

CS685: DATA MINING DATA DISCRETIZATION

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
<http://web.cse.iitk.ac.in/~cs685/>

1st semester, 2021-22
Mon 1030-1200 (online)

Data Discretization

- Transformation from continuous domain to discrete intervals, i.e., from quantitative data to qualitative data
- Reduces number of possible values of an attribute
- May be encoded in a more compact form

Data Discretization

- Transformation from continuous domain to discrete intervals, i.e., from quantitative data to qualitative data
- Reduces number of possible values of an attribute
- May be encoded in a more compact form
- If discretization uses information about groups, then it is called supervised discretization
- Otherwise unsupervised discretization

Data Discretization

- Transformation from continuous domain to discrete intervals, i.e., from quantitative data to qualitative data
- Reduces number of possible values of an attribute
- May be encoded in a more compact form
- If discretization uses information about groups, then it is called **supervised discretization**
- Otherwise **unsupervised discretization**
- **Top-down discretization:** repeatedly finds **split points** or **cut points** to divide the data
- **Bottom-up discretization:** considers all continuous values as split points and then repeatedly merges

Data Discretization

- Transformation from continuous domain to discrete intervals, i.e., from quantitative data to qualitative data
- Reduces number of possible values of an attribute
- May be encoded in a more compact form
- If discretization uses information about groups, then it is called **supervised discretization**
- Otherwise **unsupervised discretization**
- **Top-down discretization**: repeatedly finds **split points** or **cut points** to divide the data
- **Bottom-up discretization**: considers all continuous values as split points and then repeatedly merges
- Discretization often leads to **conceptualization** where each distinct group represents a particular concept

Concept Hierarchy

- Discrete intervals or concepts can be generated at multiple levels
- **Concept hierarchy** captures more basic concepts towards the top and finer details towards the leaves
- Example
 - Higher level: dark, medium, light
 - Middle level: Gray value > 160 , $80 - 160$, < 80
 - Lower level: actual gray value
- Can be used to represent knowledge hierarchies as well
- **Ontology**
- Example
 - **WordNet**: ontology for English words
 - **Gene Ontology**: three separate ontologies that capture different attributes of a gene

Methods of Discretization

- For numeric data
 - Binning and histogram analysis
 - Entropy-based discretization
 - Chi-square merging
 - Clustering
 - Intuitive partitioning

Entropy

Entropy

- Entropy captures the amount of randomness or chaos in the data

Entropy

- Entropy captures the amount of randomness or chaos in the data
- Entropy encodes the number of bits required to describe a distribution
- It is also called the information content

$$\text{entropy}(D) = - \sum_{i=1}^n (p_i \log_2 p_i)$$

Entropy

- Entropy captures the amount of randomness or chaos in the data
- Entropy encodes the number of bits required to describe a distribution
- It is also called the **information content**

$$\text{entropy}(D) = - \sum_{i=1}^n (p_i \log_2 p_i)$$

- When is entropy minimum?

Entropy

- Entropy captures the amount of randomness or chaos in the data
- Entropy encodes the number of bits required to describe a distribution
- It is also called the **information content**

$$\text{entropy}(D) = - \sum_{i=1}^n (p_i \log_2 p_i)$$

- When is entropy minimum?
 - When one p_i is 1 (others are 0)
 - Least random

Entropy

- Entropy captures the amount of randomness or chaos in the data
- Entropy encodes the number of bits required to describe a distribution
- It is also called the **information content**

$$\text{entropy}(D) = - \sum_{i=1}^n (p_i \log_2 p_i)$$

- When is entropy minimum?
 - When one p_i is 1 (others are 0)
 - Least random
 - Entropy = 0

Entropy

- Entropy captures the amount of randomness or chaos in the data
- Entropy encodes the number of bits required to describe a distribution
- It is also called the **information content**

$$\text{entropy}(D) = - \sum_{i=1}^n (p_i \log_2 p_i)$$

- When is entropy minimum?
 - When one p_i is 1 (others are 0)
 - Least random
 - Entropy = 0
- When is entropy maximum?

Entropy

- Entropy captures the amount of randomness or chaos in the data
- Entropy encodes the number of bits required to describe a distribution
- It is also called the **information content**

$$\text{entropy}(D) = - \sum_{i=1}^n (p_i \log_2 p_i)$$

- When is entropy minimum?
 - When one p_i is 1 (others are 0)
 - Least random
 - Entropy = 0
- When is entropy maximum?
 - When every p_i is $1/n$
 - Most random

Entropy

- Entropy captures the amount of randomness or chaos in the data
- Entropy encodes the number of bits required to describe a distribution
- It is also called the **information content**

$$\text{entropy}(D) = - \sum_{i=1}^n (p_i \log_2 p_i)$$

- When is entropy minimum?
 - When one p_i is 1 (others are 0)
 - Least random
 - Entropy = 0
- When is entropy maximum?
 - When every p_i is $1/n$
 - Most random
 - Entropy = $\log_2 n$
 - For $n = 2$, it is 1

Entropy-based Discretization

- Supervised
- Top-down

Entropy-based Discretization

- Supervised
- Top-down
- For attribute A , choose n partitions D_1, D_2, \dots, D_n for the dataset D
- If D_i has instances from m classes C_1, C_2, \dots, C_m , then **entropy** of partition D_i is

$$\text{entropy}(D_i) = - \sum_{j=1}^m (p_{j_i} \log_2 p_{j_i})$$

where p_{j_i} is *probability* of class C_j in partition D_i

$$p_{j_i} = \frac{|C_j \cap D_i|}{|D_i|}$$

Choosing Partitions

- Choose $n - 1$ partition values s_1, s_2, \dots, s_{n-1} such that $\forall v \in D_i, s_{i-1} < v \leq s_i$
 - Implicitly, s_0 is minimum and s_n is maximum
- How to choose these partition values?

Choosing Partitions

- Choose $n - 1$ partition values s_1, s_2, \dots, s_{n-1} such that $\forall v \in D_i, s_{i-1} < v \leq s_i$
 - Implicitly, s_0 is minimum and s_n is maximum
- How to choose these partition values?
- Information gain and expected information requirement

Information Gain

- When D is split into n partitions, the **expected information requirement** is defined as

$$info(D) = \sum_{i=1}^n \left(\frac{|D_i|}{|D|} entropy(D_i) \right)$$

Information Gain

- When D is split into n partitions, the **expected information requirement** is defined as

$$info(D) = \sum_{i=1}^n \left(\frac{|D_i|}{|D|} entropy(D_i) \right)$$

- The **information gain** obtained by this partitioning is defined as

$$gain(D) = entropy(D) - info(D)$$

Information Gain

- When D is split into n partitions, the **expected information requirement** is defined as

$$info(D) = \sum_{i=1}^n \left(\frac{|D_i|}{|D|} entropy(D_i) \right)$$

- The **information gain** obtained by this partitioning is defined as

$$gain(D) = entropy(D) - info(D)$$

- Choose partition values s_1, s_2, \dots, s_{n-1} such that information gain is *maximized* (equivalently, expected information requirement is *minimized*)

Information Gain

- When D is split into n partitions, the **expected information requirement** is defined as

$$info(D) = \sum_{i=1}^n \left(\frac{|D_i|}{|D|} entropy(D_i) \right)$$

- The **information gain** obtained by this partitioning is defined as

$$gain(D) = entropy(D) - info(D)$$

- Choose partition values s_1, s_2, \dots, s_{n-1} such that information gain is *maximized* (equivalently, expected information requirement is *minimized*)
- Keep on partitioning into two parts
- Stopping criterion

Information Gain

- When D is split into n partitions, the **expected information requirement** is defined as

$$info(D) = \sum_{i=1}^n \left(\frac{|D_i|}{|D|} entropy(D_i) \right)$$

- The **information gain** obtained by this partitioning is defined as

$$gain(D) = entropy(D) - info(D)$$

- Choose partition values s_1, s_2, \dots, s_{n-1} such that information gain is *maximized* (equivalently, expected information requirement is *minimized*)
- Keep on partitioning into two parts
- Stopping criterion
 - Number of categories greater than a threshold
 - Expected information gain is below a threshold

Example

- Dataset D consists of two classes 1 and 2

Class 1	10, 14, 22, 28
Class 2	26, 28, 34, 36, 38

- Probabilities of two classes are $p_1 = 4/9$ and $p_2 = 5/9$

$$\text{entropy}(D) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 = 0.99$$

- How to choose a splitting point?

Splitting Point 1

- Suppose splitting point $s = 24$

Class 1	10, 14, 22
Class 2	

Class 1	28
Class 2	26, 28, 34, 36, 38

- Then, probabilities of classes per partition are

$$p_{1_1} = 3/3 \quad p_{2_1} = 0/3 \quad p_{1_2} = 1/6 \quad p_{2_2} = 5/6$$

- Entropies are

$$\text{entropy}(D_1) = -(3/3) \log_2(3/3) - (0/3) \log_2(0/3) = 0.00$$

$$\text{entropy}(D_2) = -(1/6) \log_2(1/6) - (5/6) \log_2(5/6) = 0.65$$

- Expected information requirement and entropy gain are

$$\begin{aligned} \text{info}(D) &= (|D_1|/|D|) \text{entropy}(D_1) + (|D_2|/|D|) \text{entropy}(D_2) \\ &= (3/9) \times 0.00 + (6/9) \times 0.65 = 0.43 \end{aligned}$$

$$\begin{aligned} \text{gain}(D) &= \text{entropy}(D) - \text{info}(D) \\ &= 0.99 - 0.43 = 0.56 \end{aligned}$$

Splitting Point 2

- Suppose splitting point $s = 31$

Class 1	10, 14, 22, 28
Class 2	26, 28

Class 1	
Class 2	34, 36, 38

Splitting Point 2

- Suppose splitting point $s = 31$

Class 1	10, 14, 22, 28
Class 2	26, 28

Class 1	
Class 2	34, 36, 38

- Then, probabilities of classes per partition are

$$p_{1_1} = 4/6 \quad p_{2_1} = 2/6 \quad p_{1_2} = 0/3 \quad p_{2_2} = 3/3$$

- Entropies are

$$\text{entropy}(D_1) = -(4/6) \log_2(4/6) - (2/6) \log_2(2/6) = 0.92$$

$$\text{entropy}(D_2) = -(0/3) \log_2(0/3) - (3/3) \log_2(3/3) = 0.00$$

- Expected information requirement and entropy gain are

$$\begin{aligned} \text{info}(D) &= (|D_1|/|D|)\text{entropy}(D_1) + (|D_2|/|D|)\text{entropy}(D_2) \\ &= (6/9) \times 0.92 + (3/9) \times 0.00 = 0.61 \end{aligned}$$

$$\begin{aligned} \text{gain}(D) &= \text{entropy}(D) - \text{info}(D) \\ &= 0.99 - 0.61 = 0.38 \end{aligned}$$

Example (contd.)

- So, 24 is a better splitting point than 31
- What is the optimal splitting point?

Example (contd.)

- So, 24 is a better splitting point than 31
- What is the optimal splitting point?
 - Exhaustive algorithm
 - Tests all possible $n - 1$ partitions

Chi-Square Test

- Statistical test
- Pearson's chi-square test
- Uses chi-square statistic and chi-square distribution

Chi-Square Test

- Statistical test
- Pearson's chi-square test
- Uses chi-square statistic and chi-square distribution
- Chi-square statistic

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency, E_i is the expected frequency and k is the number of possible outcomes

- Chi-square statistic *asymptotically* approaches the chi-square distribution
- Chi-square distribution is characterized by a single parameter: degrees of freedom
- Here, degrees of freedom is $k - 1$

Chi-square Test for Independence

- Two distributions with k frequencies each
- Chi-square statistic

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Degrees of freedom is $(2 - 1) \times (k - 1)$

Chi-square Test for Independence

- Two distributions with k frequencies each
- Chi-square statistic

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Degrees of freedom is $(2 - 1) \times (k - 1)$
- The value of the statistic is compared against the chi-square distribution with $df = k - 1$
- Choose a **significance level**, say, 0.05
- If the statistic obtained is *less* than the theoretical level, then conclude that the two distributions are *independent* at the chosen level of significance

Chi-square Test for Independence

- Two distributions with k frequencies each
- Chi-square statistic

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Degrees of freedom is $(2 - 1) \times (k - 1)$
- The value of the statistic is compared against the chi-square distribution with $df = k - 1$
- Choose a **significance level**, say, 0.05
- If the statistic obtained is *less* than the theoretical level, then conclude that the two distributions are *independent* at the chosen level of significance
- Lower chi-square corresponds to higher **p-value**
- **Null hypothesis** is that the two distributions are independent

Test for Independence: Example

- Null hypothesis H_0 : Variables x, y are *independent*
 - Alternate hypothesis H_1 : They are *not* independent

Test for Independence: Example

- Null hypothesis H_0 : Variables x, y are *independent*
 - Alternate hypothesis H_1 : They are *not* independent
- Does riding cycles depend on age?
- Contingency table

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

Test for Independence: Example

- Null hypothesis H_0 : Variables x, y are *independent*
 - Alternate hypothesis H_1 : They are *not* independent
- Does riding cycles depend on age?
- Contingency table

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

- If H_0 is true, each cell $c_{i,j}$ is a product of $x_{i..}y_{.j}$, i.e., the **marginals**

Test for Independence: Example

- Null hypothesis H_0 : Variables x, y are *independent*
 - Alternate hypothesis H_1 : They are *not* independent
- Does riding cycles depend on age?
- Contingency table

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

- If H_0 is true, each cell $c_{i,j}$ is a product of $x_{i.}y_{.j}$, i.e., the **marginals**
 - $x_{1.} = 201/395 = 0.50$, $y_{.1} = 100/395 = 0.25$

Test for Independence: Example

- Null hypothesis H_0 : Variables x, y are *independent*
 - Alternate hypothesis H_1 : They are *not* independent
- Does riding cycles depend on age?
- Contingency table

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

- If H_0 is true, each cell $c_{i,j}$ is a product of $x_{i.}y_{.j}$, i.e., the **marginals**
 - $x_{1.} = 201/395 = 0.50$, $y_{.1} = 100/395 = 0.25$
 - Expected $c_{1,1} = 0.50 * 0.25 * 395 = 50.89$;

Test for Independence: Example

- Null hypothesis H_0 : Variables x, y are *independent*
 - Alternate hypothesis H_1 : They are *not* independent
- Does riding cycles depend on age?
- Contingency table

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

- If H_0 is true, each cell $c_{i,j}$ is a product of $x_{i.}y_{.j}$, i.e., the **marginals**
 - $x_{1.} = 201/395 = 0.50$, $y_{.1} = 100/395 = 0.25$
 - Expected $c_{1,1} = 0.50 * 0.25 * 395 = 50.89$; Observed is 60

Test for Independence: Example

- Null hypothesis H_0 : Variables x, y are *independent*
 - Alternate hypothesis H_1 : They are *not* independent
- Does riding cycles depend on age?
- Contingency table

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

- If H_0 is true, each cell $c_{i,j}$ is a product of $x_{i.}y_{.j}$, i.e., the **marginals**
 - $x_{1.} = 201/395 = 0.50$, $y_{.1} = 100/395 = 0.25$
 - Expected $c_{1,1} = 0.50 * 0.25 * 395 = 50.89$; Observed is 60
- Chi-square with degrees of freedom $(4 - 1) * (2 - 1) = 3$

$$\chi^2_{(3)} = (50.89 - 60)^2 / 50.89 + \dots = 8.006$$

- *P-value* (from chi-square distribution table) is 0.046
 - At 5% level of significance (but not 1%), cycling *does* depend on age

Chi-merge Discretization

- Supervised
- Bottom-up

Chi-merge Discretization

- Supervised
- Bottom-up
- Tests every adjacent interval
- If they are *independent* of class distributions, then they can be merged
- Otherwise, the difference in frequencies is statistically significant, and they should not be merged

Chi-merge Discretization

- Supervised
- Bottom-up
- Tests every adjacent interval
- If they are *independent* of class distributions, then they can be merged
- Otherwise, the difference in frequencies is statistically significant, and they should not be merged
- Keep on merging intervals with *lowest* chi-square value(s)

Chi-merge Discretization

- Supervised
- Bottom-up
- Tests every adjacent interval
- If they are *independent* of class distributions, then they can be merged
- Otherwise, the difference in frequencies is statistically significant, and they should not be merged
- Keep on merging intervals with *lowest* chi-square value(s)
- Stopping criterion

Chi-merge Discretization

- Supervised
- Bottom-up
- Tests every adjacent interval
- If they are *independent* of class distributions, then they can be merged
- Otherwise, the difference in frequencies is statistically significant, and they should not be merged
- Keep on merging intervals with *lowest* chi-square value(s)
- Stopping criterion
 - When lowest chi-square value is greater than threshold chosen at a particular level of significance
 - Number of categories greater than a threshold

Example

- Dataset D consists of two classes 1 and 2

Class 1	1, 7, 8, 9, 37, 45, 46, 59
Class 2	3, 11, 23, 39

Example

- Dataset D consists of two classes 1 and 2

Class 1	1, 7, 8, 9, 37, 45, 46, 59
Class 2	3, 11, 23, 39

- Test the first merging, i.e., the first two values: $(1, C_1)$ and $(3, C_2)$
- Contingency table* is

	C_1	C_2	
I_1	1	0	1
I_2	0	1	1
	1	1	2

$$\chi^2 = \frac{(1 - 1/2)^2}{1/2} + \frac{(0 - 1/2)^2}{1/2} + \frac{(1 - 1/2)^2}{1/2} + \frac{(0 - 1/2)^2}{1/2} = 2$$

Example (contd.)

- Test the second merging, i.e., the second and third values: $(3, C_2)$ and $(7, C_1)$

Example (contd.)

- Test the second merging, i.e., the second and third values: $(3, C_2)$ and $(7, C_1)$
- χ^2 is again 2
- Test the third merging, i.e., the third and fourth values: $(7, C_1)$ and $(8, C_1)$
- Contingency table is

	C_1	C_2	
I_1	1	0	1
I_2	1	0	1
	2	0	2

$$\chi^2 = \frac{(1-1)^2}{1} + \frac{(0-0)^2}{0} + \frac{(1-1)^2}{1} + \frac{(0-0)^2}{0} = 0$$

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
------	---	---	---	---	---	----	----	----	----	----	----	----

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	
Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Data	1, 3	7, 8, 9	11, 23	37, 39	45, 46, 59
------	------	---------	--------	--------	------------

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Data	1, 3		7, 8, 9			11, 23		37, 39		45, 46, 59		
χ^2	1.875		5			1.33		1.875				

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Data	1, 3		7, 8, 9			11, 23		37, 39		45, 46, 59		
χ^2	1.875		5			1.33		1.875				

Data	1, 3		7, 8, 9			11, 23, 37, 39					45, 46, 59	
------	------	--	---------	--	--	----------------	--	--	--	--	------------	--

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Data	1, 3		7, 8, 9			11, 23		37, 39		45, 46, 59		
χ^2	1.875		5			1.33		1.875				

Data	1, 3		7, 8, 9			11, 23, 37, 39				45, 46, 59		
χ^2	1.875		3.93							3.93		

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Data	1, 3		7, 8, 9			11, 23		37, 39		45, 46, 59		
χ^2	1.875		5			1.33		1.875				

Data	1, 3		7, 8, 9			11, 23, 37, 39				45, 46, 59		
χ^2	1.875		3.93							3.93		

Data	1, 3, 7, 8, 9					11, 23, 37, 39				45, 46, 59		
------	---------------	--	--	--	--	----------------	--	--	--	------------	--	--

Merging

Data	1	3	7	8	9	11	23	37	39	45	46	59
χ^2	2	2	0	0	2	0	2	2	2	0	0	

Data	1	3	7, 8, 9			11, 23		37	39	45, 46, 59		
χ^2	2	4	5			3		2	4			

- Continue till a threshold
- From chi-square distribution table, chi-square value at significance level 0.1 with degrees of freedom 1 is 2.70 (for significance level 0.05, chi-square value is 3.84)
- So, when none of the chi-square values is less than 2.70, stop

Data	1, 3		7, 8, 9			11, 23		37, 39		45, 46, 59		
χ^2	1.875		5			1.33		1.875				

Data	1, 3		7, 8, 9			11, 23, 37, 39				45, 46, 59		
χ^2	1.875		3.93							3.93		

Data	1, 3, 7, 8, 9					11, 23, 37, 39				45, 46, 59		
χ^2	2.72									3.93		

Intuitive Partitioning

- Relatively uniform, easy to remember and easy to read intervals
- For example, $(500, 600)$ is a better interval than $(512.23, 609.87)$
- Idea is to break up into “natural” partitions

Intuitive Partitioning

- Relatively uniform, easy to remember and easy to read intervals
- For example, (500, 600) is a better interval than (512.23, 609.87)
- Idea is to break up into “natural” partitions
- Uses the 3-4-5 rule

Intuitive Partitioning

- Relatively uniform, easy to remember and easy to read intervals
- For example, (500, 600) is a better interval than (512.23, 609.87)
- Idea is to break up into “natural” partitions
- Uses the 3-4-5 rule
- Hierarchical partitioning

Intuitive Partitioning

- Relatively uniform, easy to remember and easy to read intervals
- For example, (500, 600) is a better interval than (512.23, 609.87)
- Idea is to break up into “natural” partitions
- Uses the 3-4-5 rule
- Hierarchical partitioning
- At each level, examine number of distinct values of the *most significant digit* for each interval
 - If it is 3, 6, 7 or 9 ($3n$), partition into 3 equi-width intervals
 - May be 2-3-2 for 7 partitions
 - If it is 2, 4 or 8 ($2n$), partition into 4 equi-width intervals
 - If it is 1, 5 or 10 ($5n$), partition into 5 equi-width intervals
- Recursively applied at each level

Intuitive Partitioning

- Relatively uniform, easy to remember and easy to read intervals
- For example, (500, 600) is a better interval than (512.23, 609.87)
- Idea is to break up into “natural” partitions
- Uses the 3-4-5 rule
- Hierarchical partitioning
- At each level, examine number of distinct values of the *most significant digit* for each interval
 - If it is 3, 6, 7 or 9 ($3n$), partition into 3 equi-width intervals
 - May be 2-3-2 for 7 partitions
 - If it is 2, 4 or 8 ($2n$), partition into 4 equi-width intervals
 - If it is 1, 5 or 10 ($5n$), partition into 5 equi-width intervals
- Recursively applied at each level
- To avoid outliers, it is applied for data that represents the *majority*, i.e., 5th percentile to 95th percentile

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$
- Since $low' < min$, adjust boundary of first interval to $(-400, 0)$

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$
- Since $low' < min$, adjust boundary of first interval to $(-400, 0)$
- Since $max > high'$, a new interval needs to be formed: $(2000, 5000)$

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$
- Since $low' < min$, adjust boundary of first interval to $(-400, 0)$
- Since $max > high'$, a new interval needs to be formed: $(2000, 5000)$
- These partitions are broken recursively
 - $(-400, 0)$ into

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$
- Since $low' < min$, adjust boundary of first interval to $(-400, 0)$
- Since $max > high'$, a new interval needs to be formed: $(2000, 5000)$
- These partitions are broken recursively
 - $(-400, 0)$ into 4 partitions: $(-400, -300)$, $(-300, -200)$, $(-200, -100)$, $(-100, 0)$
 - $(0, 1000)$ into

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$
- Since $low' < min$, adjust boundary of first interval to $(-400, 0)$
- Since $max > high'$, a new interval needs to be formed: $(2000, 5000)$
- These partitions are broken recursively
 - $(-400, 0)$ into 4 partitions: $(-400, -300)$, $(-300, -200)$, $(-200, -100)$, $(-100, 0)$
 - $(0, 1000)$ into 5 partitions: $(0, 200)$, $(200, 400)$, $(400, 600)$, $(600, 800)$, $(800, 1000)$
 - $(1000, 2000)$ into

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$
- Since $low' < min$, adjust boundary of first interval to $(-400, 0)$
- Since $max > high'$, a new interval needs to be formed: $(2000, 5000)$
- These partitions are broken recursively
 - $(-400, 0)$ into 4 partitions: $(-400, -300)$, $(-300, -200)$, $(-200, -100)$, $(-100, 0)$
 - $(0, 1000)$ into 5 partitions: $(0, 200)$, $(200, 400)$, $(400, 600)$, $(600, 800)$, $(800, 1000)$
 - $(1000, 2000)$ into 5 partitions: $(1000, 1200)$, $(1200, 1400)$, $(1400, 1600)$, $(1600, 1800)$, $(1800, 2000)$
 - $(2000, 5000)$ into

Example

- Suppose data has range $(-351, 4700)$, i.e., $min = -351$, $max = 4700$
- 5th and 95th percentiles are $low = -159$ and $high = 1838$ respectively
- Most significant digit is for 1000
- Rounding yields $low' = -1000$ and $high' = 2000$
- Number of distinct digits is 3
- 3 equi-width partitions are $(-1000, 0)$, $(0, 1000)$ and $(1000, 2000)$
- Since $low' < min$, adjust boundary of first interval to $(-400, 0)$
- Since $max > high'$, a new interval needs to be formed: $(2000, 5000)$
- These partitions are broken recursively
 - $(-400, 0)$ into 4 partitions: $(-400, -300)$, $(-300, -200)$, $(-200, -100)$, $(-100, 0)$
 - $(0, 1000)$ into 5 partitions: $(0, 200)$, $(200, 400)$, $(400, 600)$, $(600, 800)$, $(800, 1000)$
 - $(1000, 2000)$ into 5 partitions: $(1000, 1200)$, $(1200, 1400)$, $(1400, 1600)$, $(1600, 1800)$, $(1800, 2000)$
 - $(2000, 5000)$ into 3 partitions: $(2000, 3000)$, $(3000, 4000)$, $(4000, 5000)$