Data Mining

FIMUS – Missing Data Imputation

Pranav Patel - 104417311

Feng Zhu - 104217106

Yuwie Liu - 104307195

Sample dataset

Record	Age	Education	Salary	Position
R1	27	MS	85	L
R2	45	-	145	P
R3	42	PhD	145	P
R4	25	MS	85	L
R5	50	PhD	146	P
R6	38	PhD	140	P
R7	-	MS	86	L

Step-1

B_{ij} = 1 if r_{ij} is missing.

0 if r_{ij} is available.

Missing Data Matrix B

Record	Age	Education	Salary	Position
R1	0	0	0	0
R2	0	1	0	0
R3	0	0	0	0
R4	0	0	0	0
R5	0	0	0	0
R6	0	0	0	0
R7	1	0	0	0

Step-2

Find generalize Dataset D_G

Convert numerical attribute column to the categorical attribute.

Here, Age and Salary are numerical attributes.

For Age,

Minimum value = 25

Maximum value = 50

Domain size = sqrt(max-min) = sqrt(50-25) = 5

So, bins are (25-29), (30-34), (35-39), (40-44), (45-49) and (50-54).

For Salary,

Minimum value = 85

Maximum value = 146

Domain size = sqrt(max-min) = sqrt(146-85) = 8

So, bins are (85-92), (93-100), (101-108), (109-116), (117-124), (125-132), (133-140) and (140-148).

Generalize Dataset D_G

Record	Age	Education	Salary	Position
R1	25-29	MS	85-92	L
R2	45-49	-	141-148	P
R3	40-44	PhD	141-148	P
R4	25-29	MS	85-92	L
R5	50-54	PhD	141-148	P
R6	35-39	PhD	133-140	P
R7	-	MS	85-92	L

<u>Step-3</u>

Co-appearance Matrix

	25-29	35-39	40-44	45-49	50-54	MS	PhD	85-92	133-140	141-148	L	P
25-29	-	-	-	-	-	2	0	2	0	0	2	0
35-39	-	-	-	-	-	0	1	0	1	0	0	1
40-44	-	-	-	-	-	0	1	0	0	1	0	1
45-49	-	-	-	-	-	0	0	0	0	1	0	1
50-54	-	-	-	-	-	0	1	0	0	1	0	1
MS	2	0	0	0	0	-	-	3	0	0	3	0
PhD	0	1	1	0	1	-	-	0	1	2	0	3
85-92	2	0	0	0	0	3	0	-	-	-	3	0
133-140	0	1	0	0	0	0	1	-	-	-	0	1
141-148	0	0	1	1	1	0	2	-	-	-	0	3
L	2	0	0	0	0	3	0	3	0	0	-	-
P	0	1	1	1	1	0	3	0	1	3	-	-

Frequency table

Category	Count
25-29	2
35-39	1
40-44	1
45-49	1
50-54	1
MS	3
PhD	3
85-92	3
133-140	1
141-148	3
L	3
P	4

Similarity Matrix

$S(x,y) = 1$ if $x = y$

$= 1/(1 + (\log(N/f(x)) * \log(N/f(y)))$ else

For S^{Age}

$S_{(25\text{-}29\,,\,25\text{-}29)} = 1$

$S_{(35\text{-}39\,,\,35\text{-}39)} = 1$

$S_{(40\text{-}44\,,\,40\text{-}44)} = 1$

$S_{(45\text{-}49\,,\,45\text{-}49)} = 1$

$S_{(50\text{-}54\,,\,50\text{-}54)} = 1$

$S_{(25\text{-}29\,,\,35\text{-}39)} = 1/(1 + (\log(7/2) * \log(7/1)) = 0.685 = S_{(35\text{-}39\,,\,25\text{-}29)}$

$S_{(25\text{-}29\,,\,40\text{-}44)} = 1/(1 + (\log(7/2) * \log(7/1)) = 0.685 = S_{(40\text{-}44\,,\,25\text{-}29)}$

$S_{(25\text{-}29\,,\,45\text{-}49)} = 1/(1 + (\log(7/2) * \log(7/1)) = 0.685 = S_{(45\text{-}49\,,\,25\text{-}29)}$

$S_{(25\text{-}29\,,\,50\text{-}54)} = 1/(1 + (\log(7/2) * \log(7/1)) = 0.685 = S_{(50\text{-}54\,,\,25\text{-}29)}$

$S_{(35\text{-}39\,,\,40\text{-}44)} = 1/(1 + (\log(7/1) * \log(7/1)) = 0.583 = S_{(40\text{-}44\,,\,35\text{-}39)}$

$S_{(35\text{-}39\,,\,45\text{-}49)} = 1/(1 + (\log(7/1) * \log(7/1)) = 0.583 = S_{(45\text{-}49\,,\,35\text{-}39)}$

$S_{(35\text{-}39\,,\,50\text{-}54)} = 1/(1 + (\log(7/1) * \log(7/1)) = 0.583 = S_{(50\text{-}54\,,\,35\text{-}39)}$

$S_{(40\text{-}44\,,\,44\text{-}49)} = 1/(1 + (\log(7/1) * \log(7/1)) = 0.583 = S_{(44\text{-}49\,,\,40\text{-}44)}$

$S_{(40\text{-}44\,,\,50\text{-}54)} = 1/(1 + (\log(7/1) * \log(7/1)) = 0.583 = S_{(50\text{-}54\,,\,40\text{-}44)}$

$S_{(45\text{-}49\,,\,50\text{-}54)} = 1/(1 + (\log(7/1) * \log(7/1)) = 0.583 = S_{(50\text{-}54\,,\,45\text{-}49)}$

	25-29	35-39	40-44	45-49	50-54
25-29	1	0.685	0.685	0.685	0.685
35-39	0.685	1	0.583	0.583	0.583
40-44	0.685	0.583	1	0.583	0.583
45-49	0.685	0.583	0.583	1	0.583
50-54	0.685	0.583	0.583	0.583	1

Normalize Similarity Matrix for Age

	25-29	35-39	40-44	45-49	50-54
25-29	1	0.685	0.685	0.685	0.685
35-39	0.685	1	0.583	0.583	0.583
40-44	0.685	0.583	1	0.583	0.583
45-49	0.685	0.583	0.583	1	0.583
50-54	0.685	0.583	0.583	0.583	1

Same procedure for similarity analysis for Education, Salary and Position

Normalize Similarity Matrix for Education S^{Edu}

	MS	PhD
MS	0.532	0.468
PhD	0.468	0.532

Normalize Similarity Matrix for Salary S^{Salary}

	85-92	133-140	141-148
85-92	0.378	0.289	0.333
133-140	0.302	0.396	0.302
141-148	0.333	0.289	0.378

Normalize Similarity Matrix for Position $S^{Position}$

	MS	PhD
MS	0.521	0.479
PhD	0.479	0.521

Co-relation Matrix

Pearson Contingency Co-efficient = $\sqrt{\frac{T}{N+T}}$

Where T = Chi- square

For, $K_{age, age}$ = -

$K_{edu, edu}$ = -

$K_{sal, sal}$ = -

$K_{pos, pos} = -$

For $K_{age, edu}$

	MS	PhD	Total
25-29	2	0	2
35-39	0	1	1
40-44	0	1	1
45-49	0	0	0
50-54	0	1	1
Total	2	3	5

T = 1.8 + 1.2 + 0.4 + 0.267 + 0 + 0 + 0.4 + 0.267 = 5. 001

$K_{age, edu} = 0.645$

For $K_{age, sal}$

	85-92	133-140	141-148	Total
25-29	2	0	0	2
35-39	0	1	0	1
40-44	0	0	1	1
45-49	0	0	1	1
50-54	0	0	1	1
Total	2	1	3	6

T = 2.64 + 0.33 + 1 + 0.33 + 4.88 + 0.5 + 0.33 + 0.17 + 0.5 + 0.33 + 0.17 + 0.5 + 0.33 + 0.17 + 0.5

= 12.68

$K_{age, sal} = 0.802$

For $K_{age, pos}$

	L	P	Total
25-29	2	0	2
35-39	0	1	1
40-44	0	1	1
45-49	0	1	1
50-54	0	1	1
Total	2	4	6

T = 2.64 + 1.33 + 0.33 + 0.49 + 0.33 + 0.49 + 0.33 + 0.49 + 0.33 + 0.49 = 7.25

$K_{age, pos} = 0.713$

For $K_{edu, sal}$

	85-92	133-140	141-148	Total
MS	3	0	0	3
PhD	0	1	2	3
Total	3	1	2	6

T = 1.5 + 0.5 + 0.33 + 1.5 + 0.5 + 8.45 = 12.78

$K_{edu, sal} = 0.803$

For $K_{edu, pos}$

	L	P	Total
MS	3	0	3
PhD	0	3	3
Total	3	3	6

T = 1.5 + 1.5 + 1.5 + 1.5 = 6

$K_{edu, pos} = 0.707$

For $K_{sal, pos}$

	L	P	Total
85-92	3	0	3
133-140	0	1	1
141-148	0	3	3
Total	3	4	7

T = 2.26 + 1.71 + 0.43 + 0.324 + 1.29 + 0.97 = 6.987

$K_{sal, pos} = 0.733$

Co-relation Matrix K

	Age	Education	Salary	Position
Age	-	0.645	0.802	0.713
Education	0.645	-	0.803	0.707
Salary	0.802	0.803	-	0.733
Position	0.713	0.707	0.733	-

Step-4

Missing data imputation

	Age	Education	Salary	Position
R2	45-49	?	141-148	P
R7	?	MS	85-92	L

Input : Record r_i, Attribute index j, Attribute list A, Co-appearance matrix C, Set of Similarity matrices S, Correlation matrix K, threshold λ, and data set D_G

Output : Imputed value r_{ij}

Step 1:

- **foreach** *category* $x \in A_j$ **do**
  - $V_x^T = 0$; /*V_x^T is the total voting in favour of x*/
  - /* Loop over all attributes in A excluding the jth attribute*/
  - **foreach** *attribute* $A_p \in A \backslash A_j$ **do**
    - $V_x^N = 0$, $V_x^S = 0$, and $l = r_{ip}$; /*notations are introduced in Section 3.1*/
    - **if** $\lambda > 0$ **then**
      - $V_x^N = \frac{C_{xl}}{f_l}$;
    - **end**
    - **if** $\lambda < 1$ **then**
      - **foreach** *category* $a \in A_p$ **do**
        - $H = \frac{C_{xa}}{f_a}$;
        - $V_x^S = V_x^S + H \times S_{la}^p$; /*see Section 3.1*/
      - **end**
    - **end**
    - $V_x^p = \{V_x^N \times \lambda + V_x^S \times (1 - \lambda)\} \times k_{jp}$;/*$k_{jp} \in K$ is the correlation between the jth and pth attribute */
    - $V_x^T = V_x^T + V_x^p$;
  - **end**
- **end**

end

Step 2:

- Set $r_{ij} \leftarrow Value(max(V_x^T; \forall x \in A_j))$; /*finds the attribute value x for which V_x^T is the Max*/

end

Step 3:

- Return imputed value r_{ij};

end

Calculation:

For R2

Possible candidates are : MS and PhD

For x = MS

Ap = Age

$$V_{MS}^{N} = \frac{C\ (\mathrm{MS}, 45-49)}{f\ (45-49)} = 0$$

Now for each categories of Age

1. a = 25-29

$$H = \frac{C\ (\mathrm{MS}, 25-29)}{f\ (25-29)} = 1$$

$$V_{MS}^{S} = 0 + 1 * S^{age}_{45\text{-}49,\ 25\text{-}29} = 0.2$$

2. a = 35-39

$$H = \frac{C\ (\mathrm{MS},35-39)}{f\ (35-39)} = 0$$

$V_{MS}{}^{S} = 0.2 + 0 * S^{age}{}_{35\text{-}39,\ 45\text{-}49} = 0.2$

3. a = 40-44

$$H = \frac{C\ (\mathrm{MS},40-44)}{f\ (40-44)} = 0$$

$V_{MS}{}^{S} = 0.2 + 0 * S^{age}{}_{40\text{-}44,\ 40\text{-}44} = 0.2$

4. a = 45-49

$$H = \frac{C\ (\mathrm{MS},44-49)}{f\ (44-49)} = 0$$

$V_{MS}{}^{S} = 0.2 + 0 * S^{age}{}_{40\text{-}44,\ 45\text{-}49} = 0.2$

5. a = 50-54

$$H = \frac{C\ (\mathrm{MS},50-54)}{f\ (141-148)} = 0/$$

$V_{MS}{}^{S} = 0.2 + 0 * S^{age}{}_{40\text{-}44,\ 50\text{-}54} = 0.2$

$V_{MS}{}^{age} = \{\ V_{MS}{}^{N}\ \lambda + V_{MS}{}^{S}(\ 1\text{-}\ \lambda)\ \}\ K_{edu,age}$

$= 0.01032$

Now for Ap=salary

$$V_{MS}{}^{N} = \frac{C(\mathrm{MS}-141-148)}{f\ (45-49)} = 0/3=0$$

Now for each category of salary

a = 85-92

$$H = \frac{C\ (\mathrm{MS},85-92)}{f\ (85-92)} = 3/3=1$$

$V_{MS}{}^{s} = 1 * \mathrm{S\ salary}(141-148)(85-92)\ {}^{=0.333}$

a = 133-140

$$H = \frac{C\ (\mathrm{MS},133-140)}{f\ (133-140)} =0$$

$V_{MS}{}^{s} =0.333$

$V_{MS}^{salary} = \{ V_{MS}^{N} \lambda + V_{MS}^{S}(1- \lambda) \} K_{edu,salary}$

$= (0*0.2+0.333(11-0.2)\} 0.803$

$= 0.2139$

For Ap = Position

$V_{MS}^{N} = \frac{C(\mathrm{MS-P})}{f\ (P)} = 0/4=0$

Now for each categories of position

A=L

$H = \frac{C\ (\mathrm{MS-L})}{f\ (L)} = 3/3=1$

$V_x^S = 1 * S^{pos}{}_{(L-P)}= 0.521$

A=P

$H = \frac{C\ (\mathrm{MS-P})}{f\ (P)} =0$

$V_x^S = 0.521+0=0.521$

$V_x^{pos} = \{ V_x^N \lambda + V_x^S(1- \lambda) \} K_{edu,pos}$

$=(0*0.02*0.521(1-o.2)\}* 0.707$

$=0.2946$

So, total $V_{MS}^{T} = V_{MS}^{Age} + V_{MS}^{Salary} + V_{MS}^{Pos}$

$=0.0103+0.2139+0.2946=$

$=0.5188$

Now same way we will find V_{PhD}^{T}

X=PhD

for Ap = age

$V_{PhD}^{N} = \frac{C(\mathrm{PhD-45-49})}{f\ (45-49)} = 0/1=0$

Now for each categories for age

A = 25-29

$H = \frac{C\ (\mathrm{PhD-25-29})}{f\ (25-29)} =0$

$V_x^S = 0$

A = 35-39

$H = \frac{C\ (\text{PhD}-35-39)}{f\ (35-39)}$ =1/1=1

$V_x^S = 0 + H^*\ S^{Age}{}_{35-39,45-41} = 1^*0.17 = 0.17$

A=40-44

$H = \frac{C\ (\text{PhD}-40-44)}{f\ (40-44)}$ =1/1=1

$V_x^S = 0.17+(1^*0.17) = 0.34$

A=44-49

$H = \frac{C\ (\text{PhD}-44-49)}{f\ (44-49)}$ =0

$V_x^S = 0.34$

A=50-54

$H = \frac{C\ (\text{PhD}-50-54)}{f\ (50-54)}$ =1

$V_x^S = 0.34+1(0.17) = 0.51$

$V_{PhD}^{age} = \{\ V_{PhD}^N\ \lambda + V_{PhD}^S(\ 1-\ \lambda)\ \}\ K_{edu,age}$

=(0*0.2+0.51(1-0.2)}*0.645

=0.26316

For Ap=Salary

$V_{PhD}^N = \frac{C(\text{PhD}-141-148)}{f\ (141-148)}$ = 2/3=0.67

For each categories of salary

A = 85-92

$H = \frac{C\ (\text{PhD}-85-92)}{f\ (85-92)}$ =0

$V_{PhD}^S = 0$

A = 133-140

$H = \frac{C\ (\text{PhD}-133-140)}{f\ (133-140)}$ =1/1=1

$V_{PhD}^{S} = 0+1* S^{Sal}_{(141-148,133-140)} = 0+1(0.289) = 0.289$

A = 141-148

H=0.67

$V_{PhD}^{S} = 0.289+(0.67*0.378) = 0.5422$

$V_{PhD}^{Salary} = (0*0.21+(0.5422*(1-0.2))*0.803$

=0.3483

For Ap = Pos

$V_{PhD}^{N} = \frac{C(\text{PhD}-\text{P})}{f\ (\text{P})} = 3/4=0.75$

For each categories of salary

A = L

$H = \frac{C\ (\text{PhD}-\text{L})}{f\ (\text{L})} =0$

$V_{PhD}^{S} = 0$

A = P

$H = \frac{C\ (\text{PhD}-\text{P})}{f\ (\text{P})} =0$

$V_{PhD}^{S} = 0.75$

$V_{PhD}^{S} = 0+(0.75*0.521) = 0.3908$

$V_{PhD}^{pos} = (0* 0.2)+(0.3908*(1-0.2)0.707 \quad = 0.2210$

So, total $V_{PhD}^{T} = V_{PhD}^{Age} + V_{PhD}^{Salary} + V_{PhD}^{Pos}$

=0.2631+0.3483+0.2210

=0.8325

Here $V_{PhD}^{T} > V_{MS}^{T}$

So, winner is PhD

R2	45-49	PhD	141-148	P

By following same procedure, we can imput missing data for other element

For, R7, missing value imputed is 25-29.This is derived by same process as done for R2."25-29" is categorical label. So we have to find numerical value between 25 and 29. For this, we have to repeat steps 3-4.

Step-4

Repeat step-3

	MS	PDD	85-92	133-140	141-148	L	P
25	1	0	1	0	0	1	0
26	-	-	-	-	-	-	-
27	1	0	1	0	0	1	0
28	-	-	-	-	-	-	-
29	-	-	-	-	-	-	-

Co-appearance matrix for only 25 & 27

	25	27	MS	PHD	85-92	133-140	141-148	L	P
25	-	-	1	0	1	0	0	1	0
27	-	-	1	0	1	0	0	1	0
MS	1	1	-	-	2	0	0	2	0
PHD	0	0	-	-	0	0	0	0	0
85-92	1	1	2	0	-	-	-	2	0
133-140	0	0	0	0	-	-	-	0	0
141-148	0	0	0	0	-	-	-	0	0
L	1	1	2	0	2	0	0	2	0
P	0	0	0	0	0	0	0	0	0

Frequency

25-1

27-1

MS-2

PhD-0

85-92-2

133-140-0

141-148-0

L-2

P-0

Similarity Analysis

S^{Age}

	25	27
25	1	0.917
27	0.917	1

Normalize S^{Age}

	25	27
25	0.522	0.478
27	0.478	0.522

For Education

S^{Edu}	**MS**	**PHD**
MS	1	0
PHD	0	1

For Salary

S^{Sal}	**85-92**	**133-140**	**141-148**
85-92	1	0	0
133-140	0	1	0
141-148	0	0	1

For Position

S^{Pos}	**L**	**P**
L	1	0
P	0	1

Corelation Matrix : K

	Age	**Edu**	**Sal**	**Pos**
Age	1	0	0	0
Edu	0	1	0	0
Sal	0	0	1	0
Pos	0	0	0	1

Repeat step 4

Total vote for 25

Ap= A Education

$V_{25}^{N} = \frac{C\ (25-\mathrm{MS})}{f\ (\mathrm{MS})}$ =1/2 =0.5

Now for each category for age

A=MS

$H= \frac{C\ (25-\mathrm{MS})}{f\ (\mathrm{MS})}$ =0.5

$V_X^S = 0.5 * S^{Edu}_{(MS-MS)}$ =0.5

A=PhD

$H= \frac{C\ (25-\mathrm{PhD})}{f\ (\mathrm{PhD})}$ =0

Repeat these steps and find total vote V_X^T for V_{25}^S and V_{27}^S

Final Imputed Missing data is 25 by results.

Step-6

Impute Missing Data into table data

Rec	Age	Edu	Sal	Pos
R1	27	MS	85	L
R2	45	**PHD**	145	P
R3	42	PHD	145	P
R4	25	MS	85	L
R5	50	PHD	146	P
R6	38	PHD	140	P
R7	**25**	MS	86	L

$NRMS= \frac{\mathrm{ll\ Xestimate - Xoriginal\ ll\ F}}{\mathrm{ll\ Xoriginal\ ll\ F}}$

Where $\mathrm{llAll} = \sum_{i=1}^{m} \sum_{j=1}^{n} IIaijII2$

Here, $X^{estimate} - X^{original} = 0$

Because all the imputed missing data are similar to original dataset.

NRMS = 0

And AE Table

$AE = 1/n \sum_{i=1}^{n} I(x = xi)$

Where I(.) stands for function that returns 1 if the estimated value x and real value xi are same, but otherwise 0.

AE Table

1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

AE = 1.00