Notes 01-2: Characteristics of Data

In this section, we provide some background into how data describing real-world entities or events is represented and stored.

How data is represented and stored is affected by a number of important issues, including:

- the type of the data (e.g. what phenomena does the data describe)
- the types of the attributes (e.g. numeric or non-numeric, discrete or continuous)
- the quality of the data (e.g., missing data, duplicate data)
- how the data will be used (e.g., which data mining technique is most suitable)

Types of Data

Data is used to create a *data object* that represents some abstraction of a real-world entity or event.

Depending on the problem domain, a data object may be referred to as a *record*, *point*, *vector*, *pattern*, *event*, *case*, *sample*, *observation*, or *entity*.

A data object consists of a number of *attributes* that describe its basic characteristics (e.g., customer record, land description, medical diagnosis).

Again, depending on the problem domain, an attribute may be referred to as a *variable*, *characteristic*, *field*, *feature*, or *dimension*.

A collection of data objects is called a *dataset*. For convenience, in this course, we assume there are *three* types of datasets: record data, graph-based data, and ordered data.

Record Data

A collection of records, each of which consists of a fixed set of attributes, usually stored in a flat file or relational database.

Transaction data: A collection of data objects consisting of sets of items representing a transaction.

Example

TID	Items
t_1	*bread, soda, milk*
t_2	*beer, bread*
t_3	*beer, soda, pepto bismol*
t_4	*beer, vodka, soda, stomach pump*
t_5	*soda, diaper, milk*

Data matrix: A collection of data objects that all have the same fixed set of numeric attributes, where each data object represents a point in a multidimensional space.

Example

W	X	Y	Z
10.23	5.27	27	1.2
12.65	6.25	22	1.1
13.54	7.23	23	1.2
14.27	8.43	25	0.9

Sparse matrix: A special case of the data matrix, where attributes contain binary values (i.e., 0 (1) indicates that some event has not (has) occurred).

Example

TID	bread	soda	milk	beer	vodka	stomach pump	diaper	pepto bismol
t_1	1	1	1	0	0	0	0	0
t_2	1	0	0	1	0	0	0	0
t_3	0	1	0	1	0	0	0	1
t_4	0	1	0	1	1	1	0	0
t_5	0	1	1	0	0	0	1	0

Graph-Based Data

Graphs can be used to represent the relationship between data objects or to represent the data objects themselves.

Relationships between data objects: Data objects are represented by nodes and relationships are represented by arcs.

Example: Linked web pages

Data objects are graphs: Data objects consist of components whose relationship describes some kind of structure.

Example: Molecules in chemical compounds

Ordered Data

The relationship of one data object to other data objects involves order based upon time or space.

Sequential data: An extension of record data, where each record has a time stamp associated with it.

Example

TID	Time	Customer	Items
t_1	*T1*	*C1*	*A, B*
t_2	*T2*	*C3*	*A, C*
t_3	*T3*	*C1*	*C, D*
t_4	*T4*	*C2*	*A, D*
t_5	*T5*	*C2*	*E*
t_6	*T6*	*C1*	*A, E*

C1: (T1: A, B) (T2: C, D) (T5: A, E)

Sequence data: When the positions of the individual attributes in a data object describe an ordered sequence.

Example: Genomic sequence data expressed using the four nucleotides from which DNA is constructed: A, T, G, and C.

Time series data: A special case of sequential data, where each record is a measurement of some phenomenon taken at a particular time.

Example: A data set of daily prices for stocks on the stock market.

Spatial data: When the attributes of a data object represent some phenomenon related to location.

Example: A data set of precipitation and temperature measurements for a number of geographical locations.

Types of Attributes

Operations on Numbers

- *Distinctness*: =, ≠
- *Order*: <, ≤, >, and ≥
- *Addition*: +, −
- *Multiplication*: *, /

Types of Attributes

Given the four operations shown above, *four* types of attributes can be defined (note that for each attribute type described below, it possesses all the properties of the attribute type above it, and the relevant operations include those relevant to the attribute type above it): nominal scale data, ordinal scale data, interval scale data, and ratio scale data.

Nominal Scale Data

Nominal scale data: Values essentially just name things.

Valid operations = {Distinctness}

For similar data objects, names simply enable one instance of the data object to be distinguished from another, but nothing more can be implied from the name.

Example: CS Department computers: Venus, Hercules, Grendel

For dissimilar data objects, names simply enable instances of the data objects to be categorized or grouped, but nothing more can be implied from the name.

Example: Butcher, Baker, Candlestick Maker

There is no inherently implied or meaningful order in nominal data.

Ordinal Scale Data

Ordinal scale data: Values contain information that assigns some meaningful order to the data.

Valid operations = {Distinctness, Order}

Examples: {Small, Medium, Large}, {Cold, Cool, Warm, Hot}

The ordering (or ranking) of values is done according to the notion of transitivity (i.e., if a > b and b > c, then a > c).

The ordering (or ranking) of values does not imply or require that the amount of difference between each value be specified.

Interval Scale Data

Interval scale data: The differences between the values are meaningful.

Valid operations = {Distinctness, Order, Addition}

The ordering (or ranking) implies that some unit of measurement exists and that the differences between values and changes in values can be quantified.

Example: Temperature measured using a thermometer

The zero point is often arbitrarily chosen, affecting how the differences or changes can be quantified.

Example: Can’t say that 80 Fahrenheit is twice as hot as 40 Fahrenheit.

Ratio Scale Data

Ratio scale data: Values are similar to interval scale values except the values contain enough information to determine the relative difference between them.

Valid operations = {Distinctness, Order, Addition, Multiplication}

Typically, has a “true” zero point.

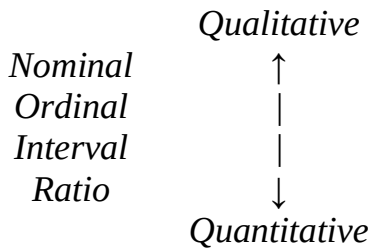
Two classes of ratio scale data: *named* and *unnamed*.

Example (named ratio scale data): 80 Kelvin is twice as hot as 40 Kelvin

Example (unnamed ratio scale data): A sensor is on for 10 seconds and off for 5 seconds. The duty cycle is 10 / 5 = 2 (with no units).

Continuums

The *qualitative – quantitative continuum* captures the low to high information content of the different types of attributes.



The *discrete – continuous continuum* distinguishes attributes by the number of values they can take.

A *discrete attribute* has at least two or more distinct values.

A *continuous attribute* is one whose values are real numbers and that, in theory, another value can be found between any two values.

Single-valued attributes are those whose value does not change (i.e., constants).

- Such an attribute is a defining characteristic of the data object.
- Such an attribute does not contribute any information for describing the object.
- Example: Extracting credit card transactions from a database for gold card members. Once these members have been extracted from the database, the gold card indicator is no longer informative.

Two-valued attributes are known as *dichotomous* attributes.

- Many mining techniques are specifically designed for dichotomous attributes.

- Example: A classification tree can classify a mortgage application as low or high risk.

- Binary attributes are a type of dichotomous attribute that takes on only the values 0 (i.e., zero) and 1 (i.e., one).

Data Quality

Data quality issues need to be considered before mining the data.

Archived data may not have originally been collected and stored with the intent of mining it in the future.

Data may not be properly documented (i.e., What is it exactly that the data describes?).

The data may not be relevant to the phenomenon that is being studied (i.e., Are the attributes appropriate for properly describing the phenomenon?).

The data may be "stale" (i.e. no longer reflects the true state of the real world).

Imperfections in Data

It is not realistic to expect "perfect" data because humans, measuring devices, and processes are subject to errors or flaws.

A ***measurement error*** occurs when the value recorded for an attribute differs from the actual value.

Example

> When a measurement is taken with a measuring device that has not been calibrated against some known quantity.
>
> When a measuring device has not been "zeroed" prior to taking a measurement.

A ***data collection*** error occurs when data objects or attributes values are omitted from the data, or when an inappropriate data object or attribute value is included in the data.

Example

When measuring diesel fuel in a large storage tank, levels can fluctuate due to changes in ambient temperature. The temperature must always be recorded at the time of the measurement.

Noise is the random component of a measurement error that may include distortion of attribute values or the addition of spurious objects.

Example

When a measuring device is affected by environmental conditions, such as lightning.

When a survey respondent completes a survey containing combinations of values that cannot be true, such as a marital status = single and spouse's job = teacher.

An ***artifact*** is a data error caused by some deterministic phenomenon (i.e., the same error occurred when collecting values for each data object).

Example

When some flaw in a lens (e.g., dirt or scratch) is systematically reproduced on a group of medical images.

When a poor pseudo-random number generator introduces regularities into some phenomenon being studied.

Precision is the closeness of values obtained by repeatedly measuring the same quantity.

Example

When a standard weight of 1g is weighed five times resulting in the values 1.015, 0.990, 1.013, 1.001, and 0.986. The precision, as measured by the standard deviation, is 0.013.

Bias is the systematic variation of the values obtained for the quantity being measured.

Example

When a standard weight of 1g is weighed five times resulting in the values 1.015, 0.990, 1.013, 1.001, and 0.986. The bias, as measured by the difference between the mean and the actual weight, is 0.001.

Accuracy is the closeness of the values obtained to the true value of the quantity being measured.

Example

When using a scale that only registers weight in 0.1g increments. Any weight registered is accurate only to within ±0.05g.

Outliers are either:

Data objects that have different characteristics from most of the other typical data objects in the data set, or

Attribute values that are different from other typical values for the attribute.

Example

When analyzing weights for 100 newborns, most of whose weights seem to fall within the 6 to 10 pound range, and then a weight of 16 pounds is encountered. The value is not typical, but it does represent a valid measurement.

A ***missing value*** is one that has not been entered into a dataset, but an actual value does exist for it in the real world.

Example

Failing to record a patient's gender.

Failing to complete all questions in a survey.

An ***empty value*** is one for which no real world value can be supposed.

Example

An attribute that records the number of miscarriages on a male patient's medical record. The value 0 (i.e., zero) seems correct, but it misrepresents the context and may not be appropriate.

Inconsistent values are values for different attributes that don't seem to belong together.

Example

> When marital status = single and spouse's job = teacher.
>
> When city = Regina and postal code = 56560.

Duplicate data is multiple copies of the same data object or nearly identical copies of different data objects.

Example

> When people living at the same address are on a mailing list and receive multiple copies of a community newsletter.
>
> When your name appears multiple times on a mailing list because it is spelled differently each time.

Data Preprocessing

Data preprocessing is a strategy or technique for preparing the data for mining

The following types of data preprocessing are considered in this class: aggregation, sampling, dimensionality reduction, feature subset selection, feature creation, discretization, and variable transformation.

Aggregation

Aggregation is the combining of two or more data objects into a single data object.

Example

> When determining the number of different vehicle models sold at dealerships in some geographic region, the sales records for each model are aggregated into one sales record.

Sampling

Sampling is the process of selecting a subset of the data set to be mined.

The sample is *representative* of the data set if it has approximately the same properties as the original data set.

Using *simple random sampling*, there is an equal probability of selecting any particular data object.

When *sampling without replacement*, as each data object is selected, it is removed from the data set (i.e., each data object can only be selected once).

When *sampling with replacement*, as each data object is selected, it is placed back into the data set (i.e., a data object could be selected multiple times).

Using *stratified sampling*, data objects are selected in proportion to their occurrence in the original dataset.

Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of data objects or attributes.

Some data objects or attributes may be irrelevant to the data mining task.

Example

> When studying fertility rates for women between the ages of 35 and 45, those whose age is not in the range from 35 to 45 are not required and males are not required.

Feature Subset Selection

Feature subset selection is the process of removing attributes that are redundant or irrelevant.

Redundant features duplicate information contained in one or more other attributes.

Example

> When completing a survey, the city/town is not required if the postal code has already been provided.

Irrelevant features contribute no useful information.

Example

> When studying fertility rates for women, the gender is inherent in the data and is no longer required.

Feature weighting is used to assign weights to features based upon their perceived importance.

Example

> When assigning weights to words that are most useful in a text classification task (e.g., articles and nouns typically have a zero and non-zero weight, respectively).

Feature Creation

Feature creation is the process of creating a new set of attributes from information stored in the original attributes.

Feature extraction is the process of refining a set of features from the original "raw" data (i.e., extracting those features that have "useful" characteristics) by mapping a large problem space into a smaller problem.

Example

> When text mining, useful keywords in a large document are summarized in a small keyword vector.

Feature construction is the process of creating one or more new features from the original features.

Example

> When body mass index is derived from height and weight.

Discretization

Discretization is the process of transforming a continuous attribute into a discrete (i.e., categorical) attribute.

Unsupervised discretization does not use class information when deciding on the *split points*.

Equal width discretization divides the range of the attribute into a user-specified number of equal width intervals.

Example

> When the values 2, 3, 4, 4, 5, 5, 6, 6, 7, 9 are divided into the intervals 2, 3, 4, 4, 5, 5 and 6, 6, 7, 9. Both intervals have a width of four.

Equal frequency discretization attempts to divide the data objects so that an equal number are placed in each interval.

Example

> When the values 2, 3, 4, 4, 5, 5, 6, 6, 7, 9 are divided into the intervals 2, 3, 4, 4, 5 and 5, 6, 6, 7, 9. Both intervals contain five values.

Supervised discretization uses class information when deciding on the split points (usually based upon some information-theoretic or statistical approach).

Variable Transformation

Variable transformation is the process of applying some function or algorithm to all the values of a variable so that the transformed values have more "appropriate" properties.

Simple mathematical functions can be applied to values.

Example

> When using the logarithm of large numbers rather than the numbers themselves.

Normalization (or standardization) can be used re-scale the set of values associated with a variable.

Example

When comparing gross national product to family income between rich and poor nations. Comparing real dollars may skew the results, but normalizing incomes on the interval (0, 1) may make the comparison more meaningful.