Sampling

→ Obtaining a small sample s to represent the whole data set N

- Simple random sampling
- Sampling w/o replacement
- " with "
- ~~Sta~~ Stratified sampling
  - ↳ partion data
  - ↳ draw samples from each partition.

Data Transformation

☆ a func. that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.

methods:

- Smoothing
- Attribute / feature construction
- Aggregation, - summary
- Normalization : data is scaled
- Discretization.
- concept hierarchy generation

Data Transformation by Normalization

→ measurement unit can affect data analysis.

→ To avoid dependence on the choice of measurement units, the data should be normalized or standardised.

Page no.:
Date: / /

(1) Min max Normalization.

perform linear transformation on og data.

13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72.

value (45) for age ~~[illegible]~~ to [0.0, 1.0]

Transform point = $v' = \frac{(v - min)}{(max - min)} \times (newmax - newmin) + newmin$

Given range → 0.0 to 1.0,

Data - transform - v = 45

min = 13

new min = 0

max = 72

newmax = 1

$v' = \frac{45 - 13}{72 - 13} \times (1 - 0) + 0$

Transform point. = v' = 0.5423, [0,1] range.

(2) Z-score normalization (zero score)

↳ transformation is done by values conversion to a common scale where an avg number = 0 & standard deviation = 1

$$z = \frac{x - \mu}{\sigma}$$

z score tells you how many SD from the mean your score is

Page no.:
Date: / /

ex: mean = 54000 (μ)
SD = 16000
x = salary = 73600

$$z = \frac{73600 - 54000}{16000}$$

$$z = 1.225$$

(3) Decimal scaling

A = -986 to 917

max abs (A) = 986.

→ divide by 1000.

= -0.986 to 0.917.

Discretization by Binning.

→ top-down splitting technique (bins).

Discretization by Histogram Analysis:

Same as prev.

Concept Hierarchy Generation for Nominal data:

PPT

Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - *street* < *city* < *state* < *country*
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only *street* < *city*, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {*street, city, state, country*}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year