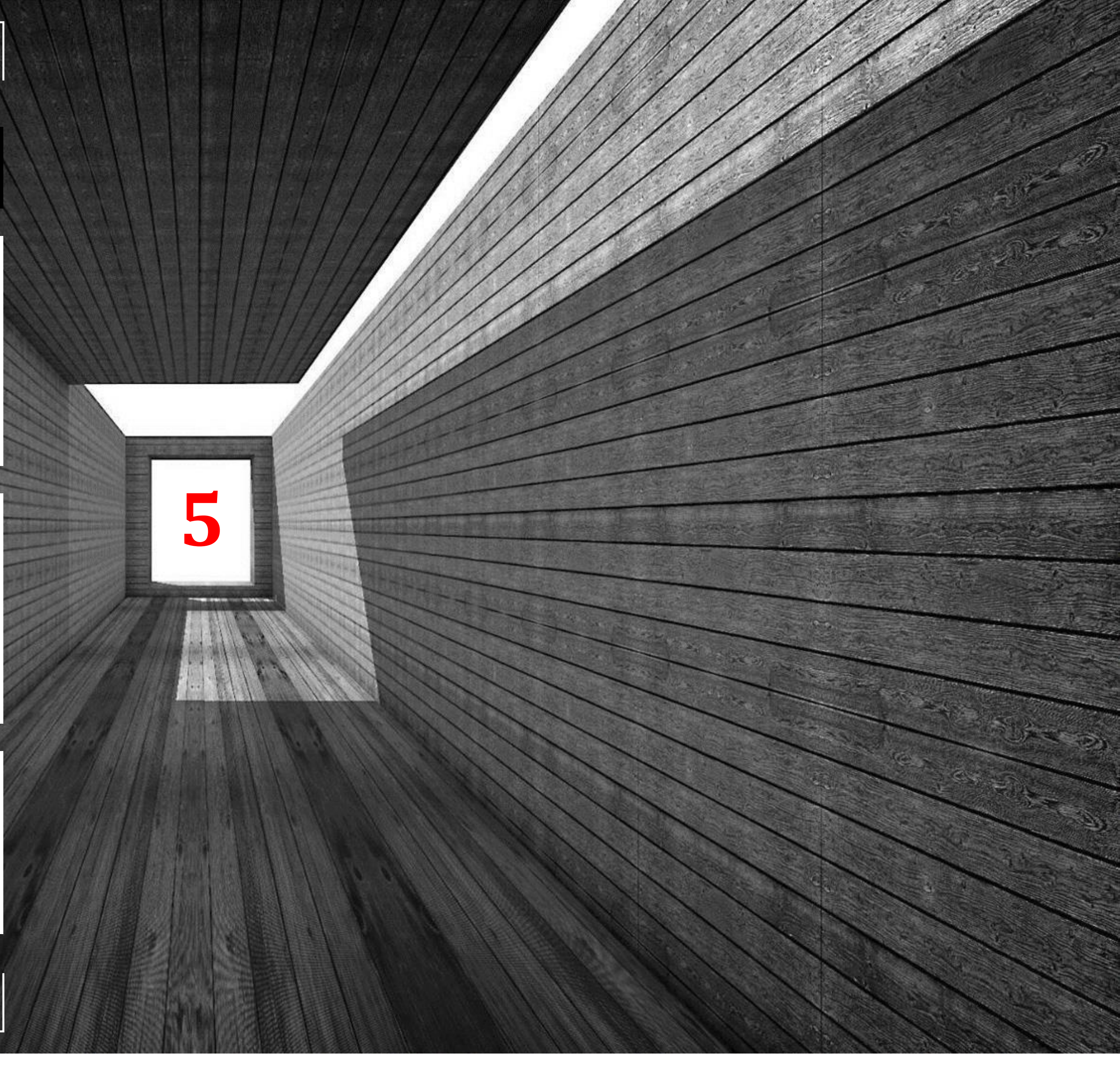
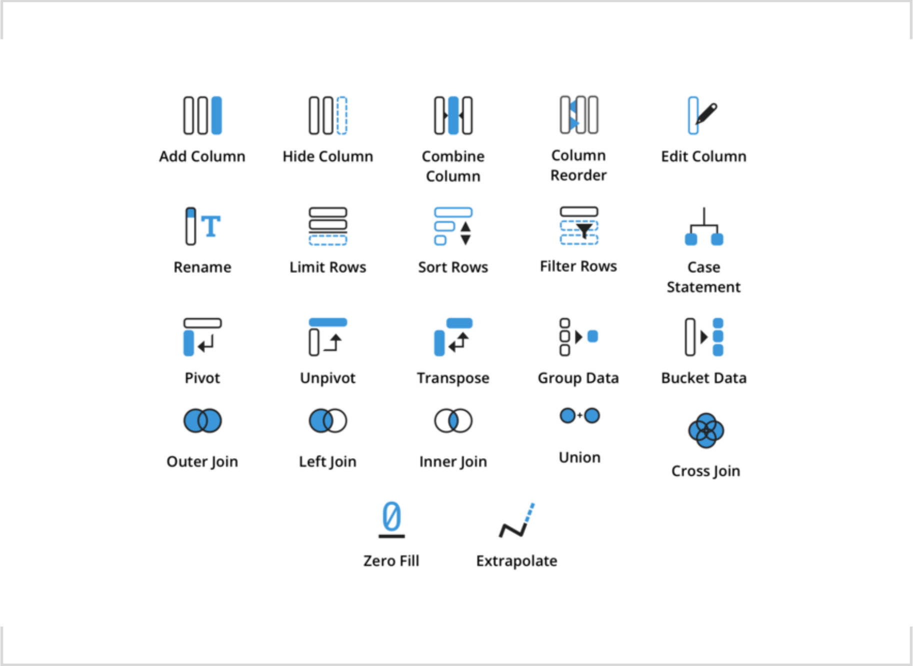
The German University in Cairo

CSEN1095 Data Engineering

Lecture 5 Data Transformation



Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg



Data Transformation

Why Transform Data?

- You need to **convert the data from one format or structure into another** format or structure
- Usually a requirement by data integration
- But also can be used to improve data quality for certain machine learning algorithms
- A batch process – has to be performed on a given attribute at one shot

Transformation and Discretization

- *Smoothing* → binning, regression
- *Aggregation* → grouping and summarization, reduction methods
- *Database Normalization* → establish PKs and FKs
- *Attribute Normalization* → attribute data scaled to fall into smaller range
- *Discretization* → raw values of an attribute (e.g. *age*) replaced by interval labels (e.g. 0–10, 11–20) or conceptual labels (e.g., *youth, adult, senior*) or encodings (0, 1, …)
- *Encoding* → e.g. replace male/female with 1/0
- *Concept hierarchy* → e.g. street generalized to higher-level concepts
- *Attribute Construction / Feature Engineering*

Transformation by **Attribute Normalization**

- To help avoid dependence on the choice of measurement units
- Give all attributes equal weight
- **Methods**
  - *min-max normalization*
  - *z-score normalization*
  - *Distribution fitting*

Transformation by **Attribute Normalization**

- *min-max normalization* → linear transformation
- For value v_i

$$v'_i = \frac{v_i - min_A}{max_A - min_A}\boxed{(new\_max_A - new\_min_A) + new\_min_A}$$

- *Example* → income *min* and *max* are \$12000 and \$98000. Mapping to [0.0, 1.0], a value of \$73600 for income is transformed to

$$\frac{73600 - 12000}{98000 - 12000}(1.0 - 0) + 0 = 0.716$$

Transformation by **Attribute Normalization**

z-score normalization → attribute value normalized based on mean and SD

- For value v_i

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

- *Example* → income *mean* and *SD* are \$54000 and \$16000. z-score for a value of \$73600 for income is

$$\frac{73600 - 54000}{16000} = 1.225$$

Transformation by **Attribute Normalization**

- *Distribution fitting* → transform the *statistical distribution* of the attribute to fit the **normal distribution**
  - Some ML algorithms assume attributes with normal distribution
- **Box-Cox transform** is one such method for distribution transformation

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & if\ \lambda \neq 0 \\ \log y & if\ \lambda = 0 \end{cases}$$

- λ varies from -5 to 5
  - When $\lambda = 0$, the transform is called the **log transform**
- Box-Cox transform works only for positive attribute values

Transformation by **Encoding**

- Encoding involves converting categorical or text attributes/features into numerical representations (not numeric values!)
  - Many ML techniques work better with numerical values
- Two major methods:
  - Label Encoding
  - One-hot Encoding
- Performance of models will be greatly impacted by choice of encoding method!

Transformation by **Encoding**

Label Encoding

- Assign each category in an attribute a numerical value

ID	Country	Population
1	Japan	127185332
2	U.S	326766748
3	India	1354051854
4	China	1415045928
5	U.S	326766748
6	India	1354051854

ID	Country	Population
1	0	127185332
2	1	326766748
3	2	1354051854
4	3	1415045928
5	1	326766748
6	2	1354051854

- Problem: algorithms will assume differences in numeric values mean something (e.g. when country number increases the population increases?)
- If categorical attribute is ordinal, pay attention to numeric assignment to maintain order!

Transformation by **Encoding**

One-hot Encoding

- Create new columns/attributes indicating presence or absence of each possible categorical value in the original data

ID	Country	Population
1	Japan	127185332
2	U.S	326766748
3	India	1354051854
4	China	1415045928
5	U.S	326766748
6	India	1354051854

ID	Country_Japan	Country_U.S	Country_India	Country_China	Population
1	1	0	0	0	127185332
2	0	1	0	0	326766748
3	0	0	1	0	1354051854
4	0	0	0	1	1415045928
5	0	1	0	0	326766748
6	0	0	1	0	1354051854

  - Does not perform well if the categorical variable has a large number of values
  - Usually used for *text analysis*

	female	male	0 - 17	18 - 24	25 - 29	30 - 34	35 - 44	45 - 54	55+	A1	...	SK	SV	TR	TW	US	UY	ZW	acct_age_weeks	user_id	sum
0	0	1	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	329	97f47c9fba714ca68320b8a80e010a1a	29398352
1	1	0	0	0	0	0	0	1	0	0	...	0	0	0	0	1	0	0	178	d615ca85849d458e9a5d755ec4727e8f	31999
2	1	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	68	6c83a5bf63b74f85b106ac7e7e015a1b	29986258
3	1	0	0	0	0	1	0	0	0	0	...	0	0	0	0	1	0	0	8	530fcedb3f244e6f91ecb326740005eb	24333
4	1	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	42	d2ed6a815eda4f61aa346b7936d03ef7	5395128

Transformation by Discretization

- Divide range of a continuous attribute into discrete intervals
  - Interval labels can then be used to replace actual data values
- Some algorithms only accept categorical attributes
- Works also as data reduction mechanism
- Supervised vs. unsupervised
- Split (top-down) vs. merge (bottom-up)
  - Can be **performed recursively** on an attribute

Transformation by **Discretization**

Typical methods (All can be applied recursively)

- Binning, Histograms
  - Both are top-down split, unsupervised
- Clustering
  - Either top-down split or bottom-up merge, unsupervised
- χ^2 analysis
  - bottom-up merge, unsupervised
- Entropy-based
  - Entropy (or information content) is calculated based on a class label
    - Best split so that the bins are as pure as possible that is the majority of the values in a bin correspond to have the same class label
  - top-down split, supervised

Transformation by **Clustering**

- **Partitioning** a set of data objects into subsets or **clusters**
  - Objects in a cluster are similar, yet dissimilar to objects in other clusters
- Clustering can be used also for *outlier detection* and *prediction/analysis*

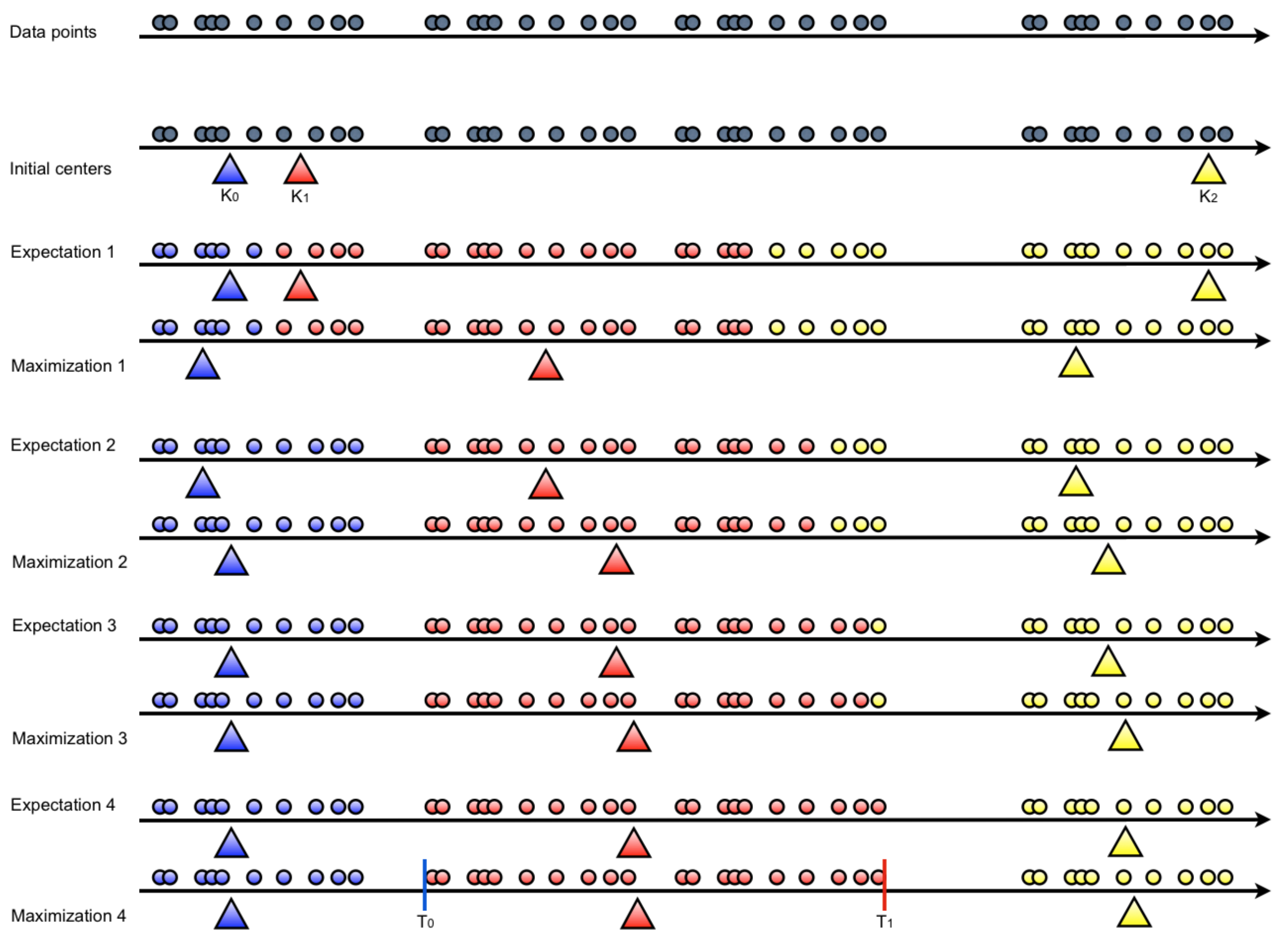
➢ How do we cluster objects together? How do we identify similar objects?

- Similarity/Dissimilarity measures objects *proximity*
- Similarity of i and j → 0 if totally unalike, larger means more alike
- Dissimilarity (distance) of i and j → 0 if totally alike, larger means less alike

Note on Clustering for Discretization

Univariate

- Done per a single numeric attribute



Source: ML Engineering Book

Transformation by **Chi Merge**

- χ^2 is a statistical measure used to test the **null hypothesis** that two discrete attributes are statistically independent
- Premise: Relative class frequencies should be fairly consistent within an interval (otherwise we should split)
- For two adjacent intervals, if χ^2 test concludes that each interval is independent from the class, intervals should be merged
- If χ^2 test concludes that they are not independent, i.e., the difference in relative class frequency is statistically significant, the two intervals should remain separate
- $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(a_{ij} - e_{ij})^2}{e_{ij}}$
  - a_{ij} → # observations in ith interval and jth class
  - e_{ij} → expected # observations $\left(\frac{count\ in\ interval\ i \times count\ in\ class\ j}{total\ count\ in\ the\ two\ intervals}\right)$

Data Discretization using ChiMerge – Example

- Compute the χ^2 value for each pair of adjacent intervals
- Merge the pair of adjacent intervals with the lowest χ^2 value
- Repeat the above steps until χ^2 values of all adjacent pairs exceeds a threshold
- **Threshold**: determined by the significance level and degrees of freedom
  - $df \ = \ number\ of\ classes - 1$

X	Y	Class
1	2	A
3	4	B
5	6	A
7	8	A
9	10	A
11	12	B
13	14	A

Dataset 1		Dataset 2	
X	Class	Y	Class
1	A	2	A
3	B	4	B
5	A	6	A
7	A	8	A
9	A	10	A
11	B	12	B
13	A	14	A

Dataset 1			Dataset 2		
X	Class	Interval	Y	Class	Interval
1	A	$\frac{1+3}{2} = 2 = [0,2]$	2	A	$\frac{2+4}{2} = 3 = [1,3]$
3	B	$[2,4]$	4	B	$[3,5]$
5	A	$[4,6]$	6	A	$[5,7]$
7	A	$[6,8]$	8	A	$[7,9]$
9	A	$[8,10]$	10	A	$[9,11]$
11	B	$[10,12]$	12	B	$[11,13]$
13	A	$[12,14]$	14	A	$[13,15]$

Data Discretization using ChiMerge – Example

- Compute the χ^2 value for each pair of adjacent intervals
- Merge the pair of adjacent intervals with the lowest χ^2 value
  - The higher the χ^2 value the greater the belief that the difference between the two intervals is statistically significant
- Repeat the above steps and until χ^2 values of all adjacent pairs exceeds a threshold

	Class A	Class B	Sums
[0,2]	1 $\left(\frac{1\times1}{2}=0.5\right)$	0 $\left(\frac{1\times1}{2}=0.5\right)$	1
[2,4]	0 $\left(\frac{1\times1}{2}=0.5\right)$	1 $\left(\frac{1\times1}{2}=0.5\right)$	1
Sums	1	1	2

$$\chi^2=\frac{(1-0.5)^2}{0.5}+\frac{(0-0.5)^2}{0.5}+\frac{(0-0.5)^2}{0.5}+\frac{(1-0.5)^2}{0.5}=2$$

Dataset 1

X	Class	Interval
1	A	$\frac{1+3}{2}=2=[0,2]$
3	B	[2,4]
5	A	[4,6]
7	A	[6,8]
9	A	[8,10]
11	B	[10,12]
13	A	[12,14]

Data Discretization using ChiMerge – Example

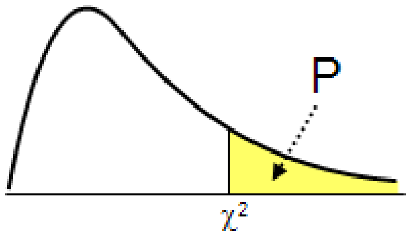
- Threshold: determined by the significance
  - $df = number\ of\ classes - 1$
- $e_{ij} = \dfrac{count\ in\ interval\ i \times count\ in\ class\ j}{total\ count\ in\ the\ two\ intervals}$

	Class A	Class B	Sums
[0,2]	1 $\left(\frac{1\times1}{2} = 0.5\right)$	0 $\left(\frac{1\times1}{2} = 0.5\right)$	1
[2,4]	0 $\left(\frac{1\times1}{2} = 0.5\right)$	1 $\left(\frac{1\times1}{2} = 0.5\right)$	1
Sums	1	1	2

$$\chi^2 = \frac{(1-0.5)^2}{0.5} + \frac{(0-0.5)^2}{0.5} + \frac{(0-0.5)^2}{0.5} + \frac{(1-0.5)^2}{0.5} = \mathbf{2}$$

For one degree of freedom at p-value = 0.1 significance level, the χ^2 value needed to reject the null hypothesis is 2.702

Chi-square distribution table



P

DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	$\boxed{2.706}$	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32.000	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.790
18	6.265	8.231	22.760	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.900	27.204	30.144	32.852	33.687	36.191	38.582	41.610	43.820
20	7.434	9.591	25.038	28.412	31.410	34.170	35.020	37.566	39.997	43.072	45.315
21	8.034	10.283	26.171	29.615	32.671	35.479	36.343	38.932	41.401	44.522	46.797

Data Discretization using ChiMerge – Example

	Class A	Class B	Sums
[0,2]	1 $\left(\frac{1\times1}{2}=0.5\right)$	0 $\left(\frac{1\times1}{2}=0.5\right)$	1
[2,4]	0 $\left(\frac{1\times1}{2}=0.5\right)$	1 $\left(\frac{1\times1}{2}=0.5\right)$	1
Sums	1	1	2

→ χ^2 =2

	Class A	Class B	Sums
[2,4]	0 (0.5)	1 (0.5)	1
[4,6]	1 (0.5)	0 (0.5)	1
Sums	1	1	2

→ χ^2 =2

	Class A	Class B	Sums
[4,6]	1 $\left(\frac{2\times1}{2}=1\right)$	0 $\left(\frac{0\times1}{2}=0\right)$	1
[6,8]	1 $\left(\frac{2\times1}{2}=1\right)$	0 $\left(\frac{0\times1}{2}=0\right)$	1
Sums	2	0	2

$\boxed{\rightarrow \chi^2 = 0}$

[4,10]

	Class A	Class B	Sums
[6,8]	1 (1)	0 (0)	1
[8,10]	1 (1)	0 (0)	1
Sums	2	0	2

$\boxed{\rightarrow \chi^2 = 0}$

	Class A	Class B	Sums
[8,10]	1 (0.5)	0 (0.5)	1
[10,12]	0 (0.5)	1 (0.5)	1
Sums	1	1	2

→ χ^2 =2

	Class A	Class B	Sums
[10,12]	0 (0.5)	1 (0.5)	1
[12,14]	1 (0.5)	0 (0.5)	1
Sums	1	1	2

→ χ^2 =2

Data Discretization using ChiMerge – Example

- Compute the χ^2 value for each pair of adjacent intervals
- Merge the pair of adjacent intervals with the lowest χ^2 value
  - The higher the χ^2 value the greater the belief that the difference between the two intervals is statistically significant
- **Repeat** the above steps and until χ^2 values of all adjacent pairs exceeds 2.7

	Class A	**Class B**	Sums
[0,2]	1 $\left(\frac{1\times1}{2}=0.5\right)$	0 $\left(\frac{1\times1}{2}=0.5\right)$	1
[2,4]	0 $\left(\frac{1\times1}{2}=0.5\right)$	1 $\left(\frac{1\times1}{2}=0.5\right)$	1
Sums	1	1	2

$\chi^2 = 2$

Dataset 1		
X	**Class**	**Interval**
1	A	[0,2]
3	B	[2,4]
5	A	[4,10]
7	B	[10,12]
9	A	[12,14]

Thank You