CSCE 566 Assignment #1, Spring 2021

Dr. Vijay Raghavan
Assigned: Feb 02, 2020
Due: Feb 13, 2020

Note:

1. *You must show all details of work for each question. Any answer without explanation will not be accepted.*
2. *Submit the answer sheet as a **single** pdf/word file (if you scan handwritten answers, convert those to pdf/word, before submission).*
3. *Make a cover with Name, CLID.*
4. *Number all pages and give an index to each question on the cover page.*
5. *Most importantly, any sort of cheating will **NOT** be tolerated. More information can be found on class Web page on cheating policy.*

Q1.

Partition the given data into 4 bins using the Equi-depth Binning method and perform smoothing according to the following methods:

[12 points]

a. smoothing by bin means,
b. smoothing by bin median,
c. smoothing by bin boundaries.

11, 13, 13, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 45, 71, 72, 73, 75.

Q2.

Reduce the given data using histogram with the following algorithms:

[12 points]

a. equi-width histogram,
b. equi-depth histogram, but use V-Optimal criterion to choose between the two options,
c. Max-Diff histogram.

You should partition the data into 4 buckets for each method.

Data: 1,1,1,2,2,3, 5, 6, 6, 9, 10, 10, 11, 12, 12, 19,19,19,19, 20, 21, 23, 24, 24, 25, 25, 25, 30, 35, 37, 38, 40,40,41,50,68,69,70,71,71,71,71.

Q3.

Object Identifier	Binary	Categorical	Ordinal	Ratio
1	Y	A	Excellent	6
2	N	C	Ordinary	15
3	Y	B	Ordinary	56
4	N	C	Good	89

A Sample data table Containing Variables of Mixed type

For the given table

i) Normalize each attribute, [8 points]

ii) Compute dissimilarity between tuples according to each individual attribute, [12 points]

iii) Compute distance between tuples by combining results of ii), [12 points]

iv) Repeat ii) and iii), assuming that for object-1, the categorical attribute is NULL. [12 points]

[Note: The reference material is section 7.2 in the 2nd edition (section 2.4, 3rd edition). The same material is section is section 8.2 in the online as well as the 1st editions of the book.]

Q4.

We are given Sample survey data, which tells us regarding what kind of films are liked by people in a club, as shown below. [12 points]

Fiction film– 22.7%
Documentary film– 58.0%
Action film – 35.2%
Newsreel film– 43.1%
Musical film–45.9%
Animation film– 51.6%

Using Naïve Prediction

i) What is the predicted class (like or dislike) for each type of film?
ii) What is the percentage error associated with each of your predictions?
iii) How the prediction error will influence deciding about more sophisticated models?

Q5.

Classify following attributes as binary, categorical or numeric. Also classify them as qualitative (nominal or ordinal) or quantitative (Interval or ratio) with reasoning. Some cases may have more than one Interpretation. [20 points]

a. Gender
b. Military rank
c. Angles measured in degrees between 0 and 360°
d. Times in terms of AM or PM
e. Temperature in Kelvin degree (absolute temperature)