CS 412, HW #2

Pouya Akbarzadeh (pa2)

March 18th, 2021

Question 1

1. **Work / Explanation**

We divide up the data into multiple bins with 3 in each (depth = 3). Usually we should order these values, however, the array is already in order, thus we just split it up into separate arrays.

Data Set 1

```
Data Set 1: [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]
Bin 1:  [13 15 16]
Bin 2:  [16 19 20]
Bin 3:  [20 21 22]
Bin 4:  [22 25 25]
Bin 5:  [25 25 30]
Bin 6:  [33 33 35]
Bin 7:  [35 35 35]
Bin 8:  [36 40 45]
Bin 9:  [46 52 70]
```

Data Set 2

```
Data Set 2:  [  5  10  11  13  15  35  50  55  72  92 204 215]
Bin 1:  [ 5 10 11]
Bin 2:  [13 15 35]
Bin 3:  [50 55 72]
Bin 4:  [ 92 204 215]
```

Then we will find the mean of each bin.

Mean of each bin for data set 1

```
Bin 1:  14.666666666666666
Bin 2:  18.333333333333332
Bin 3:  21.0
Bin 4:  24.0
Bin 5:  26.666666666666668
Bin 6:  33.666666666666664
Bin 7:  35.0
Bin 8:  40.333333333333336
Bin 9:  56.0
```

Mean of each bin for data set 2

```
Bin 1:  8.666666666666666
Bin 2:  21.0
Bin 3:  59.0
Bin 4:  170.33333333333334
```

We will replace all the values in the bin with its mean to achieve the following

Data Set 1

```
Bin 1:  [14.67 14.67 14.67]
Bin 2:  [18.33 18.33 18.33]
Bin 3:  [21. 21. 21.]
Bin 4:  [24. 24. 24.]
Bin 5:  [26.67 26.67 26.67]
Bin 6:  [33.67 33.67 33.67]
Bin 7:  [35. 35. 35.]
Bin 8:  [40.33 40.33 40.33]
Bin 9:  [56. 56. 56.]
```

Data Set 2

```
Bin 1:  [8.67 8.67 8.67]
Bin 2:  [21. 21. 21.]
Bin 3:  [59. 59. 59.]
Bin 4:  [170.33 170.33 170.33]
```

We join the bins back together then take variance

Variances

```
Variance Original Data 1:  161.29492
Variance Original Data 2:  4880.6875
Variance New Data 1:  146.71692
Variance New Data 2:  4059.8066
```

We now find the means

Means

```
DS1 Mean OG:  29.962962962962962
DS2 Mean OG:  64.75
DS1 Mean New:  29.963333333333335
DS2 Mean New:  64.75000000000001
```

Now looking at the means of each data set and comparing it with after smoothing by means to smooth, we see that we had the same means after we put all the data back together. The variance is different. Having variance is the average of the squared differences of the mean, it makes sense.

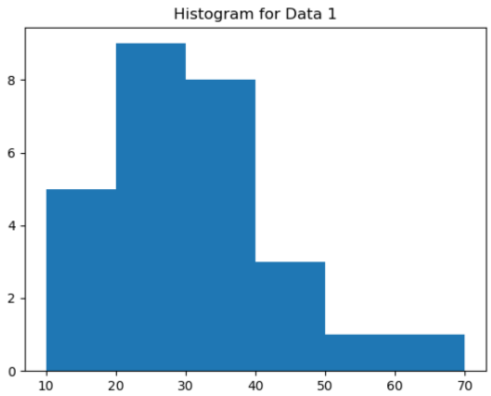
$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2,$$

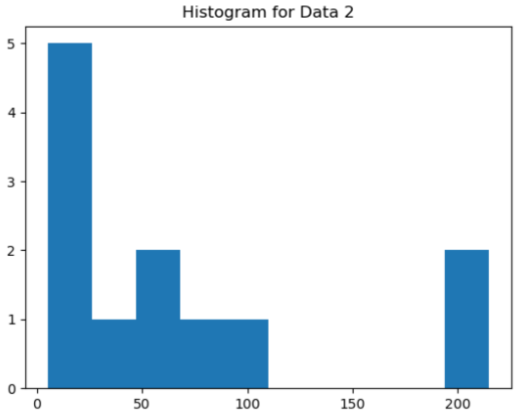
Since the variance measures how far our set of number are spread out from their average value, it makes sense how our values changed unlike the means.

Code (see file attached)

Output (see image attached)

2. **Work / Explanation**
We follow the same procedure as we did in part A, however, this time we only make 3 bins as instructed. Originally the histograms for each data set look like the following:





Here we can see that data set 1 will allow us to make a nicer data set in 3 bins. We can divide it up into 10-30, 31-50, and 51-70. However, in data set 2, it will be not as helpful. We will divide data set 2 in the following ranges 5-75, 76-145, 146-215. This range will allow all of our data points to be included with the same distant from each bin. However, it should be noted, under other circumstances with unknown data it would have been better to start the ranges from 0-X. Since we know our data here, we will proceed as mentioned before.

Every, other step now resembles part A for the equal frequency. Instead, the depth now will be different now that we have 3 bins. In the next page, you can see all intermediate steps, and values talked about. For data set 1, it seems like both methods worked great. While Equal Frequency did not perform as well as equal width in overall scheme. This makes sense because looking at our data, and histogram above, we can see how majority of data is concentrated in the 10-39 range.

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2,$$

Understanding that variance can show us how close the data is to the mean and to other data points allows us to results for data set 1. Where we see a major difference is with data set 2. Having data so spread out really put the difference of 2 methods to the test. Having our ranges the way they are it caused us to have only 1 value in 1 bin, and thus having variance of 0. However, the equal width allowed us to have a lower variance as well, mainly due to the fact that it grouped numbers close to each other.

I believe that each technique can have its own benefits. If the range of data is always known, equal width can help us a lot more, however, in some cases I can see how equal frequency can be useful too.

Data Set 1 for EQUAL FREQ
Bin 1: [13 15 16 16 19 20 20 21 22]
Bin 2: [22 25 25 25 25 30 33 33 35]
Bin 3: [35 35 35 36 40 45 46 52 70]

Data Set 2 for EQUAL FREQ
Bin 1: [5 10 11 13]
Bin 2: [15 35 50 55]
Bin 3: [72 92 204 215]

Data Set 1 for EQUAL WIDTH
Bin 1: [13 15 16 16 19 20 20 21 22 22 25 25 25 25 30]
Bin 2: [33 33 35 35 35 35 36 40 45 46]
Bin 3: [52 70]

Data Set 2 for EQUAL WIDTH
Bin 1: [5 10 11 13 15 35 50 55 72]
Bin 2: [92]
Bin 3: [204 215]

Variances of Original Data
Variance Original Data 1: 161.29492
Variance Original Data 2: 4880.6875

Variances of Each Bin for Equal Freq
Variance Equal-Freq Data 1, Bin 1: 8.444445
Variance Equal-Freq Data 1, Bin 2: 19.432098
Variance Equal-Freq Data 1, Bin 3: 118.61728
Average Variance for Equal-Freq Data 1: 48.8312733968099

Variance Equal-Freq Data 2, Bin 1: 8.6875
Variance Equal-Freq Data 2, Bin 2: 242.1875
Variance Equal-Freq Data 2, Bin 3: 4129.1875
Average Variance for Equal-Freq Data 2: 1460.0208333333333

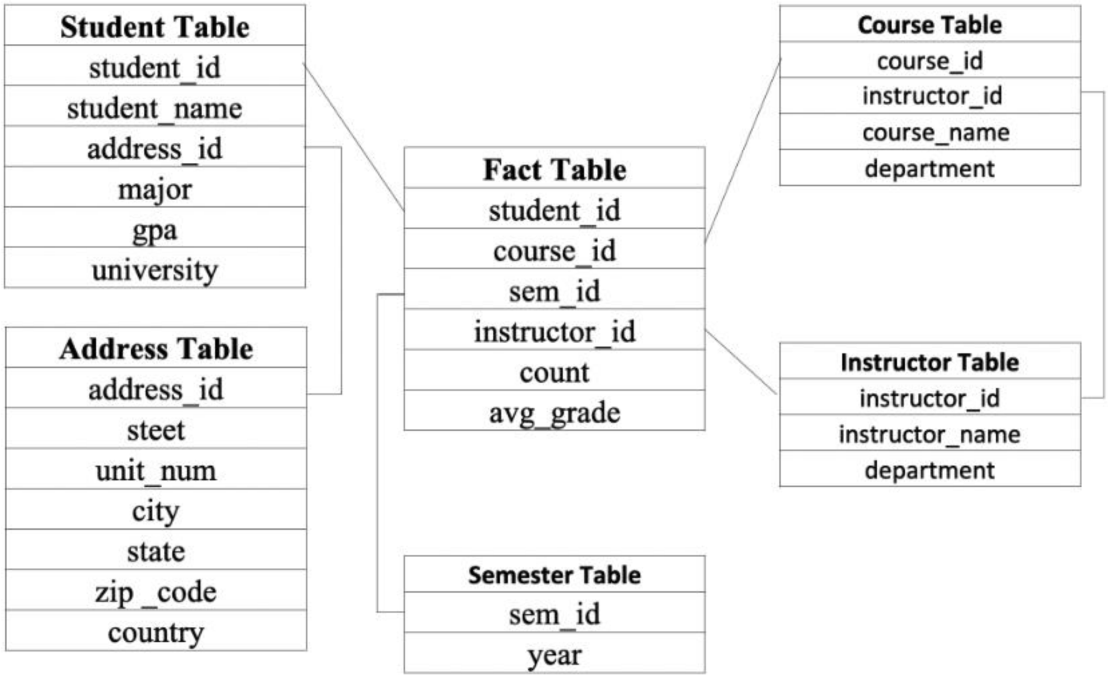
Variances of Each Bin for Equal Width
Variance Equal-Width Data 1, Bin 1: 20.195555
Variance Equal-Width Data 1, Bin 2: 20.210001
Variance Equal-Width Data 1, Bin 3: 81.0
Average Variance for Equal-Freq Data 2: 40.468518575032554

Variance Equal-Width Data 2, Bin 1: 523.58026
Variance Equal-Width Data 2, Bin 2: 0.0
Variance Equal-Width Data 2, Bin 3: 30.25
Average Variance for Equal-Freq Data 2: 184.6100870768229

Code (see file attached)

Output / Plots (see images attached)

Question 2
a)



NOTE: Here, I chose to have more meaningful names other than S1, S2, etc.
Also, I made the I made S1,Key -> S3,Key instead of S3,4

S1: student_id   S2: student_name $S3_{key}$: address_id
S4: major   S5: gpa   S6: university

$S3,1_{key}$: address_id   S3,2: street   S3,3: unit_num
S3,4: city   S3,5: state   S3,6: zip_code
S3,7: country

$C1_{key}$: course_id $C2_{key}$: instructor_id   C3: course_name
C4: department

$C2,1_{key}$ = I1: instructor_id   C2,2 = I2: instructor_name   C2,3 = I3: department

$T1_{key}$: sem_id   T2: year

b)

i. Roll-up on course from course_id to department
ii. Roll-up on student from student_id to University
iii. Dice on course, student with department set to CS and University set to Big University
iv. Drill-down on student from university to student_name

c) 5^4 = 625 Cuboids

Question 3

T1 : {a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12}
T2 : {a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20}
T3 : {a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20}
T4 : {a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a23, a24, a25, a26, a27, a28, a29, a30}

Where we define support as support(A⇒B) = P(A ∪ B)

a) We will have 4 closed patterns: T1, T2, T3, T4. We will have 1 maximal pattern and that will be T4. T4s elements are never repeated by T1-3. However, we can see that T1, T2, and T3 are all in T4.

b) We will look at each transactions and based on the definition above.

T1 : {a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12} : 3
T2 : {a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20} : 3
T3 : {a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20} : 2
T4 : {a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a23, a24, a25, a26, a27, a28, a29, a30} : 1

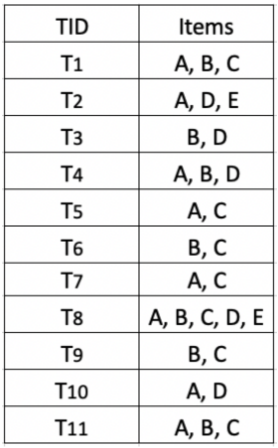
3 closed patterns here, T1, T2, and T3. 1 Closed maximal pattern, T3. We observe that T1 and T2 both have the support 3. In the other hand, T3 is repeated only within T4. Whish means, T3's support will be 2. Thus, we declare T3 as maximal pattern.

c) Here, we have 1 closed and 1 maximal pattern. T2 has a support of 4 and thus we can conclude that it is the maximal pattern.

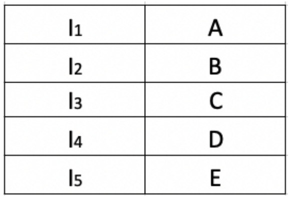
Question 4

a) Support = 4/11 = 0.36 * 100 = 36%
Confidence = 4/8 = 0.5 * 100 = 50%

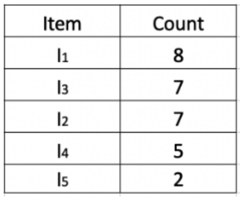
b) I first made a table containing support count of each item in our dataset. (I1-I5)



TID	Items
T1	A, B, C
T2	A, D, E
T3	B, D
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, D, E
T9	B, C
T10	A, D
T11	A, B, C

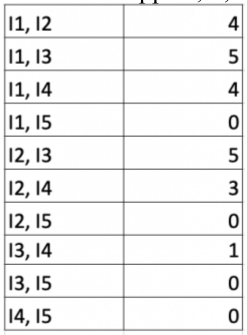


I1	A
I2	B
I3	C
I4	D
I5	E

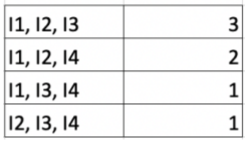


Item	Count
I1	8
I3	7
I2	7
I4	5
I5	2

Created a table containing support count of each item present in dataset. Then I compared those items support count with minimum support. (We see E / I5 has the lowest support, 3, thus we can remove it)



I1, I2	4
I1, I3	5
I1, I4	4
I1, I5	0
I2, I3	5
I2, I4	3
I2, I5	0
I3, I4	1
I3, I5	0
I4, I5	0

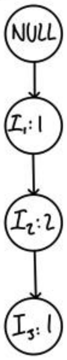


I1, I2, I3	3
I1, I2, I4	2
I1, I3, I4	1
I2, I3, I4	1

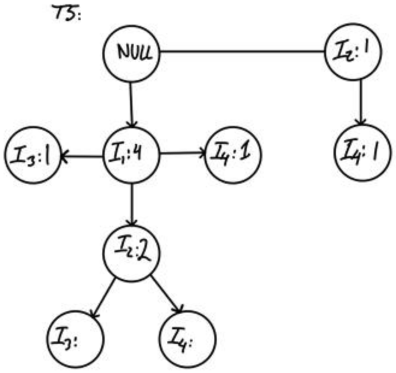
Then I joined the data and created table above. Then I found the support count of those itemset by sifting through the data. I checked the subsets and removed the ones that were not frequent. Then found support count of the remaining itemset by searching through the dataset. We then can't go further after this. We are then left with {I1, I2, I3} or simply, {A, B, C} with min support of 3.

c)

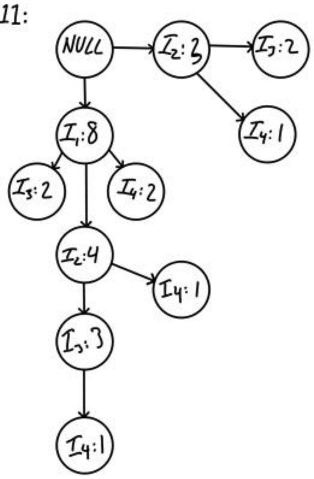
T1:



T5:



T11:



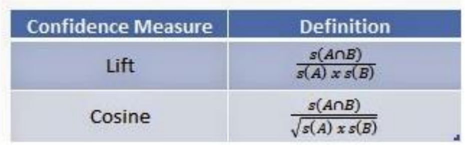
d)

Item	Conditional Pattern	FP-Tree	Freq Pattern
D / I4	{{A:2}, {A,B:1}, {A,B,C:1}, {B:1}}	<A:5>	{A,D:5}
C / I3	{{A:2}, {A,B:3}, {B:2}}	<A: 5,B:3>, <B:2>	{A,C:5}, {B,C:5}, {A,B,C:3}
B / I2	{{A:4}}	<A:4>	{A,B: 4}
A / I1			

Question 5

a) We use the following:
Kulczynski=1/2(P(A|B)+P(B|A))
Expanding it we will get
P(A|B) = P(A) x P(B) / P(B)
P(B|A) = P(B) x P(A) / P(A)
We can see that value is NULL invariant, because it can not be impacted by NULL values, and thus results in avg of the given probabilities.

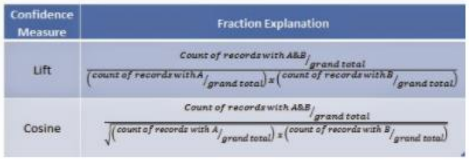
b)



Confidence Measure	Definition
Lift	$\frac{s(A \cap B)}{s(A) \times s(B)}$
Cosine	$\frac{s(A \cap B)}{\sqrt{s(A) \times s(B)}}$

The equation above is a measure of the performance of a targeting model at predicting cases as having an enhanced response measured against a random choice targeting model.

c) Looking at the table above and comparing the two equations we can see the difference is the square root in the denominator. This difference is obviously intentional. If a mistake is made in calculation of lift by not using the support as a % of the total records in that set of database this changing the number of null records. The square root allows correction due to some “math tricks”.



Confidence Measure	Fraction Explanation
Lift	$\frac{\text{Count of records with A\&B}/\text{grand total}}{(\text{count of records with A}/\text{grand total}) \times (\text{count of records with B}/\text{grand total})}$
Cosine	$\frac{\text{Count of records with A\&B}/\text{grand total}}{\sqrt{(\text{count of records with A}/\text{grand total}) \times (\text{count of records with B}/\text{grand total})}}$

Following the math above, we can see that we can cancel out the grand totals, but not the second one. We are left with AB/A*B/Grand_Total. Cosine equation fixes this using that square root. Using the square root we can cancel out the grand total from the function. Making the measure **null-invariant**.