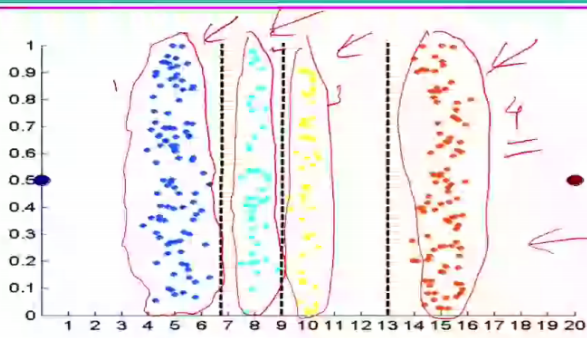


## Discretization Without Using Class Labels



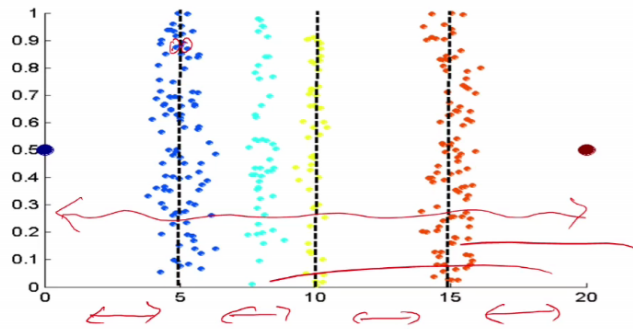
K-means approach to obtain 4 values.

01/22/2018

Introduction to Data Mining, 2nd Edition

60

## Discretization Without Using Class Labels



Equal interval width approach used to obtain 4 values.

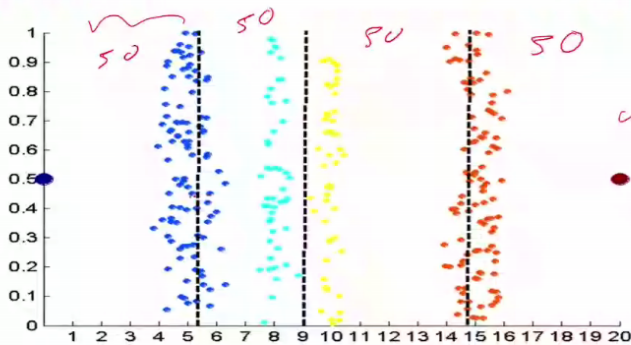
01/22/2018

Introduction to Data Mining, 2nd Edition

58

It's, there's equal frequency

## Discretization Without Using Class Labels



Equal frequency approach used to obtain 4 values.

01/22/2018

Introduction to Data Mining, 2nd Edition

59

(no. of points in a group is same)

$f_1 \rightarrow 10^{20} \rightarrow D_1$   
 $10^{25} \rightarrow D_2$   
 $10^{25} \rightarrow D_3$   
 $10^5 \rightarrow D_4$

You can't do arithmetic  
 for learning small changes  
 machine gets biased data as well

You can apply log

It's for small value

15    18    25    - - - - -    285

$$\left[ \frac{15 - 180}{10} \right]$$

$$\frac{x_i - \bar{x}}{s}$$

← standardization

this just brings values to zero mean  
& 1 std, but doesn't compress

Compress between 0-10 } normalize

$$\left[ \frac{x_i - 180}{300 - 0} \times 10 \right]$$

So we can do several transformations  
(missed the first & last parts)

→ This tells how much you are away from  
mean. etc.

There is no metric which is best. All of these  
need to be used heuristically.