Data Preparation for Data Mining

Lesson 4

Lesson 4 Overview

- Sampling
- Variability
- Confidence
- Numeric vs. Nominal Attributes
  - Dealing with Numeric Variables

Lesson 4 Overview

- Sampling
- Variability
- Confidence
- Numeric vs. Nominal Attributes
  - Dealing with Numeric Variables
  - *Dealing with Nominal Variables*

Purpose of Sampling

- Get enough data so that all of the relationships at all levels:
  - Superstructure
  - Macrostructure
  - Microstructure

are captured

Data Population

- All the data as the whole is called the *Data Population*
- The data is not the population
- The Data is simply a set of measurements about the population of objects

Why not use all the data?

- Problems with using all of the data
  - The whole data not available
  - Too much data
  - Necessary to sample the data when building models

Purpose of Sampling

- Capture a Sample:
  - To represent only some part of the population

Sampled Datasets

- Modeling requires at least two datasets to be sampled, and sometimes three
- Essential that each of the samples represents the full set of relationships that are present in the whole population
- If not, the model does not represent what would be found in the population

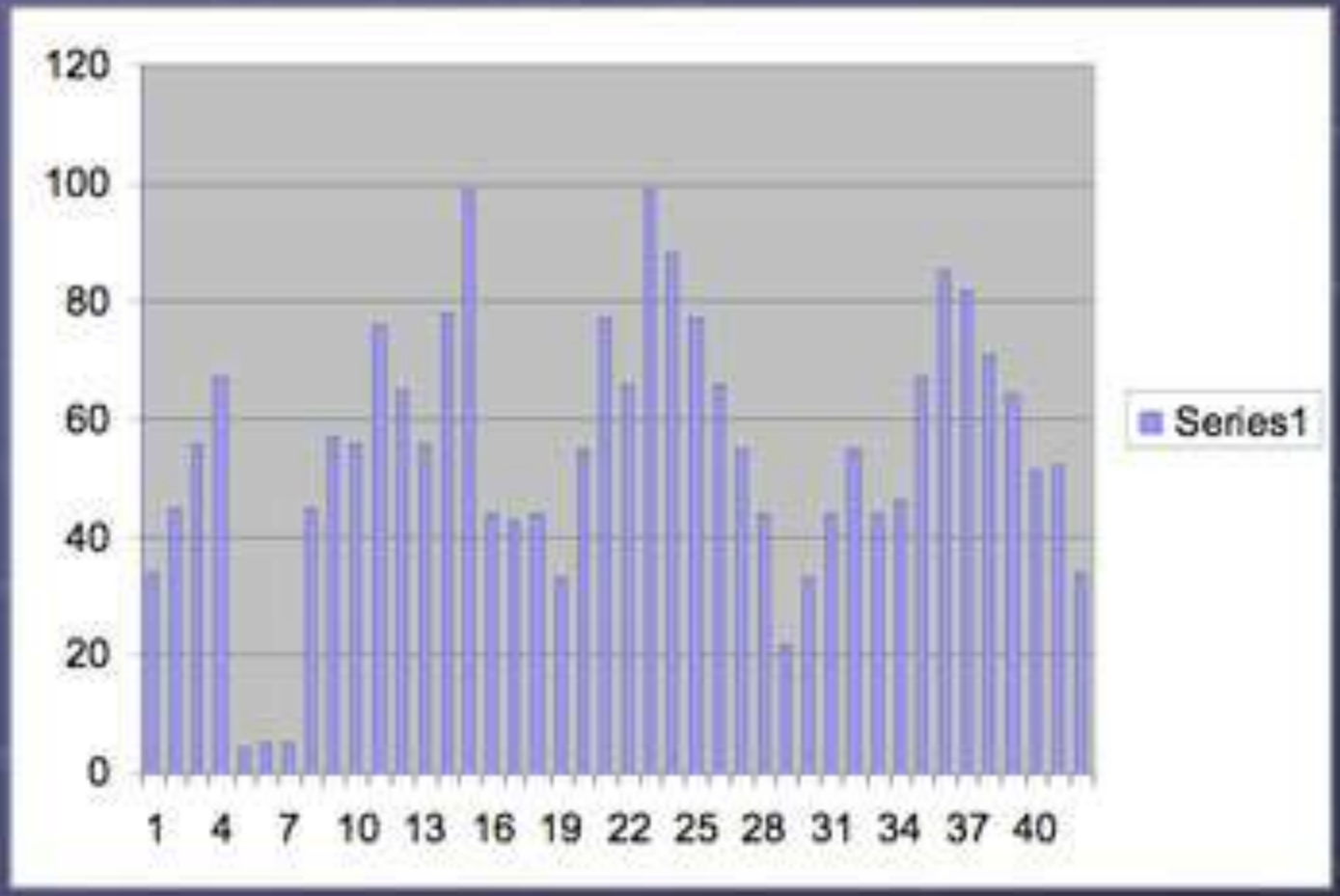
Variability of Variables

- Main Feature of a Variable
  - Takes on a variety of values
  - Contains Pattern distribution
    - Numerical variables
    - Categorical variables

Tools for Examining Variability

- Example:
  34,45,56,67,4,5,5,45,57,56,76,65,56,78,59,44,43,44,33,55,77,66,99,88,77,66,55,44,22,33,44,55,44,46,67,85,82,71,64,51,52,34, etc.
- Is there any pattern?
- Graphical Display of a Pattern Distribution
  - Histogram, Curve

Histogram



Curve

- Can be very jagged
- Smoothing
- Each smoothing method gives a slightly different curve

- Which is the correct pattern shape for that particular distribution, if any?

Issues

- Until a representative sample is obtained, it is impossible to know if the pattern in some particular random sample dies represent the “true” variability
- While it is obvious there is some sort of pattern to a distribution, various ways of looking at it seem to produce slightly different patterns

The Question

-
- Which of these shapes, if any, is the right one to use?

Variability of Variables

- Problems
  - Convergence: True Population Distribution Pattern Unknown
  - Measuring Variability: Which Distribution Curve is the Right one to use?

Converging

- To Create a Distribution Curve for the Sample
  - Selecting instance values, one at a time at random
  - Recalculated when adding a new instance value

Converging

- Converge
  - At first: a large change
  - After a while: settled down -> Converges to the final shape
- Summary
  - What is measured not the shape of the curve, but the variability of the sample

Measuring Variability

- Require Some Method of Measuring Variability
  - Without being sensitive to column width or smoothing method
- What is Variability
  - How far the individual instances from the Mean of the sample

Measuring Variability

- Standard Deviation --- One Popular Measure

- One Formula:

$$\textbf{Standard deviation} = \sqrt{\left(\sum (x-m)^2/(n-1)\right)}$$

- Another Formula: Important for data preparation process

$$\mathbf{s} = \sqrt{\left(\left(\sum x^2 - nm^2\right)/(n-1)\right)}$$

Variability of Numeric and Alpha Variables

- Why Confidence
  - An alternative of sampling the whole population
  - To establish some acceptable degree of confidence,
    - 95% as a satisfactory level of confidence

Variability of Numeric and Alpha Variables

- Distinction
  - Alpha: for nominal / categorical; measured in nonnumeric scales
  - Numeric: measured in numeric scales
  - Different when measuring variability

Measuring Variability of Numeric Variables

-
- Covered earlier
- Random sampling without introducing bias

Measuring Variability of Alpha Variables

- Instead of standard deviation
- Rate of Discovery (ROD):
  - Measure the rate of change of the relative proportion of values discovered
  - Sample size increases, the ROD of new alpha values falls

Confidence

- Measuring and testing for confidence
- Confidence in capturing variability
- Problems with sampling
- Confidence and instance count

Measuring confidence

- Required confidence level arbitrarily chosen
- Sampling vs. entire population
- Testing for confidence from a sample
- Level of confidence vs. number of tests

Confidence with entire population

- Sampling and modeling unnecessary
- Inferential modeling could be used to find interrelationships
- No training necessary - no risk of overtraining either

Confidence with entire population

- Confidence levels easy to calculate if sampling is used and population (or size of it) is known
- Otherwise, assumptions necessary to determine LOC

Testing for confidence

- If entire population is not available, we need either assumptions about
  - the randomness of the sample
  - the distribution of the data

OR

  - the success ratio of tests, assuming the tests are independent

i.e. the size of the population is not needed.

Testing for confidence (cont.)

- Assumption: LOC = ”error rate”, i.e. (1-confidence)
- No. of tests necessary to achieve desired LOC:

  $c = 1 - e^n \Rightarrow n = \log(1-c)/\log(c)$

  Note that no knowledge of the size of the population is required.

Example of repetitive tests

Skepticism	Error rate	No. of tests
0.9	0.9	1
0.81	0.9	2
0.729	0.9	3
0.6561	0.9	4
0.59049	0.9	5

Confidence in variability

- How to determine that the variability of the sample is similar to that of the population?

- convergence: if variability remains within a particular range, variability is assumed captured (to a particular level of confidence)
  - How to measure convergence?
  - How to discover the range?

Capturing variability

- Relies on normal distribution
- If variability not normally distributed, can be adjusted to resemble normal distribution
  - this relies on convergence of changes in variance around the mean

Example

- Example: CREDIT data, DAS record variability:

  95% certainty that 95% of variability captured

Problems with sampling

- Missing values
  - ignored
  - null vs. 0
  - density thresholds - to keep or not to keep?

Problems with sampling (cont.)

- Missing values
- Constants
  - not necessarily easy to spot
  - discard if found

Problems with sampling (cont.)

- Missing values
- Constants
- Representative samples?
  - problems with categorical variables

Problems with sampling (cont.)

- Missing values
- Constants
- Representative samples?
- Monotonic variables
  - detection may be difficult because of sampling
  - two methods to use for detection:
    - interstitial linearity
    - rate of discovery

Problems with sampling (cont.)

- Interstitial linearity:
  - intervals between values are evaluated
  - if spacing is consistent, monotonicity is assumed
- Rate of discover

Problems with sampling (cont.)

- Interstitial linearity:
  - intervals between values are evaluated
  - if spacing is consistent, monotonicity is assumed
- Rate of discover
  - every sample will contain a new value
  - may be legitimate, but using both characteristics together makes detection of monotonicity likely

Summary

- Deciding how much data the miner needs to make sure that variables have their variability represented
- Variability important
- Never perfect confidence
- Sampling plausible

Summary

- We can either select enough data to establish the needed level of confidence
- Or, determine how much confidence is justified in a limited dataset on hand

Summary

- Selecting the appropriate level of confidence requires:
  - Problem knowledge
  - Domain knowledge
  - Cannot be automatically determined
  - 95% works reasonably well
- Confidence decisions must be made by the problem owner, problem holder, domain expert and the miner

Next

- Assignment II

- Handling non-numeric variables
- Fix various problems in the variables

Assignment II