studocu

PS2 sol - Assignment 2 solution

Data Mining (Siksha 'O' Anusandhan University)

Scan to open on Studocu

Studocu is not sponsored or endorsed by any college or university
Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)

Problem Solving
Assignment 2
Data Mining (CSE4052)

1. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   (a) What is the *mean* of the data? What is the *median*?
   (b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal..)
   (c) What is the *midrange* of the data?
   (d) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of the data? What is the interquartile range?

   Solution:

   a) The mean of the data $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = {}^{809}/_{27} = 30$. The median of the data is the middle value of the ordered set which is 25.

   b) Mode of data refers to the value with highest frequency among others. In this example 25 and 35 both are having the same highest frequency and hence the data is bimodal in nature.

   c) The midrange of the data is the average of the largest (70) and smallest (13) values in the data set. ${}^{(70+13)}/_{2} = 41.5$

   d) First Quartile(Q1)=((n+1)/4)th=((27+1)/4)th=7th term which is 20.It is also known as the lower quartile.

   -The second quartile or the 50th percentile or the Median is given as: Second Quartile(Q2)=((n+1)/2)th Term=25

   -The third Quartile of the 75th Percentile (Q3) is given as: Third Quartile(Q3)=(3(n+1)/4)th Term=35 also known as the upper quartile.

   -The interquartile range is calculated as: Upper Quartile - Lower Quartile=35-20=15

2. Suppose a group of 12 sales price records has been sorted as follows:
   5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three bins by each of the following methods.
(a) equal-frequency (equidepth) partitioning
(b) equal-width partitioning
(c) clustering

Solution:

This document is available free of charge on studocu

Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)

(a) equal-frequency (equidepth) partitioning
Partition the data into equidepth bins of depth 4:

Bin 1: 1: 5, 10, 11, 13 Bin 2: 15, 35, 50, 55 Bin 3: 72, 92, 204, 215

(b) equal-width partitioning
Partitioning the data into 3 equi-width bins will require the width to be $(215 - 5)/3 = 70$. We get:

Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72 Bin 2: 92 Bin 3: 204, 215

(c) clustering
Using K-means clustering to partition the data into three bins we get:

Bin 1: 5, 10, 11, 13, 15, 35 Bin 2: 50, 55, 72, 92 Bin 3: 204, 215

3. Use *smoothing by bin means, median, and boundaries* to smooth the following data, using a bin depth of 6.
Data: 11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,72,73,75
Solution:

Divide the data into bins of depth 6

bin 1:11,13,13,15,15,16 (mean=13.83, median=(13+15)/2=14)

bin 2:19,20,20,20,21,21 (mean=20.16, median=(20+20)/2)

bin3:22,23,24,30,40,45 (mean=30.67, median=(24+30)/2=27)

bin4: 45,45,71,72,73,75 (mean=63.5, median=(71+72)/2=71.5)

smoothing by means

bin 1-13.83,13.83,13.83,13.83,13.83,13.83

bin 2-20.16,20.16,20.16,20.16,20.16

bin 3-30.67,30.67,30.67,30.67,30.67,30.67

bin 4-63.5,63.5,63.5,63.5,63.5,63.5,63.5

smoothing by boundaries

bin 1:11,11,11,16,16,16

bin 2:19,19,19,21,21,21

bin3:22,22,22,22,45,45

bin4:45,45,75,75,75,75

smoothing by median

bin 1:14,14,14,14,14,14

bin 2:20,20,20,20,20,20

bin3: 27,27,27,27,27,27

Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)

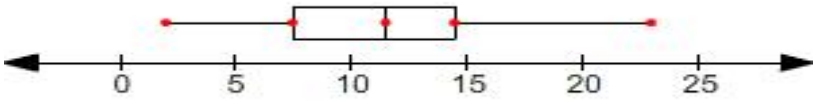
bin4:71.5,71.5,71.5,71.5,71.5,71.5

4. Find Q1, Q2, and Q3 for the following data set, and draw a box-and-whisker plot.
{2,6,7,8,8,11,12,13,14,15,22,23}

Solution: There are 12 data points. The middle two are 11 and 12. So the median, Q2, is 11.5.

The "lower half" of the data set is the set {2,6,7,8,8,11}. The median here is 7.5. So Q1=7.5.

The "upper half" of the data set is the set {12,13,14,15,22,23}. The median here is 14.5.
So Q3=14.5.
A box-and-whisker plot displays the values Q1, Q2, and Q3, along with the extreme values of the data set (2 and 23, in this case):



A box & whisker plot shows a "box" with left edge at Q1, right edge at Q3, the "middle" of the box at Q2 (the median) and the maximum and minimum as "whiskers"

5. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
(a) Compute the Euclidean distance between the two objects.
(b) Compute the Manhattan distance between the two objects.
(c) Compute the Minkowski distance between the two objects, using h = 3.

Solution: (a) Compute the Euclidean distance between the two objects.
The Euclidean distance is computed using Equation
$\sqrt{(22-20)^2+(1-0)^2+(42-36)^2+(10-8)^2}$ = $\sqrt{45}$ = 6.7082.

(b) Compute the Manhattan distance between the two objects.
The Manhattan distance is computed using Equation
|22 - 20| + |1 -0| + |42 - 36| + |10 - 8| = 11.

(c) Compute the Minkowski distance between the two objects, using h = 3.
The Minkowski disance is

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \ldots + |x_{ip}-x_{jp}|^h}$$

so,

$$d(i,j) = \sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3} = \sqrt[3]{233} = 6.1534$$

6. Use the methods below to normalize the following group of data
200, 300, 400, 600, 1000

a) min-max normalization by setting min = 0 and max = 1

This document is available free of charge on studocu

Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)

b) z-score normalization

c) z-score normalization using the mean absolute deviation instead of standard deviation

d) normalization by decimal scaling

Solution:

(a) *min-max normalization* by setting $min = 0$ and $max = 1$ get the new value by computing

$$v_i' = \frac{v_i - 200}{1000 - 200}(1 - 0) + 0.$$

The normalized data are:

$$0, 0.125, 0.25, 0.5, 1$$

(b) In *z-score normalization*, a value v_i of A is normalized to v_i' by computing

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A},$$

where

$$\bar{A} = \frac{1}{5}(200 + 300 + 400 + 600 + 1000) = 500,$$

$$\sigma_A = \sqrt{\frac{1}{5}(200^2 + 300^2 + ... + 1000^2) - \bar{A}^2} = 282.8.$$

The normalized data are:

$$-1.06, -0.707, -0.354, 0.354, 1.77$$

(c) *z-score normalization* using the *mean absolute deviation* instead of standard deviation replaces σ_A with s_A, where

$$s_A = \frac{1}{5}(|200 - 500| + |300 - 500| + ... + |1000 - 500|) = 240$$

The normalized data are:

$$-1.25, -0.833, -0.417, 0.417, 2.08$$

(d) The smallest integer j such that $Max(|\frac{v_i}{10^j}|) < 1$ is 3. After *normalization by decimal scaling*, the data become:

$$0.2, 0.3, 0.4, 0.6, 1.0$$

7. Compute the pearson correlation of the following data-

Weight (kg)	Length (cm)
3.63	53.1
3.02	49.7
3.82	48.4
3.42	54.2
3.59	54.9

Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)

Weight (kg)	Length (cm)
2.87	43.7
3.03	47.2
3.46	45.2
3.36	54.4
3.3	50.4

Solution:

$$\overline{A} = \frac{3.63 + 3.02 + 3.82 + 3.42 + 3.59 + 2.87 + 3.03 + 3.46 + 3.36 + 3.3}{10} = \frac{33.5}{10}$$
$$= 3.35$$

$$\overline{B} = \frac{53.1 + 49.7 + 48.4 + 54.2 + 54.9 + 43.7 + 47.2 + 45.2 + 54.4 + 50.4}{10} = \frac{501.2}{10}$$
$$= 50.12$$

A	B	$a_i b_i$	$(a_i - \overline{A})^2$	$(b_i - \overline{B})^2$
3.63	53.1	192.753	0.0784	8.8804
3.02	49.7	150.094	0.1089	0.1764
3.82	48.4	184.888	0.2209	2.9584
3.42	54.2	185.364	0.0049	16.6464
3.59	54.9	197.091	0.0576	22.8484
2.87	43.7	125.419	0.2304	41.2164
3.03	47.2	143.016	0.1024	8.5264
3.46	45.2	156.392	0.0121	24.2064
3.36	54.4	182.784	0.0001	18.3184
3.3	50.4	166.32	0.0025	0.0784
SUM		1684.121	0.8182	143.856

This document is available free of charge on studocu

Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)

$$r_{ab} = \frac{\sum_i^n (a_i b_i) - n\overline{A}\ \overline{B}}{n\sigma_A \sigma_B} = \frac{1684.121 - 10 * 3.35 * 50.12}{10 * \sqrt{\frac{(a_i - \overline{A})^2}{10}} \sqrt{\frac{(b_i - \overline{B})^2}{10}}} = \frac{5.101}{10 * \sqrt{\frac{0.8182}{10}} \sqrt{\frac{143.856}{10}}}$$

$$= 0.47$$

8. Perform the chi-square test for correlation for the following observation of survey where 256 peopople shared the month of their birth where the expected distribution of moths are evenly distributed.

January	29
February	24
March	22
April	19
May	21
June	18
July	19
August	20
September	23
October	18
November	20
December	23

Solution: Chi square formula is given as-

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Now, prepare the below table:

Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29				
Taurus	24				
Gemini	22				
Cancer	19				
Leo	21				
Virgo	18				
Libra	19				
Scorpio	20				
Sagittarius	23				
Capricorn	18				
Aquarius	20				
Pisces	23				

Since the expected month is evenly distributed among all the participant in the survey hence the value is $256/12 = 21.33$

So after computing all the entries the table looks like-

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

The value of chi square statistic is 5.094

This document is available free of charge on studocu

Downloaded by 531Binod Kumar (23531binod.2021@gmail.com)