 2.

  d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)
  - First, calculate vector dot product

    $d_1 \bullet d_2$ = 5 X 3 + 0 X 0 + 3 X 2 + 0 X 0 + 2 X 1 + 0 X 1 + 0 X 1 + 2 X 1 + 0 X 0 + 0 X 1 = 25
  - Then, calculate $||d_1||$ and $||d_2||$

$$\| d_1 \| = \sqrt{5\times5+0\times0+3\times3+0\times0+2\times2+0\times0+0\times0+2\times2+0\times0+0\times0} = 6.481$$

$$\| d_2 \| = \sqrt{3\times3+0\times0+2\times2+0\times0+1\times1+1\times1+0\times0+1\times1+0\times0+1\times1} = 4.12$$

  - Calculate cosine similarity: cos(d_1, d_2) = 25/ (6.481 X 4.12) = 0.94

KL Divergence: Comparing Two Probability Distributions

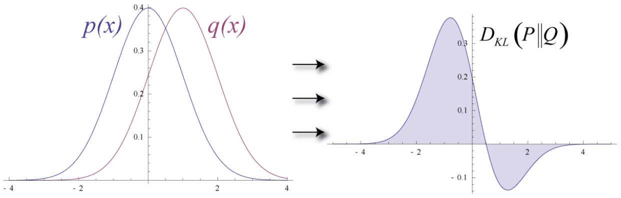
- *The Kullback-Leibler (KL) divergence:* Measure the difference between two probability distributions over the same variable *x*
  - From information theory, closely related to *relative entropy*, *information divergence*, and *information for discrimination*
- $D_{KL}(p(x) \,||\, q(x))$: divergence of $q(x)$ from $p(x)$, measuring the information lost when $q(x)$ is used to approximate $p(x)$

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Discrete form

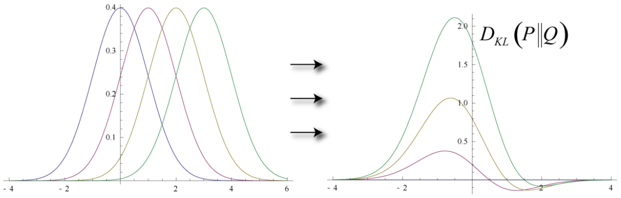
$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

Continuous form



Original Gaussian PDF's

KL Area to be Integrated



Ack.: Wikipedia entry: *The Kullback-Leibler (KL) divergence*

More on KL Divergence

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

- The KL divergence measures the expected number of extra bits required to code samples from *p*(*x*) (“true” distribution) when using a code based on *q*(*x*), which represents a theory, model, description, or approximation of *p*(*x*)
- The KL divergence is not a distance measure, not a metric: asymmetric, not satisfy triangular inequality ($D_{KL}(P\|Q)$ does not equal $D_{KL}(Q\|P)$)
- In applications, *P* typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while *Q* typically represents a theory, model, description, or approximation of *P*.
- The Kullback–Leibler divergence from *Q* to *P*, denoted $D_{KL}(P\|Q)$, is a measure of the information gained when one revises one's beliefs from the prior probability distribution *Q* to the posterior probability distribution *P*. In other words, it is the amount of information lost when *Q* is used to approximate *P*.
- The KL divergence is sometimes also called the information gain achieved if *P* is used instead of *Q*. It is also called the relative entropy of *P* with respect to *Q*.

Subtletly at Computing the KL Divergence

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

- *Base on the formula,* $D_{KL}(P,Q) \geq 0$ and $D_{KL}(P \,||\, Q) = 0$ if and only if $P = Q$
- How about when p = 0 or q = 0?
  - $\lim_{p \to 0} p \log p = 0$
  - when p != 0 but $q = 0$, $D_{KL}(p \,||\, q)$ is defined as ∞, i.e., if one event *e* is possible (*i.e.*, $p(e) > 0$), and the other predicts it is absolutely impossible (*i.e.*, $q(e) = 0$), then the two distributions are absolutely different.
- However, in practice, *P* and *Q* are derived from frequency distributions, not counting the possibility of unseen events. Thus *smoothing* is needed
- Example: $P : (a : 3/5, b : 1/5, c : 1/5)$. $Q : (a : 5/9, b : 3/9, d : 1/9)$
  - need to introduce a small constant ϵ, e.g., $\epsilon = 10^{-3}$
  - The sample set observed in *P*, $SP = \{a, b, c\}$, $SQ = \{a, b, d\}$, $SU = \{a, b, c, d\}$
  - Smoothing, add missing symbols to each distribution, with probability ϵ
  - $P' : (a : 3/5 - \epsilon/3, b : 1/5 - \epsilon/3, c : 1/5 - \epsilon/3, d : \epsilon)$
  - $Q' : (a : 5/9 - \epsilon/3, b : 3/9 - \epsilon/3, c : \epsilon, d : 1/9 - \epsilon/3)$
  - $D_{KL}(P' \,||\, Q')$ can then be computed easily

Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, etc.
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - **Measure data similarity**
- Above steps are the beginning of data preprocessing
- Many methods have been developed but still an active area of research

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu, et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009