Data Scales and representation

Prof. Asim Tewari
IIT Bombay

Data Mining

- Data mining is a process of discovering patterns in data sets to achieve some specific objective. This involving methods at the intersection of machine learning, statistics, and database systems.

- In the 1960s, statisticians and economists used terms like *data fishing* or *data dredging* to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis.

Data Mining Skill Set

- Statistics
- Programming Languages

Pre-processing

- Data Extraction & processing
- Data wrangling and exploration
- Machine Learning models
- Data Visualization

Post-processing

Business acumen

Data Mining Tasks

- Gathering Business objectives
- Data acquisition

Pre-processing

- Data processing
- Data exploration
- Data Modeling
- Data Visualization
- Model deployment

Post-processing

Data Mining job profiles

Designation	Role
Data analyst manager	Manage the data mining group
Data Scientist	Design, develop and deploy data models
Data analyst	
Data Architecture	Provide secure and efficient access to data.
Data Engineer	
Database administrator	
Business analyst	Provide business objectives
Statistician	Provide statistical insights

Data Type

- Discrete data:
  - Discrete non-ordered numbers
  - Random collection of words
  - Unrelated audio sounds
  - Random music notes
- Sequential (temporal) data:
  - Stochastic process
  - Sequence of words in a sentence
  - Audio speech data
  - Music
- Spatial data:
  - Image data
  - Geo-spatial data

Sequential Spatio-temporal data

Other classifications include

- Categorical vs numerical
- Qualitative vs Quantitative

Data Scales

- Same numerical data may have different semantic meanings

- Depending on the semantic meaning different types of mathematical operations are appropriate

Data Scales

- Based on semantic meanings there are four different *scales*

Scale	Operations		Example	Statistics
Ratio	$\cdot$	$/$	21 years, 273 K	Generalized mean
Interval	$+$	$-$	2015 A.D., 20 °C	Mean
Ordinal	$>$	$<$	A, B, C, D, F	Median
Nominal	$=$	$\neq$	Alice, Bob, Carol	Mode

- For each scale level the operations and statistics of the lower scale levels are also valid

Data Scales

Scale	Operations		Example	Statistics
Ratio	$\cdot$	$/$	21 years, 273 K	Generalized mean
Interval	$+$	$-$	2015 A.D., 20 °C	Mean
Ordinal	$>$	$<$	A, B, C, D, F	Median
Nominal	$=$	$\neq$	Alice, Bob, Carol	Mode

For each scale level the operations and statistics of the lower scale levels are also valid

- Nominal scaled data
  - Only tests for equality or non-equality are valid.
  - Data of a nominal feature can be represented by the mode (value that occurs most frequently.)

Data Scales

Scale	Operations		Example	Statistics
Ratio	$\cdot$	$/$	21 years, 273 K	Generalized mean
Interval	$+$	$-$	2015 A.D., 20 °C	Mean
Ordinal	$>$	$<$	A, B, C, D, F	Median
Nominal	$=$	$\neq$	Alice, Bob, Carol	Mode

For each scale level the operations and statistics of the lower scale levels are also valid

- Ordinal scaled data
  - The operations "greater than" and "less than" are valid
  - inequality, and the combinations "greater than or equal" (≥)and "less than or equal" (≤).
  - The relation "less than or equal" (≤) defines a *total order*, such that for any *x*; *y*; *z* we have
    - Antisymmetry

$$(x \leq y) \wedge (y \leq x) \Rightarrow (x = y)$$

    - Transitivity

$$(x \leq y) \wedge (y \leq z) \Rightarrow (x \leq z)$$

    - Totality

$$(x \leq y) \vee (y \leq x)$$

  - Represented by the median (the value for which (almost) as many smaller as larger values exist)

Data Scales

Scale	Operations		Example	Statistics
Ratio	$\cdot$	$/$	21 years, 273 K	Generalized mean
Interval	$+$	$-$	2015 A.D., 20 °C	Mean
Ordinal	$>$	$<$	A, B, C, D, F	Median
Nominal	$=$	$\neq$	Alice, Bob, Carol	Mode

For each scale level the operations and statistics of the lower scale levels are also valid

- Interval scaled data
  - addition and subtraction are valid
  - have arbitrary zero points
  - represented by the (arithmetic) mean

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

Data Scales

Scale	Operations		Example	Statistics
Ratio	$\cdot$	$/$	21 years, 273 K	Generalized mean
Interval	$+$	$-$	2015 A.D., 20 °C	Mean
Ordinal	$>$	$<$	A, B, C, D, F	Median
Nominal	$=$	$\neq$	Alice, Bob, Carol	Mode

For each scale level the operations and statistics of the lower scale levels are also valid

- Ratio scaled data
  - multiplication and division are valid
  - represented by the generalized mean

$$m_\alpha(X) = \sqrt[\alpha]{\frac{1}{n}\sum_{k=1}^{n} x_k^\alpha}$$

Data Type, Data Scale, Data value

Date Type, Data Scale and Data values are three different concepts

- Data Type:
  - Discrete Type
    - Order of collection does not matter
  - Sequential Type
    - One directional order of collection
  - Spatio-temporal Type
    - Multidimensional order of collection

These can be of any Data Scale

- Data Scale
  - Ratio
  - Interval
  - Ordinal
  - Nominal

- Data value
  - Discrete (numerical or non-numerical)
  - Continuous (numerical also called quantitative)

Data Type, Data Scale, Data value

Date Type, Data Scale and Data values are three different concepts

- Data Type:
  - Discrete Type
    - Order of collection does not matter
  - Sequential Type
    - One directional order of collection
  - Spatio-temporal Type
    - Multidimensional order of collection

These can be of any Data Scale

- Data Scale
  - Ratio ->Can be only numerical (also called quantitative)
  - Interval -> Can be only numerical (also called quantitative)
  - Ordinal -> Can be categorical or Qualitative
  - Nominal -> Can be only categorical (?)

- Data value
  - Discrete (numerical or non-numerical)
  - Continuous (numerical also called quantitative)

1985 Auto Imports Database

	Attribute:	Attribute Range:
1.	symboling:	-3, -2, -1, 0, 1, 2, 3.
2.	normalized-losses:	continuous from 65 to 256.
3.	make:	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4.	fuel-type:	diesel, gas.
5.	aspiration:	std, turbo.
6.	num-of-doors:	four, two.
7.	body-style:	hardtop, wagon, sedan, hatchback, convertible.
8.	drive-wheels:	4wd, fwd, rwd.
9.	engine-location:	front, rear.
10.	wheel-base:	continuous from 86.6 120.9.
11.	length:	continuous from 141.1 to 208.1.
12.	width:	continuous from 60.3 to 72.3.
13.	height:	continuous from 47.8 to 59.8.
14.	curb-weight:	continuous from 1488 to 4066.
15.	engine-type:	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16.	num-of-cylinders:	eight, five, four, six, three, twelve, two.
17.	engine-size:	continuous from 61 to 326.
18.	fuel-system:	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19.	bore:	continuous from 2.54 to 3.94.
20.	stroke:	continuous from 2.07 to 4.17.
21.	compression-ratio:	continuous from 7 to 23.
22.	horsepower:	continuous from 48 to 288.
23.	peak-rpm:	continuous from 4150 to 6600.
24.	city-mpg:	continuous from 13 to 49.
25.	highway-mpg:	continuous from 16 to 54.
26.	price:	continuous from 5118 to 45400.

Abalone (sea snails) data

Name	Data	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

Census bureau database

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K

Variables in ML

- The **inputs** go by different names, such as *predictors*, *independent variables*, *features*, or sometimes just *variables* and is typically denoted using the symbol *X*
- The **output** variable is often called the *response* or *dependent variable*, and is typically denoted using the symbol *Y*

Supervised Machine Learning

- Our goal in supervised machine learning is to extract a relationship from data (ordered pairs of *(y,x)*)

The real relation is

$$y = f(x) + \epsilon$$

ϵ is noise with zero mean.

What we get from learning from data is

$$\hat{y} = h(x)$$

Regression vs Classification

$$y = f(x) + \epsilon$$

- If *y* is in interval or ratio scale, then it is regression
- If *y* is in Nominal or ordinal (?) scale, then it is classification

Regression vs Classification

$$y = f(x) + \epsilon$$

- The task of <u>classification</u> differs from <u>regression</u> in that we assign a <u>discrete number</u> of classes (nominal scale or ordinal scale), instead of assigning it a <u>continuous value</u> (interval or ratio scale).

Data Set vs Matrix Representations

We can denote numerical feature data as a **set**

$$X=\{x_1,x_2, ..,x_n\} \in R^{pxn}$$

- with *n* elements, where
- each element is a *p*-dimensional real-valued feature vector, where *n* and *p* are positive integers. For *p* = 1 we call *X* a *scalar* data set.

Data Set and Matrix Representations

- As an alternative to the set representation, numerical feature data are also often represented as a **matrix**

$$X = \begin{pmatrix} x_1^{(1)} & \cdots & x_1^{(p)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \cdots & x_n^{(p)} \end{pmatrix}$$

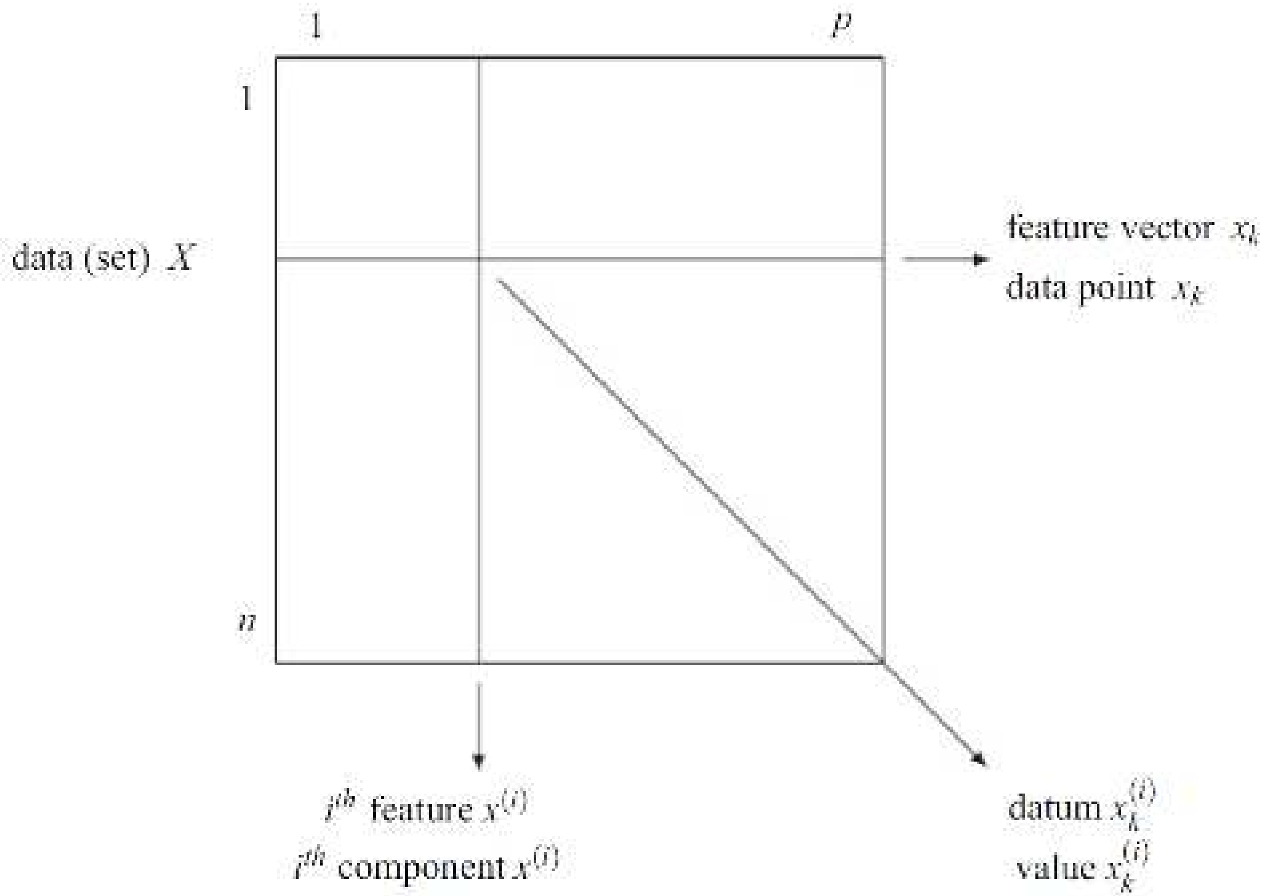
- Each row of the data matrix corresponds to an element of the data set. It is called *feature vector* or *data point* x_k, $k = 1, \ldots, n$.

- Each column of the data matrix corresponds to one component of all elements of the data set. It is called i^{th} *feature* or i^{th} *component* $x^{(i)}$, $i = 1, \ldots, p$.

- A single matrix element is a component of an element of the data set. It is called *datum* or *value* $x_k^{(i)}$, $k = 1, \ldots, n$; $i = 1, \ldots, p$.

Data Set and Matrix Representations

- Matrix representation of a data set



Data Relations

- Consider a set of (abstract categorical) elements, with no feature vector representation for the objects.

$$O = \{o_1, \ldots, o_n\}$$

- So conventional feature-based data analysis methods are not applicable. Instead, the relation of all pairs of objects can often be quantified and written as a square matrix

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Data Relations

- Each relation value r_{ij}, $i; j = 1,\ldots., n$, may refer to a degree of similarity, dissimilarity, compatibility, incompatibility, proximity or distance between the pair of objects o_i and o_j.
- R may be symmetric, so $r_{ij} = r_{ji}$ for all $i, j = 1,\ldots.,n$.
- R may be manually defined or computed from features. If numerical features X are available, then R may be computed from X using an appropriate function $f : \mathrm{R}^n \times \mathrm{R}^n \rightarrow \mathrm{R}$.