Assignment 1.1

University of San Diego

ADS 502

Dingyi Duan

Introduction to Data Mining: Exercises 1.7 – Page 21: Question #3

3. For each of the following data sets, explain whether or not data privacy is an important issue.

a. Census data collected from 1900–1950.

A: Data privacy should be an important issue for census data, even by law. Because it is sensitive information involves characteristics about people in the community.

b. IP addresses and visit times of web users who visit your website.

A: Data privacy is also an important issue for those IP addresses because they can be used to track down to individual computers for their locations.

c. Images from Earth-orbiting satellites.

A: Data privacy is important for those images. Each satellite belongs to a specific country and the images taken by those satellites can have strategically significance between countries.

d. Names and addresses of people from the telephone book.

A: Data privacy is important for the yellow book as well. If those data are misused, this can cause identity theft and to the lighter case, the people with their numbers on the book can be disturbed in their daily lives.

e. Names and email addresses collected from the Web.

A: These are extremely private. No one should be allowed to use the names and email addresses from the web without proper permissions. Names and email addresses can be considered as part of identities and use without proper permission can result in legal issues.

Exercises 2.6 – Page 105-112: Questions #2 & 16

2. 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.

Example: Age in years. **Answer:** Discrete, quantitative, ratio

a. Time in terms of AM or PM.

A: Binary (two types); Qualitative; Nominal.

Can be interval if it is considered from 1:00 AM to 2:00 AM, from one time stamp to another.

b. Brightness as measured by a light meter.

A: Continuous (can be measured); Quantitative; Ratio (0.1, 0.2, 0.3…).

If it can be measured, then it is continuous; if it can be counted, then discrete.

c. Brightness as measured by people's judgments.

A: Discrete (people can categorize different brightness, can be counted); Qualitative; Ordinal (i.e.: very bright, bright, dimed, dark, very dark).

d. Angles as measured in degrees between 0 and 360.

A:

A: Continuous (can be measured); Quantitative; Ratio

e. Bronze, Silver, and Gold medals as awarded at the Olympics.

A: Discrete (can be counted); Qualitative; Ordinal (very clear order goes from the 3rd, 2nd, 1st).

f. Height above sea level.

A: Continuous (can be measured); Quantitative; Ratio or interval (depends how it is measured, whether using a reference or from a point to a point)

g. Number of patients in a hospital.

A: Discrete (can be counted); Quantitative (does not make a difference in quality or order); Ratio.

h. ISBN numbers for books. (Look up the format on the Web.)

A: Discrete (can be counted); Qualitative; Nominal (if treat all books with no orders).

i. Ability to pass light in terms of the following values: opaque, translucent, transparent.

A: Discrete (can be counted); Qualitative; Ordinal (very clear order).

j. Military rank.

A: Discrete (can be counted); Qualitative; Ordinal (military rank structure has a very strong hierarchy).

k. Distance from the center of campus.

A: Continuous (can be measured); Quantitative; Ratio or interval (can be either: in metric system measuring or from one point to another)

l. Density of a substance in grams per cubic centimeter.

A: Continuous (can be measured); Quantitative; Ratio.

m. 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.)

A: Discrete (can be counted); Qualitative; Ordinal (can be nominal if it does not go by order).

16. Consider a document-term matrix, where is the frequency of the word (term) in the document and m is the number of documents. Consider the variable transformation that is defined by where is the

number of documents in which the term appears, which is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

a. **What is the effect of this transformation if a term occurs in one document? In every document?**

A: If it only appears in one document, then this effect is a high frequency count for the transformation; If for all/every document, then it will approach to 0.

b. **What might be the purpose of this transformation?**

A: It can help to find books with certain word(s) with high frequency in a large database.

Data Science Using Python and R: Chapter 12 - Page 184: Questions #1, 2, 3, 4, & 10

1. **What do we mean by high dimensionality in data science?**

   A: We can simply think a dataset with a large number of rows and column that the dataset is so large that makes calculation and computation very difficult.

2. **Why do we need dimension reduction methods?**

   A: To overcome the curse of dimensionality and makes data process easier for us.

3. **What does principal components replace the original set of *m* predictors with?**

   A: With a set of k < m components that are uncorrelated linear with the m components, which is dimensionality reduction (Larose, 2019).

4. **Which principal component accounts for the most variability?**

   A: The first principal component is usually the most important. It accounts for greater variability among the predictors than any other component (Larose, 2019).

10. When we use the principal components as predictors in a regression model, what value do the VIFs take? What does this indicate?

A:

The VIF for the ith predictor is given by:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 represents the R^2 value obtained by regressing x_i on the other predictor variables. Note that R_i^2 will be large when x_i is highly correlated with the other predictors, thus making VIF_i large (Larose, 2019).

Reference:

- Larose, C. D. & Larose, D. T. (2019). *Data science using Python and R*. John Wiley & Sons, Inc.
- Tan, P., Steinbach, M., Karpatne, A. & Kumar, V. (2019). *Introduction to Data Mining*. Pearson.