Computer Science 477

Data Set Quality and Sources of Data, Descriptive Statistics

Lecture 3

March 28 Exercise

You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: “It’s so simple that I can’t believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?”

(a) Who is right, the marketing director or his boss?

(b) The boss is right. A better measure is given by

$$\text{Satisfaction}(\text{product}) = \frac{\text{number of complaints for the product}}{\text{total number of sales for the product}}$$

March 28 Exercise

- What can you say about the attribute type of the original product satisfaction attribute?

- We know that two products that have the same level of customer satisfaction may have different numbers of complaints and vice-versa.

March 30 Exercise

- Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.
- Number of patients in a hospital.
  - Ratio
- ISBN numbers for books. (Look up the format on the Web.)
  - Nominal

ISBN Numbers - Digression

- Each ISBN consists of 5 elements with each section being separated by spaces or hyphens. Three of the five elements may be of varying length:
- **Prefix element** – currently this can only be either 978 or 979. It is always 3 digits in length
- **Registration group element** – this identifies the particular country, geographical region, or language area participating in the ISBN system. This element may be between 1 and 5 digits in length
- **Registrant element** - this identifies the particular publisher or imprint. This may be up to 7 digits in length
- **Publication element** – this identifies the particular edition and format of a specific title. This may be up to 6 digits in length
- **Check digit** – this is always the final single digit that mathematically validates the rest of the number. It is calculated using a Modulus 10 system with alternate weights of 1 and 3.

March 30 Exercise

- Ability to pass light in terms of the following values: opaque, translucent, transparent.
  - Ordinal (ratio?)
- Military rank.
  - Ordinal
- Distance from the center of campus.
  - Ratio
- Density of a substance in grams per cubic centimeter.
  - Ratio
- Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)
  - Nominal (probably)

Reducing the Number of Attributes

- Ever-larger storage capacity at a steadily reducing unit price has led to large numbers of attribute values being stored for every instance,
  - Information about all the purchases made by a supermarket customer for three months
  - A large amount of detailed information about every patient in a hospital
- For some datasets there can be substantially more attributes than there are instances, perhaps as many as 10 or even 100 to one.

Reducing the Number of Attributes

- Suppose 10,000 pieces of information about each supermarket customer
- Want to predict which customers will buy a new brand of dog food.
- Most attributes irrelevant.
- Best case: unnecessary computational overhead
- Worst case: may cause algorithm to give poor results
- (Do not necessarily know what is relevant or will come to be recognized as relevant in the future)
- Special Techniques: The term *feature reduction* or *dimension reduction* is generally used for this process

Noise detection without special tools

- – We may observe that some values occur an abnormally large number of times.
- Users registering for a webbased service give ‘country’ as ‘Albania’ in 10% of the time.
- A service that is particularly attractive to inhabitants of Albania?
- – If we are analyzing the results of an online survey of student satisfaction collected in 2002, may notice that the age recorded for a high proportion of the respondents was 93.
- Possible explanation: had a ‘date of birth’ field, with subfields for day, month and year
- Many of the respondents did not bother to override the default values of 01 (day), 01 (month) and 1930
- (year). A poorly designed program then converted the date of birth to an age of 93 before storing it in the database.

Noise detection without special tools

- Anomalous values such as 22654.8, 38597 and 44625.7
- Might be errors (as suggested).
- Alternatively, might be outliers, i.e. genuine values that are significantly different from the others.
- The recognition of outliers and their significance may be the key to major discoveries, especially in fields such as medicine and physics
- Care should be taken before discarding them.

Noise detection without special tools

- Another possibility is that users who registered either failed to choose from the choices in the country field,
  - Causing a (not very sensible) default value to be taken,
  - Did not wish to supply their country details and simply selected the first value in a list of options.
- In either case likely that the rest of the address data provided for those users may be suspect too.

Missing Values

- In many real-world datasets data values are not recorded for all attributes.
- Not all data meaning for all patients
  - Certain medical data may only be meaningful for female patients or patients over a certain age.
  - The best approach here may be to divide the dataset into two (or more) parts
- Missing data that should be recorded.
- Because:
  - – a malfunction of the equipment used to record the data
  - – a data collection form to which additional fields were added after some data
  - had been collected
  - – information that could not be obtained, e.g. about a hospital patient.

Missing Values

- Discard Instances
  - Missing instances might happen to be significant
- Replace by Most Frequent/Average Value
  - Nominal values?
  - Do they have an average value?

UCI Repository of Datasets

- There are a number of 'libraries' of datasets that are readily available for downloading
- Best known of these is the 'Repository' of datasets maintained by the University of California at Irvine, generally known as the 'UCI Repository'
  - http://archive.ics.uci.edu/ml/index.php
- UCI site also has links to other repositories of both datasets and programs, maintained by a variety of organizations such as the (US) National Space Science Center, the US Bureau of Census and the University of Toronto

UCI Repository

- Also hosts ‘Knowledge Discovery in Databases Archive’ at http://kdd.ics.uci.edu.

- Large and complex datasets as a challenge to the data mining research community

Data Sets Galore

- In identifying Quarter Project be (1) inquisitive, (2) ambitious
- Don't be afraid to be speculative
  - You will learn a lot, regardless of the outcome.
- Don't just settle on something!
- Preliminary progress report due next Tuesday, specifying
  - Participants
  - What you are investigating, looking for
  - Don't worry: can be subject to change.

Descriptive Statistics: Measures of Central Tendency

- Mean
  - "Average" $= \frac{\text{sum of all values}}{\text{number of distinc observations}}$
- Median
  - Observation that lies at the center of middle of a distribution when the observations are ranked in size order
    - When there are an even number of observations, the median is the average of the two ranking values
- Mode
  - Most commonly occurring observation
    - Only average that can be used with nominal data
- Geometric Mean $\sqrt[n]{X_1 \cdot X_2 \cdot \ \ldots \cdot X_n} = (X_1 \cdot X_2 \cdot \ \ldots \cdot X_n)^{1/n}$
  - Less sensitive to distortion caused by extreme values
  - But no corresponding measure of dispersion

Weighted Means

- Also call expectation
- Weighted by the number of occurrences of each value averaged

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{N}$$

- where
  - X_i is the value for group i
  - f_i is the frequency with which which those values occur
  - k is the number of groups
  - N is the number of cases from which the frequency distribution has been compiled

Weighted Mean Example

- Basic data:

Table 3.10 Return of convicts confined in Portland Prison, 1849

Name	*Age (years)*	*Offence*	*Place of committal*	*Sentence (years)*
James Hackett	21	Felony	Salford	7
John Taylor	20	Stealing a file and monies	Leicester	7
John Brown	20	Larceny (PC)	CC Court	7
James Barker	47	Stealing fowls, two indictments	Exeter	14
William Johnson	25	Setting fire to sacks of straw	Stafford	20
James Sweeney	58	Uttering counterfeit coin (PC)	Caernarvon	15
George Williams	21	Burglary (PC)	CC Court	10
Francis Best	35	Housebreaking, larceny	Worcester	15
John Henry	36	Uttering forged notes	Glasgow	20
Thomas Hartshorn	33	Robbery with violence	Liverpool	15
Samuel Laughton	22	Burglary, stealing silver spoons, etc.	Nottingham	14
Thomas Robinson	23	Burglary and theft, two indictments	Maidstone	14
Martin Stone	22	Horse stealing	Dorchester	15
Richard Ashford	58	Stealing 3lbs of pork	Exeter	10
John Dobson	28	Stealing money from the person (PC)	Stafford	14
Samuel Diggle	36	Burglary	Liverpool	15
George Goult	22	Robbery (PC)	Chelmsford	12
Robert Holder	23	Stealing from a dwelling £15 and pair of pistols	Portsmouth	15
Richard Jones	36	Warehouse breaking, and stealing malt and hops	Reading	15
Hugh King alias Cameron	36	Theft by housebreaking	Glasgow	14
Austin Montroe	34	Larceny in a dwelling to the value £5 (PC)	CC Court	15

Summarized:

Table 3.11 Simple frequency distribution of sentence lengths (unrelated to type of crime) of prisoners in Portland Prison, 1849

Sentence length (years)	*Number of prisoners*
7	3
10	2
12	1
14	5
15	8
20	2
Total	21

Source: see Table 3.10.

Weighted average sentence:

$$= \frac{(3 \times 7) + (2 \times 10) + (1 \times 12)(5 \times 14) + (8 \times 15) + (2 \times 20)}{21}$$

$$= \frac{(21 + 20 + 12 + 70 + 120 + 40)}{21}$$

$$= \frac{283}{21}$$

$$= 13.5$$

Measures of Dispersion, *et cetera*

- Read pages 97-117 of *History by Numbers*
- Measures of Dispersion
  - Variance
  - Standard Deviation
- *Z* Score
- Coefficient of Variance
- Rank Order Dispersal Measures
- Quartiles, quintiles, deciles, percentiles

- Homework 1: Due Thursday, April 6

Now

- Do problems 1 and 2
- Be prepared to discuss them