Data Mining and Machine Learning

Lecture 4:
Feature Enginering: Discretization(Binning)

Dr. Edgar Acuna
Department of Mathematical Sciences

Universidad de Puerto Rico- Mayaguez

website:academic.uprm.edu/eacuna
Github: github.com/eacunafer

Discretization

- Discretization: A process that transforms quantitative data into qualitative data.
  - Some classification algorithm only accept categorical attributes (Rough sets, Naïve Bayes, Bayesian Networks).
  - The learning process is often less efficient and less effective when the data has only quantitative features.

Example:

Original data

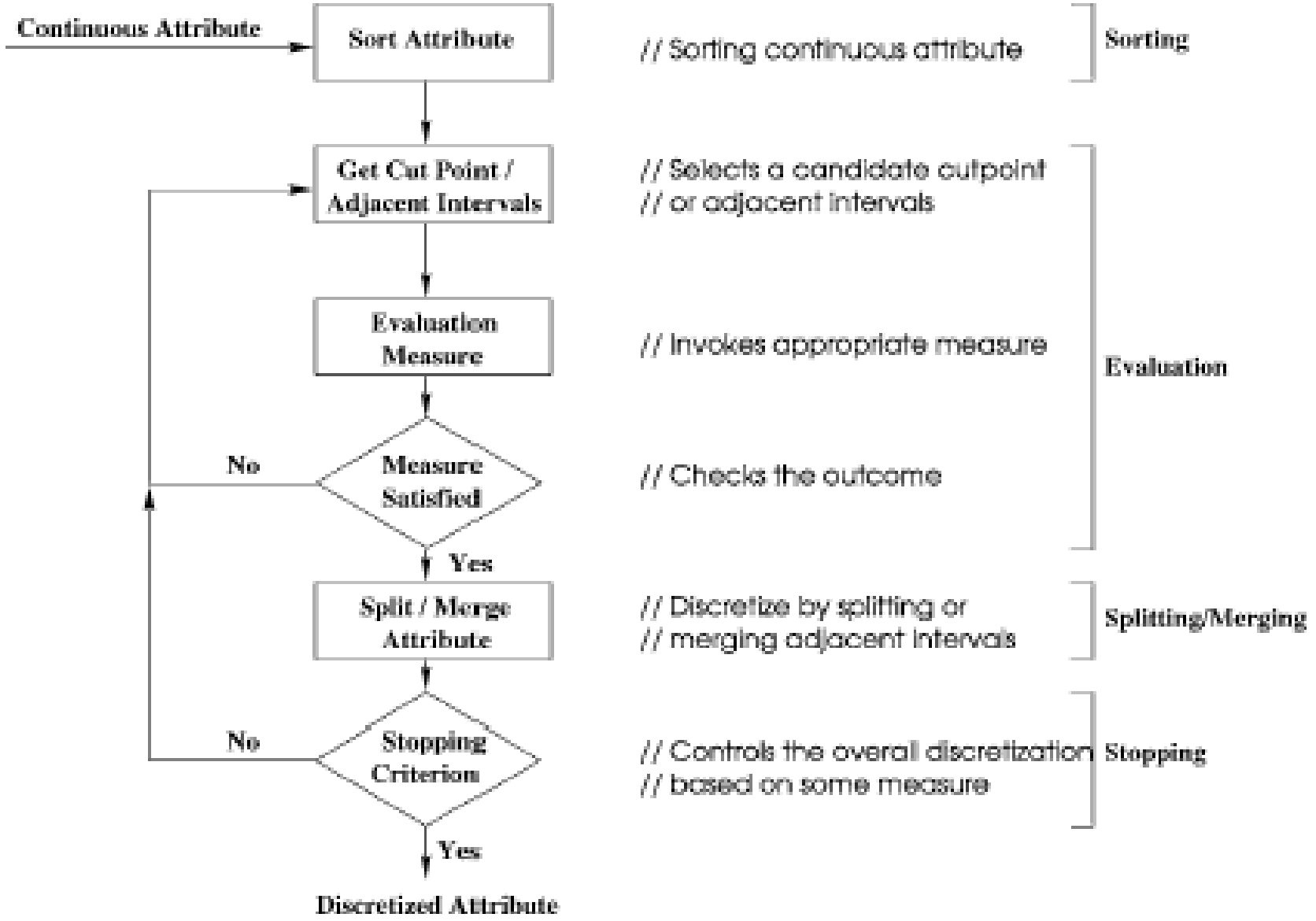
	V1	V2	V3	V4	V5
45	5.1	3.8	1.9	0.4	1
46	4.8	3.0	1.4	0.3	1
47	5.1	3.8	1.6	0.2	1
48	4.6	3.2	1.4	0.2	1
49	5.3	3.7	1.5	0.2	1
50	5.0	3.3	1.4	0.2	1
51	7.0	3.2	4.7	1.4	2
52	6.4	3.2	4.5	1.5	2
53	6.9	3.1	4.9	1.5	2
54	5.5	2.3	4.0	1.3	2
55	6.5	2.8	4.6	1.5	2

Discretized data

	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

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.



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

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).

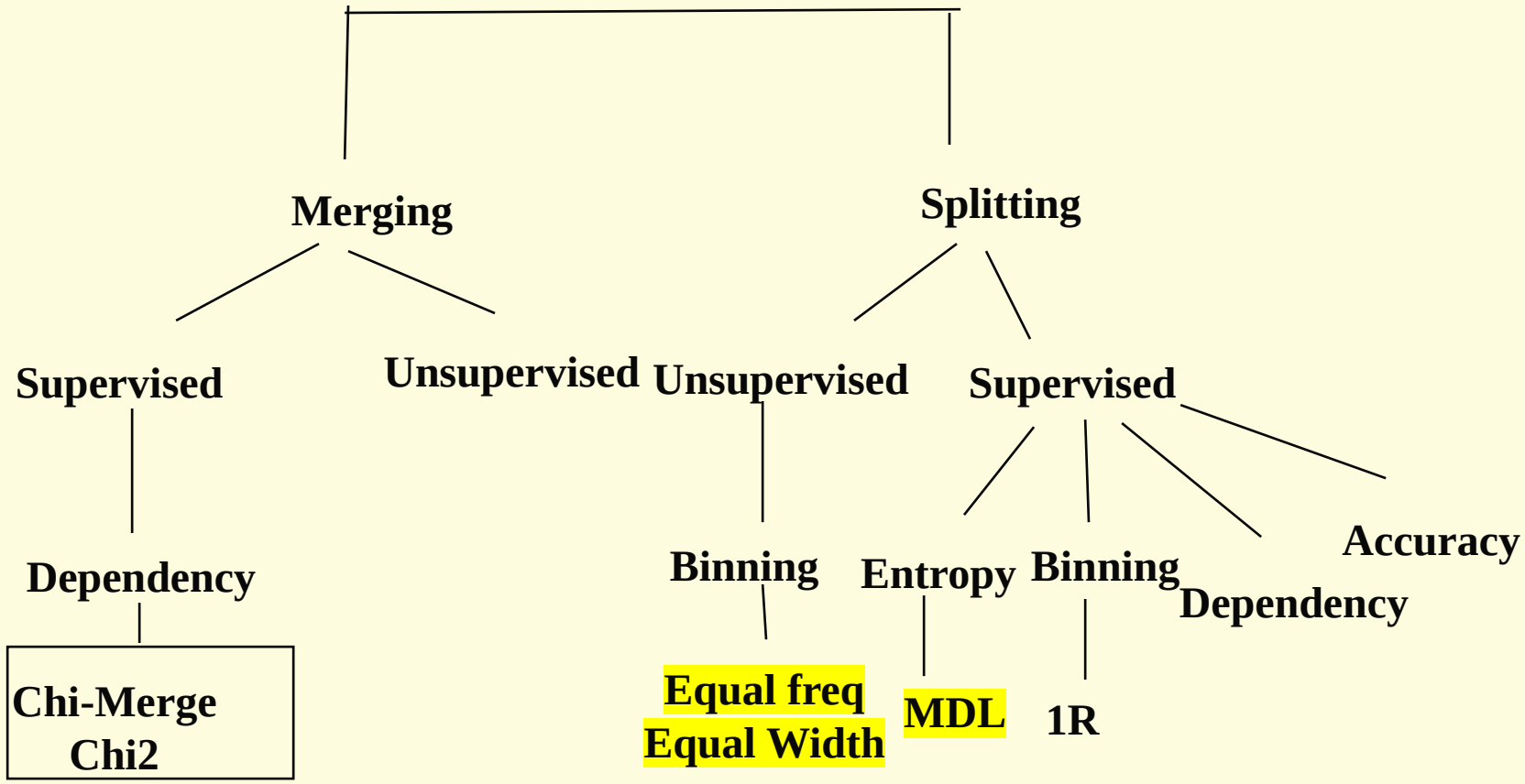
Supervised versus Unsupervised

- 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



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

$$W = (B-A) / k$$

- The interval boundaries are at

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

- 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)/W.
- Problems
  - (a) Unsupervised
  - (b) Where does k come from?
  - (c) Sensitive to outliers

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.

Both, binning using equal width and equal frequency intervals can be done using Pandas and by scikit-learn using KbinsDiscretizer.

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 (subset of Bupa dataset)

Ordering the data of the first column

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

Asigning the corresponding labels to each value and the majority class in each bin

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

Joining the adjacent intervals with the same majority class and assigning a new label.

Discretized data

1 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 4 4 4 4

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:

  Entropy(S) = - p_1*$\log_2(p_1)$ – p_2*$\log_2(p_2)$.

In this case Entropy(S)=.991076.

- If the entropy is 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 attributes values, i e: {2, 8, 14, 16, 17, 21, 25, 27}

For instance if T: attribute value=14

S_1= (0,Y), (4,Y), (12,Y) and S_2= (16,N), (16,N), (18,Y), (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)

Effects of Discretization

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