COMP 6838: Data Mining

Lecture 4:
Data preprocessing: Data Reduction-Discretization

Dr. Edgar Acuna
Department of Mathematics

University of Puerto Rico- Mayaguez

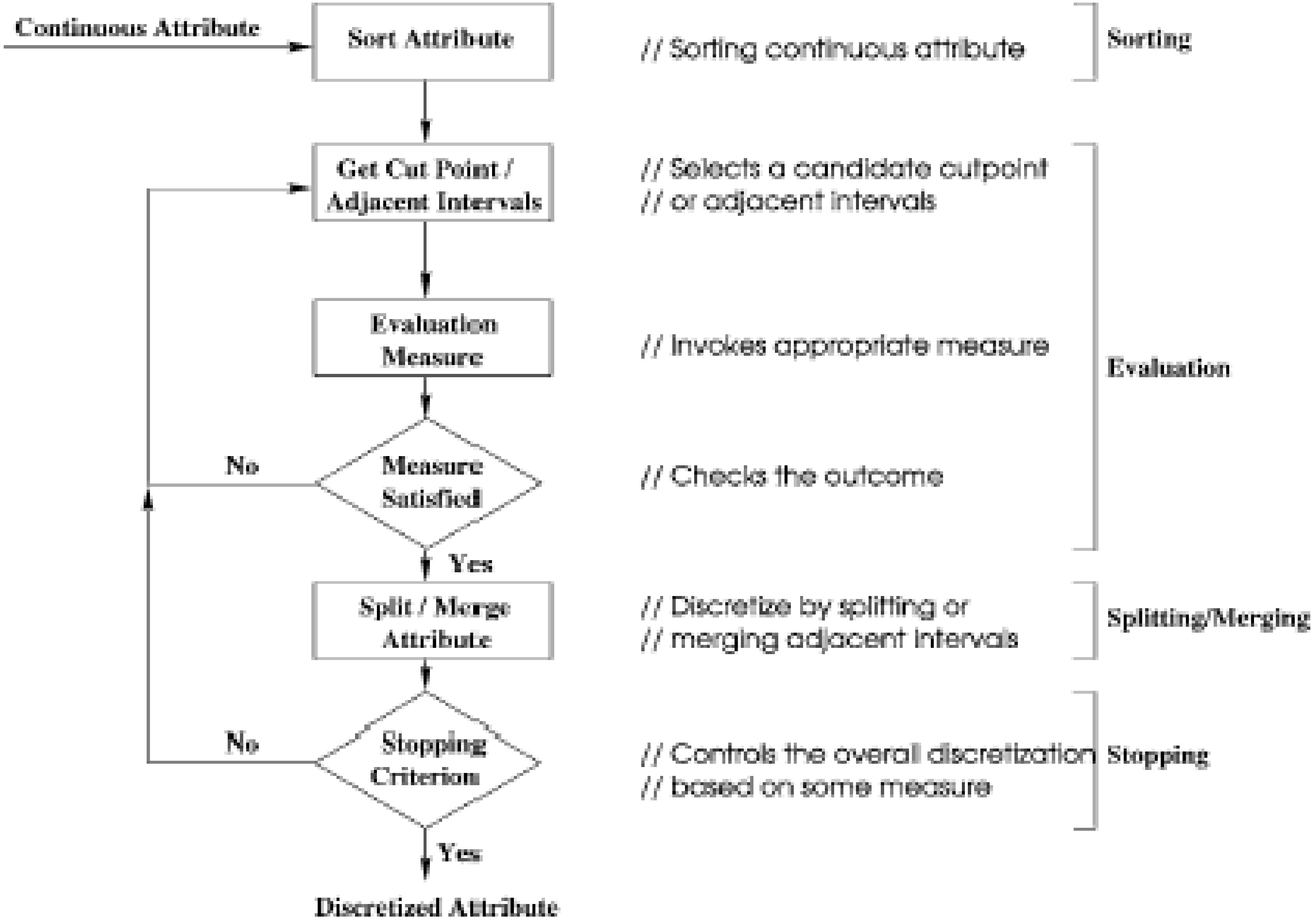
math.uprm.edu/~edgar

Discretization

- Discretization: A process that transforms quantitative data into qualitative data.
  - Some data mining algorithms only accept categorical attributes (LVF, FINCO, Naïve Bayes).
  - The learning process is often less efficient and less effective when the data has only quantitative features.

```
> m
     V1  V2  V3  V4 V5
45 5.1 3.8 1.9 0.4  1
46 4.8 3.0 1.4 0.3  1
47 5.1 3.8 1.6 0.2  1
48 4.6 3.2 1.4 0.2  1
49 5.3 3.7 1.5 0.2  1
50 5.0 3.3 1.4 0.2  1
51 7.0 3.2 4.7 1.4  2
52 6.4 3.2 4.5 1.5  2
53 6.9 3.1 4.9 1.5  2
54 5.5 2.3 4.0 1.3  2
55 6.5 2.8 4.6 1.5  2
```

```
> disc.ew(m,1:4)
   V1 V2 V3 V4 V5
45  1  3  1  1  1
46  1  2  1  1  1
47  1  3  1  1  1
48  1  2  1  1  1
49  1  3  1  1  1
50  1  2  1  1  1
51  2  2  2  2  2
52  2  2  2  2  2
53  2  2  2  2  2
54  1  1  2  2  2
55  2  2  2  2  2
```



The Discretization process. Liu et al. DM and KDD(2002)

Top-down (Splitting) versus Bottom-up(Merging)

- Top-down methods start with an empty list of cut-points (or split-points) and keep on adding new ones to the list by ‘splitting’ intervals as the discretization progresses.
- Bottom-up methods start with the complete list of all the continuous values of the feature as cut-points and remove some of them by ‘merging’ intervals as the discretization progresses.

Static vs. Dynamic Discretization

- Dynamic discretization: some classification algorithms has built in mechanism to discretize continuous attributes (for instance, decision trees: CART, C4.5). The continuous features are discretized during the classification process.
- Static discretization: a pre-preprocessing step in the process of data mining. The continuous features are discretized prior to the classification task.
- There is not a clear advantage of either method (Dougherty, Kohavi, and Sahami, 1995).

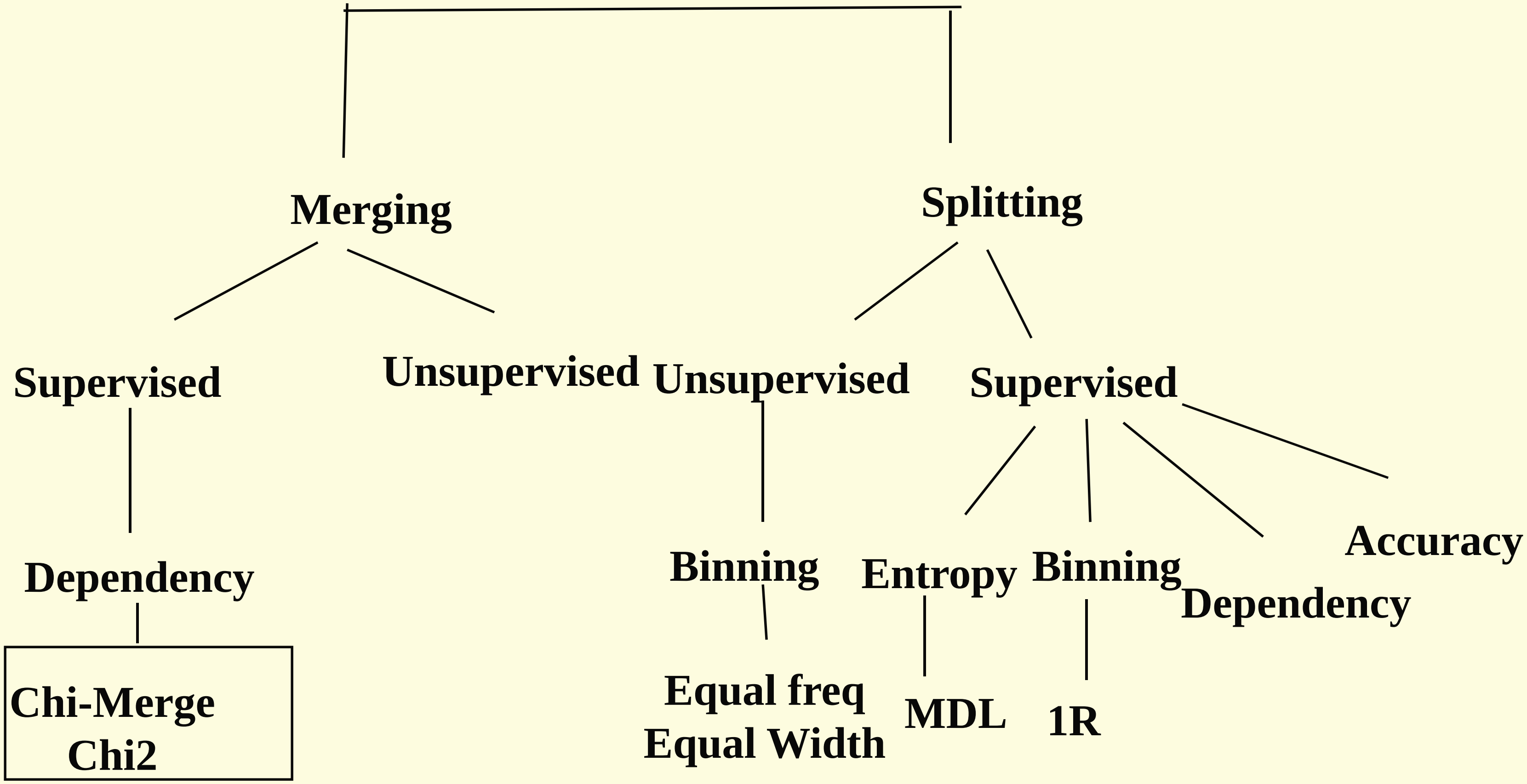
Supervised versus Unsupervised

- Supervised methods are only applicable when mining data that are divided into classes. These methods refer to the class information when selecting discretization cut points.
- Unsupervised methods do not use the class information. An unsupervised technique would not.
- Supervised methods can be further characterized as *error-based*, *entropy-based* or *statistics-based*. Error-based methods apply a learner to the transformed data and select the intervals that minimize error on the training data. In contrast, entropy-based and statistics-based methods assess respectively the class entropy or some other statistic regarding the relationship between the intervals and the class.

Global versus Local

- Global methods use all the space of instances for the discretization process.
- Local methods use only a subset of instances for the discretization process. It is related to dynamic discretization. A single attribute may be discretized into different intervals (Trees).
- Global techniques are more efficient, because only one discretization is used throughout the entire data mining process, but local techniques may result in the discovery of more useful cut points.

A classification of discretization methods



Merging

Splitting

Supervised

Unsupervised

Unsupervised

Supervised

Dependency

Binning

Entropy

Binning

Dependency

Accuracy

Chi-Merge
Chi2

Equal freq
Equal Width

MDL

1R

Evaluating a discretization method

- The total number of intervals generated. A small number of intervals is good up to certain point.
- The number of inconsistencies in the discretized dataset. The inconsistency must decrease.
- The predictive accuracy. The discretization process must not have a major effect in the misclassification error rate.

Equal width intervals (binning)

- Divide the range of each feature into k intervals of equal size
- if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be

$$W = (B\text{-}A) / \mathrm{k}$$

- The interval boundaries are at

$$A+W, A+2W, \ldots , A + (k\text{-}1)W$$

- Ways to determine k:
  - Sturges' Formula: $k=\log_2(n+1)$, n: number of observations.
  - Friedman-Diaconis' Formula: $W=2*IQR*n^{-1/3}$, where IQR=Q3-Q1. Then k=(B-A)/W
  - Scott's Formula: W=3.5*s*n-1/3, where s is the standard deviation. Then k=(B-A)/n.
- Problems
  - (a) Unsupervised
  - (b) Where does k come from?
  - (c) Sensitive to outliers

Ejemplo: Equal width intervals

```
> args(disc.ew)
function (data, varcon)
NULL
> disc.ew(m,1:4)
   V1 V2 V3 V4 V5
45  1  3  1  1  1
46  1  2  1  1  1
47  1  3  1  1  1
48  1  2  1  1  1
49  1  3  1  1  1
50  1  2  1  1  1
51  2  2  2  2  2
52  2  2  2  2  2
53  2  2  2  2  2
54  1  1  2  2  2
55  2  2  2  2  2
```

Equal Frequency Intervals

- Divide the range into *k* intervals
- Each interval will contain approximately same number of samples.
- The discretization process ignores the class information.

Ejemplo: Equal Frequency Intervals

```
> args(disc.ef)
function (data, varcon, k)
NULL
> disc.ef(m,1:4,2)
   V1 V2 V3 V4 V5
45  1  2  1  1  1
46  1  1  1  1  1
47  1  2  1  1  1
48  1  1  1  1  1
49  1  2  1  1  1
50  1  2  1  1  1
51  2  1  2  2  2
52  2  2  2  2  2
53  2  1  2  2  2
54  2  1  2  2  2
55  2  1  2  2  2
```

Method 1R

- Developed by Holte (1993)
- It is a supervised discretization method using binning.
- After sorting the data, the range of continuous values is divided into a number of disjoint intervals and the boundaries of those intervals is adjusted based on the class labels associated with the values of the feature.
- Each interval should contain a given minimum of instances (6 by default) with the exception of the last one.
- The adjustment of the boundary continues until the next values belongs to a class different to the majority class in the adjacent interval.

Example of 1R

Datos ordenados

```
bupat[1:50,1]
 [1] 65 78 79 79 81 81 82 82 82 82 82 82 82 83 83 83 83 83 83 84 84 84 84 84 84
[26] 84 84 84 85 85 85 85 85 85 85 85 85 85 85 85 85 85 85 85 85 85 86 86 86 86 86
```

- Asignando las clases y la clase mayoritaria
- bupat[1:50,2]

```
[1] 2 1 2 2 2 1* 1 2 1 2 2 2 2 2 2 *1 2 2 2 1 2 2 *1 1 2 2 1 2 1 *2 2 2 2 2 2 2 2 2*
        2              2                 2              1               2
[39] 1 1 2 2 2 2 2 2 *1 1 2 1
         2             1
```

Joint the adjacent intervals with the same majority class.

Discretized data

```
1 1  1 1 1 1 1 1 1 1  1 1 1  1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4
```

Ejemplo: Discretizacion 1R

```
> args(disc.1r)
function (data, convar, binsize = 6)
NULL
> disc.1r(m,1:4)
   V1 V2 V3 V4 V5
45  1  2  1  1  1
46  1  1  1  1  1
47  1  2  1  1  1
48  1  1  1  1  1
49  1  2  1  1  1
50  1  2  1  1  1
51  2  1  2  2  2
52  2  1  2  2  2
53  2  1  2  2  2
54  2  1  2  2  2
55  2  1  2  2  2
```

Entropy Based Discretization

- Fayyad and Irani (1993)
- Entropy based methods use the class-information present in the data.
- The entropy (or the information content) is calculated on the basis of the class label. Intuitively, it finds the best split so that the bins are as pure as possible, i.e. the majority of the values in a bin correspond to having the same class label. Formally, it is characterized by finding the split with the maximal information gain.

Entropy-based Discretization (cont)

- Suppose we have the following (attribute-value/class) pairs. Let S denotes the 9 pairs given here. S = (0,Y), (4,Y), (12,Y), (16,N), (16,N), (18,Y), (24,N), (26,N), (28,N).
- Let $p_1 = 4/9$ be the fraction of pairs with class=Y, and $p_2 = 5/9$ be the fraction of pairs with class=N.
- The Entropy (or the information content) for S is defined as:

$$\text{Entropy}(S) = - p_1 * \log_2(p_1) - p_2 * \log_2(p_2) .$$

In this case Entropy(S)=.991076.

- If the entropy small, then the set is relatively pure. The smallest possible value is 0.
- If the entropy is larger, then the set is mixed. The largest possible value is 1, which is obtained when $p_1=p_2=.5$

Entropy Based Discretization(cont)

- Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

where | | denotes cardinality. The boundary T are chosen from the midpoints of the atributes values, i e: {2, 8, 14, 16, 17, 21, 25, 27}

For instance if T: attribute value=14

S_1= (0,P), (4,P), (12,P) and S_2= (16,N), (16,N), (18,P), (24,N), (26,N), (28,N)

E(S,T)=(3/9)*E(S_1)+(6/9)*E(S_2)=3/9*0+(6/9)* 0.6500224

E(S,T)=.4333

Information gain of the split, Gain(S,T) = Entropy(S) - E(S,T).

Gain=.9910-.4333=.5577

Entropy Based Discretization (cont)

Simlarly, for T: v=21 one obtains

Information Gain=.9910-.6121=.2789. Therefore v=14 is a better partition.

- The goal of this algorithm is to find the split with the maximum information gain. Maximal gain is obtained when E(S,T) is minimal.
- The best split(s) are found by examining all possible splits and then selecting the optimal split. The boundary that minimize the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

$$Ent(S) - E(T,S) > \delta$$

Entropy Based Discretization(cont)

where

$$\partial > \frac{\log(N-1)}{N} + \frac{\Delta(T,S)}{N}$$

and,

$$\Delta(S,T) = \log_2(3^c - 2) - [cEnt(S) - c_1 Ent(S_1)$$
$$-c_2 Ent(S_2)]$$

Here c is the number of classes in S, c_1 is the number of classes in S_1 and c_2 is the number of classes in S_2. This is called the Minimum Description Length Principle (MDLP)

Ejemplo: Discretizacion usando Entropia con MDL

```
> args(disc.mentr)
function (data, vars)
NULL
> disc.mentr(bupa,1:7)
The number of partitions for var 1   is : 1
The cut points are: [1] 0
The number of partitions for var 2   is : 1
The cut points are: [1] 0
The number of partitions for var 3   is : 1
The cut points are: [1] 0
The number of partitions for var 4   is : 1
The cut points are: [1] 0
The number of partitions for var 5   is : 2
The cut points are: [1] 20.5
The number of partitions for var 6   is : 1
The cut points are: [1] 0
   V1 V2 V3 V4 V5 V6 V7
1   1  1  1  1  2  1  1
2   1  1  1  1  2  1  2
3   1  1  1  1  2  1  2
4    1  1  1  1  2  1  2
5    ……………………
```

ChiMerge (Kerber92)

- This discretization method uses a merging approach.
- ChiMerge's view:
  - relative class frequencies should be fairly consistent within an interval (otherwise should split)
  - two adjacent intervals should not have similar relative class frequencies (otherwise should merge)

χ^2 Test and Discretization

- χ^2 is a statistical measure used to test the hypothesis that two discrete attributes are statistically independent.
- For two adjacent intervals, if χ^2 test concludes that the class is independent of the intervals, the 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.

The contingency table

	Class 1	Class 2	Sums
Interval I	A_{11}	A_{12}	R_1
Interval II	A_{21}	A_{22}	R_2
Sums	C_1	C_2	N

Computing χ^2

- Value can be computed as follows:

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

k = number of classes

A_{ij} = number of samples in ith interval, jth class

E_{ij} = expected frequency of A_{ij}

$= (R_i \ * C_j) / N$

R_i = number of samples in ith interval

C_j = number of samples in jth class

N = total number of samples on the two intervals

If Eij=0 then set Eij to an small value for instance .1

ChiMerge – El algoritmo

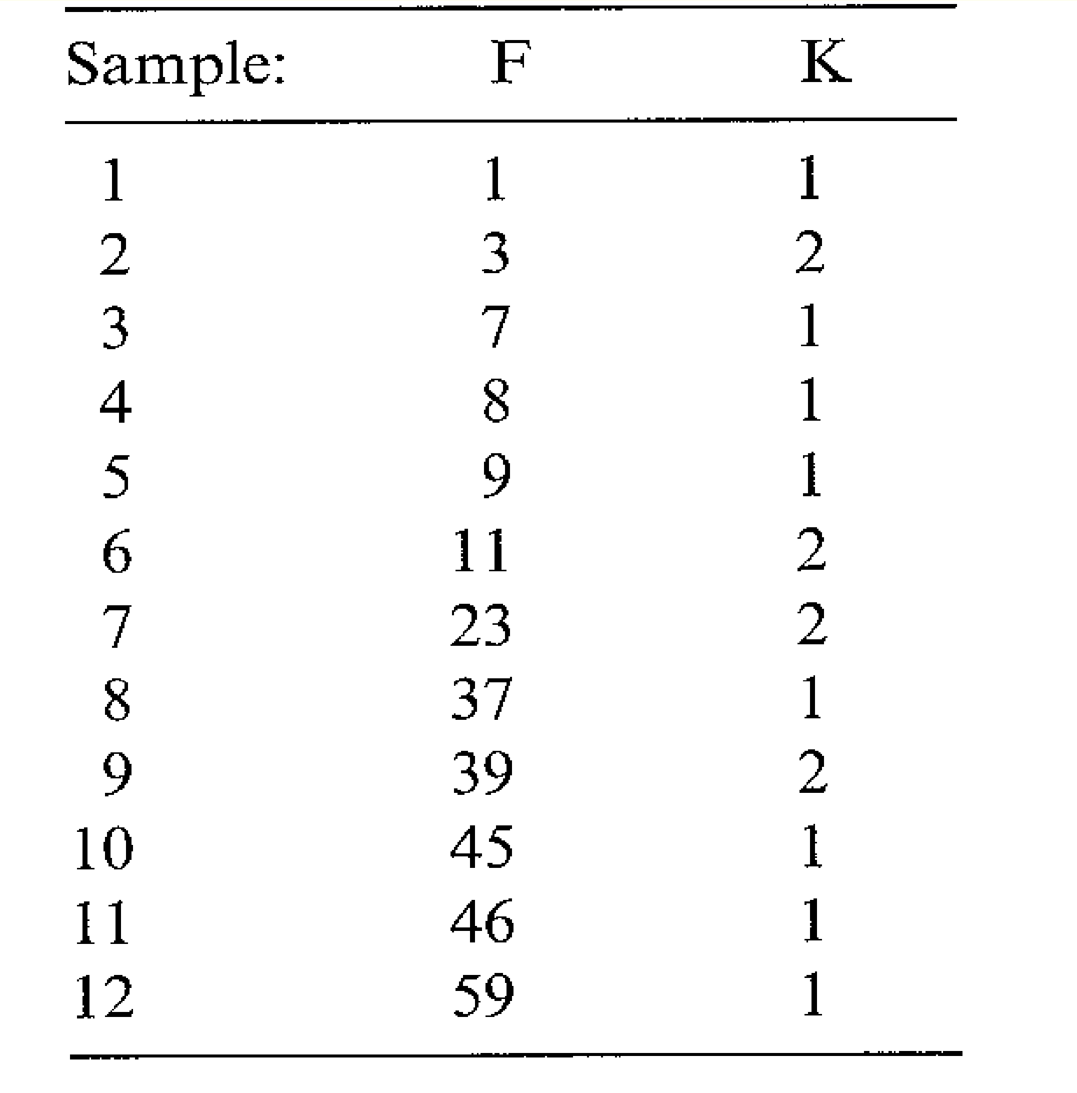
☆ Compute the χ^2 value for each pair of adjacent intervals

🕐 Merge the pair of adjacent intervals with the lowest χ^2 value

🕑 Repeat ☆ and 🕐 until χ^2 values of all adjacent pairs exceeds a threshold

❖ Threshold: determined by the *significance level* and *degrees of freedom* =number of classes -1

Ejemplo



Sample:	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

Ejemplo (cont.)

Splitting initial values are middle between F-points

Minimum $\chi 2$ is on [7.5,8.5] & [8.5,10], with class K=1

	Class 1	Class2	Sums
Interval [7.5,8.5]	1	0	1
Interval [8.5,10]	1	0	1
Sums	2	0	2

Thus E11=1, E12=0~0.1, E21=1, E22=0~0.1, d=degrees of freedom=1

Threshold (for α=10%)=2.706

X^2=0.2. No significant differences →merge

Ejemplo (cont.)

Contigency Tables for the intervals [0,10] and [10,42]

	Class 1	Class2	Sums
Interval [0,10]	4	1	5
Interval [10,42]	1	3	4
Sums	5	4	9

Thus E11=2.78, E12=2.22, E21=2.22 E22=1.78, d=degrees of freedom=1 Threshold (for α=10%)=2.706

X^2=2.72. Significant differences → No merging

FINAL RESULT: 3 intervals [0,10],[10,42],[42,60]

Ejemplo: Discretizacion de Bupa

```
> args(chiMerge)
function (data, varcon, alpha = 0.1)
NULL
> dbupa=chiMerge(bupa,1:6,.05)
> table(dbupa[,1])
  1   2   3
 90 250   5
> table(dbupa[,2])
  1   2   3   4   5   6   7   8   9  10  11  12
  3   4   3  42   9  46 100  30   7   6  16  79
> table(dbupa[,3])
  1   2   3   4   5
 24  21 284   7   9
> table(dbupa[,4])
  1   2   3   4   5   6   7   8
208  20  58   9  35   9   1   5
> table(dbupa[,5])
  1   2   3   4   5   6   7   8   9
  9  69  11  14  37 113  34   3  55
> table(dbupa[,6])
  1   2   3   4
190  67  83   5
```

Effects of Discretization

- Experimental results indicate that after discretization
  - data size can be reduced (Rough sets).
  - classification accuracy can be improved