CMSC828G **Principles of Data Mining** Lecture #2

- Today's Reading:
  - HMS, chapter 2
- Today's Lecture:
  - Project
  - Data and Measurement
- Upcoming Due Dates:
  - P0 due 2/7

Class Project

http://www.cs.umd.edu/class/spring2002/cmsc828g/project.htm

Some possible data sets

- Biomedical research literature
  - 6000 full-text articles
  - several large term lists identifying terms of various categories, for example "4-oxalocrotonate_tautomerase" is associated with function (it's an enzyme)
  - fledgling toolset for working with biomedical text
  - contact Phillip Resnik, resnik@umiacs.umd.edu

Some possible data sets, cont

- Land cover data
  - http://glcf.umiacs.umd.edu
  - contact Benjamin White, bgwhite@umiacs.umd.edu
- Web Access logs
  - access logs for Umiacs server
  - contact Atif Memon, atif@cs.umd.edu

Some possible data sets, cont

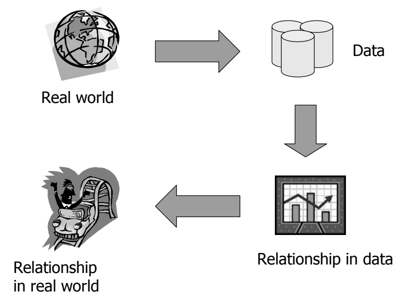
- Web Access logs
  - access logs for cs server
  - 84 files, each corresponds to a week's requests (about 1 million requests/week)
  - each record has:
    IP/domain name, timestamp, GET/POST, page url,
    http protocl, status code, length in bytes
  - In addition, for some of the periods, there are auxiliary logs which have referrer information.
  - contact Alexandrous Labrinidis, labrinid@cs.umd.edu

Some possible data sets, cont

- Stock data
  - closing prices and volume for all stocks for past year contact Alexandrous Labrinidis, labrinid@cs.umd.edu

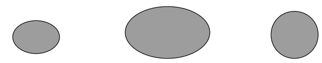
Measurement

Mapping domain entities to symbolic representations



Measurement cont.

- Rocks…



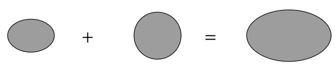
WEIGHT

weight	1	5	2	**Ranking**
weight	2	5	4	

Any monotonic (order preserving) transformation is legitimate

Measurement cont.

- Rocks…

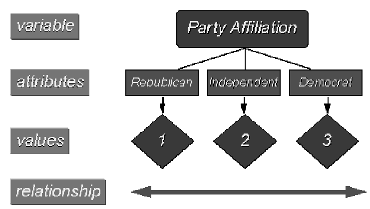


WEIGHT

weight	1	2	3	**Order and Additive**
weight	2	4	6	

- Numeric properties reflect empirical real world properties
- Allows inferences about physical system

Nominal attribute



Here, numerical values just "name" the attribute uniquely.
No ordering implied
I.e. jersey numbers in basketball; a player with number 30 is not more of anything than a player with number 15; certainly not twice whatever number 15 is.

http://trochim.human.cornell.edu/kb/measlevl.htm

Measurements, cont.

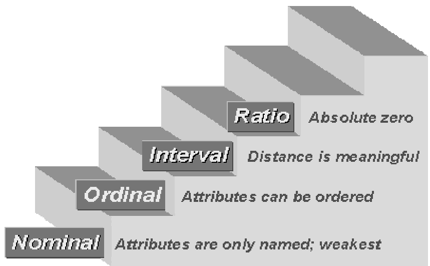
ordinal measurement - attributes can be rank-ordered. Distances between attributes do not have any meaning.
i.e., on a survey you might code Educational Attainment as 0=less than H.S.; 1=some H.S.; 2=H.S. degree; 3=some college; 4=college degree; 5=post college.
In this measure, higher numbers mean *more* education.
But is distance from 0 to 1 same as 3 to 4? No.
The interval between values is not interpretable in an ordinal measure.

interval measurement - distance between attributes *does* have meaning. i.e., when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. average makes sense, however ratios don't - 80 degrees is not twice as hot as 40 degrees

Measurements, cont.

ratio measurement - an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count" variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "...we had twice as many clients in the past six months as we did in the previous six months."

Hierarchy of Measurements



Scales

scale	Legal transforms	example
nominal	Any one-one mapping	Hair color, employment
ordinal	Any order preserving transform	Severity, preference
interval	Multiply by constant, add a constant	Temperature, calendar time
ratio	Multiply by constant	Weight, income

Why is this important?

- Make sure patterns found are genuine and not just an artifact of encoding
- Consider two sets of patients… pain 1-10

	P1	P2	P3	AVG	MEDIAN
Group 1	1	2	6	3	2
Group 2	3	4	5	4	4

consider new order preserving mapping: pain 1-20; 1→1, 2→2, 3→3, 4→4, 5→5, 6→12

	P1	P2	P3	AVG	MEDIAN
Group 1	1	2	12	**5**	2
Group 2	3	4	5	4	4

Other Kinds of Measurements

- A *representational* measurement: relationship between numbers in empirical system correspond to relationships between numbers in a numerical system.
- An *operational* or *non-representational* measurement: both defines the property *and* assigns a number to it.
  - Examples: quality of life in medicine, effort in software engineering

a = # of unique operators in program
b = # of unique operands
n = total # of operator occurrences
m = total # of operand occurrences

Programming effort: e = am(n+m)log(a+b)/2b

Distance Measures

- Many DM techniques (NN, clustering) are based on similarity measures between objects.
- Two methods for computing similarity:
  1. Explicit similarity measurement for each pair of objects
  2. Similarity obtained indirectly based on vector of object attributes.
- Metric: d(i,j) is a metric iff
  1. $d(i,j) \geq 0$ for all i, j and $d(i,j) = 0$ iff $i = j$
  2. $d(i,j) = d(j,i)$ for all i and j
  3. $d(i,j) \leq d(i,k) + d(k,i)$ for all i, j and k

Distance

- Notation: n objects with p measurements

$$x(i) = (x_1(i), x_2(i), \ldots, x_p(i))$$

- Most common distance metric is *Euclidean* distance:

$$d_E(i,j) = \left(\sum_{k=1}^{p} (x_k(i) - x_k(j))^2 \right)^{\frac{1}{2}}$$

- Makes sense in the case where the different measurements are commensurate; each is variable measured in the same units. If the measurements are different, say length and weight, it is not clear.

Standardization

When variables are not commensurate, we can standardize them by dividing by the sample standard deviation. This makes them all equally important.

The estimate for the standard deviation of x_k :

$$\hat{\sigma}_k = \left(\frac{1}{n} \sum_{i=1}^{n} \left(x_k(i) - \overline{x}_k \right)^2 \right)^{\frac{1}{2}}$$

where $\overline{x}_k$ is the sample mean:

$$\overline{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_k(i)$$

Weighted Euclidean distance

Finally, if we have some idea of the relative importance of each variable, we can weight them:

$$d_{WE}(i,j) = \left(\sum_{k=1}^{p} w_k (x_k(i) - x_k(j))^2 \right)^{\frac{1}{2}}$$

Additive Distances

- Each variable contributes independently to the measure of distance.
- May not always be appropriate…

cup i

cup j

diameter(i)
height(i)
height_2(i)
⋮
height_{100}(i)

diameter(j)
height(j)
height_2(j)
⋮
height_{100}(j)

Dealing with Correlation

- Standardize the variables, not just in direction of variable, but also taking into account covariances.
- Assume we have two variables or attributes X and Y and n objects taking on values x(1), …, x(n) and y(1), …, y(n). The sample covariance of X and Y is:

$$\text{Cov}\,(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x(i) - \overline{x})(y(i) - \overline{y})$$

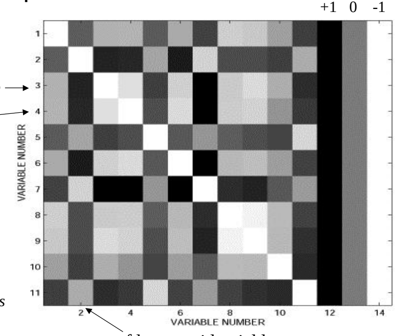
- The covariance is a measure of how X and Y vary together.
  - it will be large and positive if large values of X are associated with large values of Y, and small X $\Rightarrow$ small Y
  - it will be large and negative if large X $\Rightarrow$ small Y

Sample correlation coefficient

- Covariance depends on ranges of X and Y
- Standardize by dividing by standard deviation
- Sample correlation coefficient

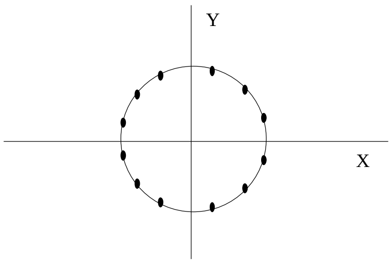
$$\rho(X,Y) = \frac{\sum_{i=1}^{n} (x(i) - \overline{x})(y(i) - \overline{y})}{\left(\sum_{i=1}^{n} (x(i) - \overline{x})^2 (y(i) - \overline{y})^2 \right)^{\frac{1}{2}}}$$

Sample Correlation Matrix



data on characteristics of Boston surburbs

What about…



Are X and Y dependent?

ρ(X,Y) = ?

linear covariance, correlation

Mahalanobis distance

$$d_{MH}(i, j) = \left(\left(x(i) - x(j)\right)^T \Sigma^{-1} \left(x(i) - x(j)\right)\right)^{\frac{1}{2}}$$

1. It automatically accounts for the scaling of the coordinate axes
2. It corrects for correlation between the different features

Price:

1. The covariance matrices can be hard to determine accurately
2. The memory and time requirements grow quadratically rather than linearly with the number of features.

Other Distance Metrics

- Minkowski or L_λ metric:

$$d(i, j) = \left(\sum_{k=1}^{p} (x_k(i) - x_k(j))^\lambda\right)^{\frac{1}{\lambda}}$$

- Manhattan, city block or L_1 metric:

$$d(i, j) = \sum_{k=1}^{p} \left|x_k(i) - x_k(j)\right|$$

- L_∞

$$d(i, j) = \max_k \left|x_k(i) - x_k(j)\right|$$

Binary Data

	j=1	j=0
i=1	n_{11}	n_{10}
i=0	n_{01}	n_{00}

- matching coefficient

$$\frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

- Jaccard coefficient

$$\frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

Transforming Data

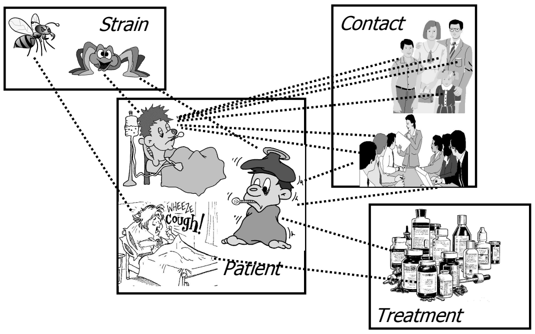
- Duality between form of the data and the model
- Common transforms: square root, reciprocal, logarithm, raising to a power.
- logit

$$\log it(p) = \frac{p}{1 - p}$$

Schema

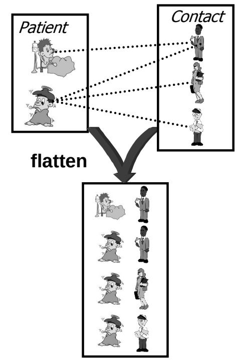
- multirelational data
- time series data
- string data
- text
- event sequence

Discovering Patterns in Structured Data



Multirelational Data

Traditional approaches
- work well with flat representations
- fixed length attribute-value vectors
- assume independent (IID) sample



Problems:
- introduces statistical skew
- loses relational structure
  - incapable of detecting link-based patterns
- must fix attributes in advance

Data Quality

- GIGO
- sources of error: human carelessness, instrument failure, data corruption, ….
- measurement error:
  - precise, reliable
  - accurate
- Data quality for collections of data
  - representative sample?
  - opportunity or convenience sample
  - population drift

Next Time

- Reading:
  - HMS, chapter 3
- Topic:
  - Visualization

References

- *Principles of Data Mining*, Hand, Mannila, Smyth. MIT Press, 2001.
- Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. (version current as of 2001).
- Pattern Recognition for HCI. Richard Duda, 1996.