Feature Creation & Discretization

Lesson

AI Academy

Data Preprocessing

- Sampling
- Feature subset selection
- Dimensionality Reduction
- **Feature creation**
- Discretization and binarization
- Attribute Transformation

Feature Creation

Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

Three general methodologies

Feature Extraction
- Domain-specific

Mapping Data to New Space

Feature Construction
- Combining Features

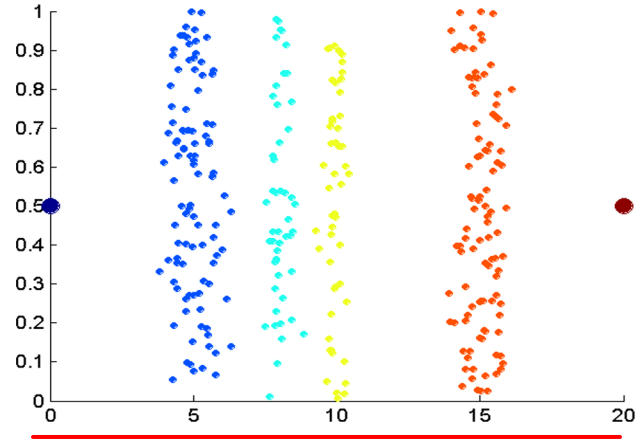
Data Preprocessing

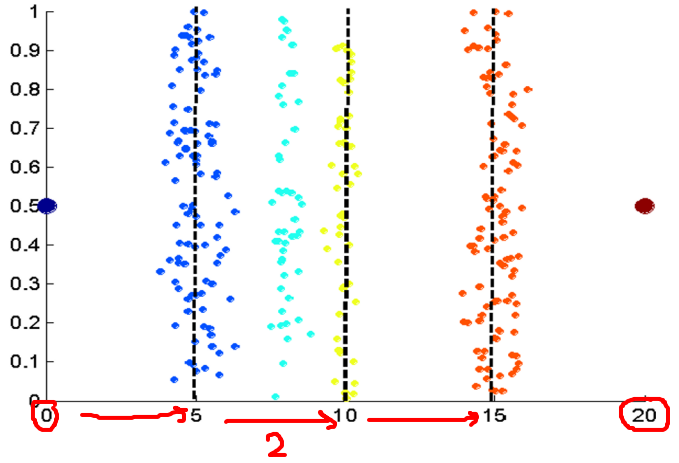
- Sampling
- Feature subset selection
- Dimensionality Reduction
- Feature creation
- **Discretization and binarization**
- Attribute Transformation

Discretization: Unsupervised

Unsupervised: When we don't know the class labels (or don't want to use them)

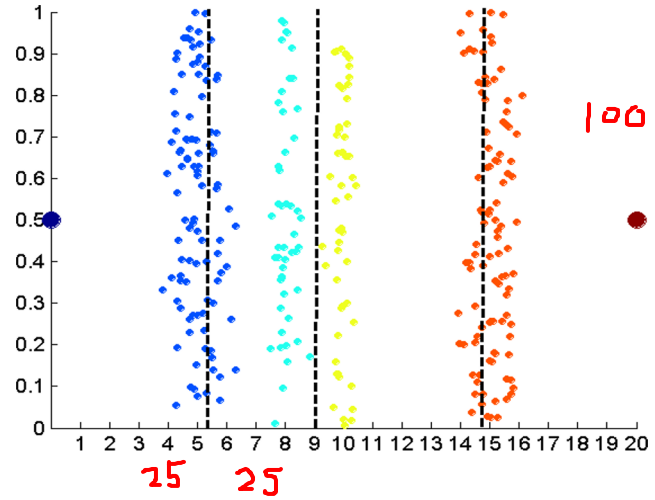
Original Data

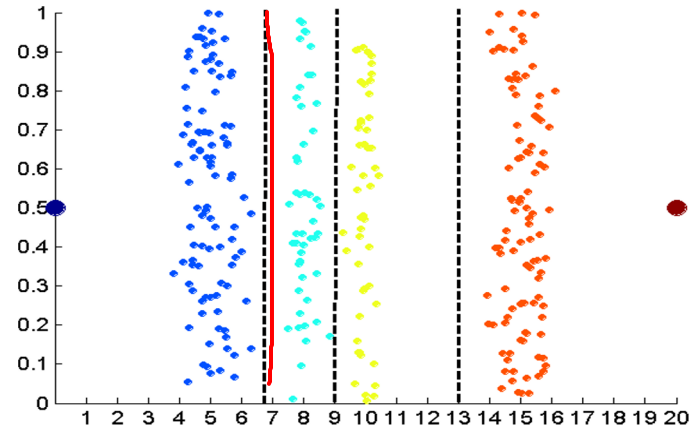




Equal Interval

Equal Frequency





Clustering

AI Academy

NC STATE

Example: Equal Frequency v.s. Interval

Data:

1	2	4	4	6	25	30	80	100

3 Equal-frequency bins:

1	2	4	4	6	25	30	80	100

3 Equal-interval bins: (100 - 1) / 3 = 33

1	2	4	4	6	25	30	80	100

Which depends on the data and your goals.

Binarization: Using 2 Bins

Data:

1	2	4	4	6	25	30	80	100

2 Equal-frequency bins:

1	2	4	4	6	25	30	80	100

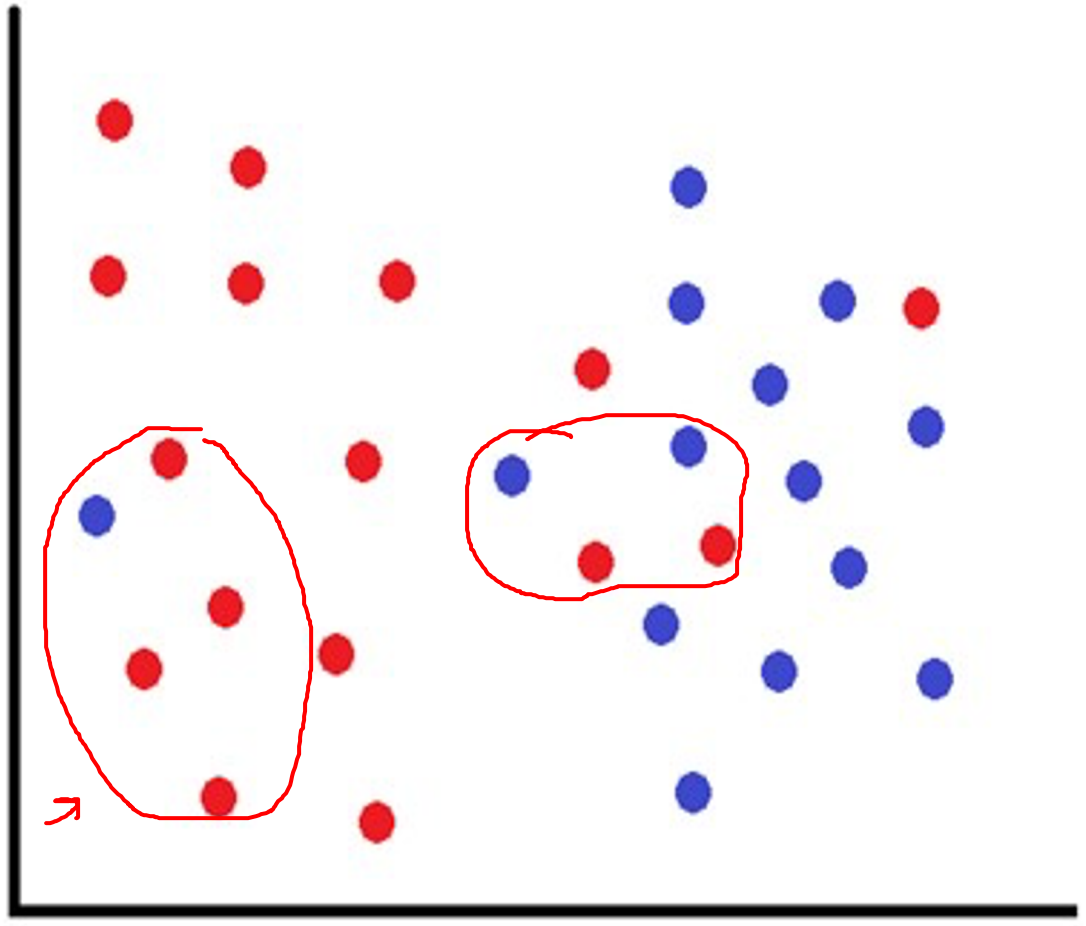
2 Equal-interval bins: (100 - 1) / 2 = 49.5

1	2	4	4	6	25	30	80	100

Which depends on the data and your goals.

Discretization: Supervised

- **Supervised:** Class labels are available
- **Idea:** Place the splits to maximize the "purity"
- Most data objects in a bin should share a class



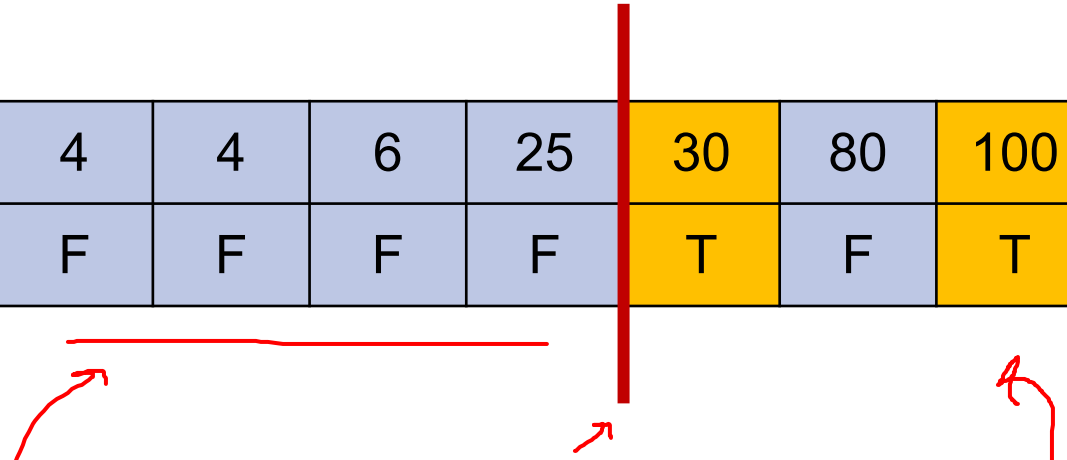
Supervised Discretization: Example

x	1	2	4	4	6	25	30	80	100
y	T	T	F	F	F	F	T	F	T

How can we binarize this attribute to maximize purity?

Supervised Discretization: Example

x	1	2	4	4	6	25	30	80	100
y	T	T	F	F	F	F	T	F	T



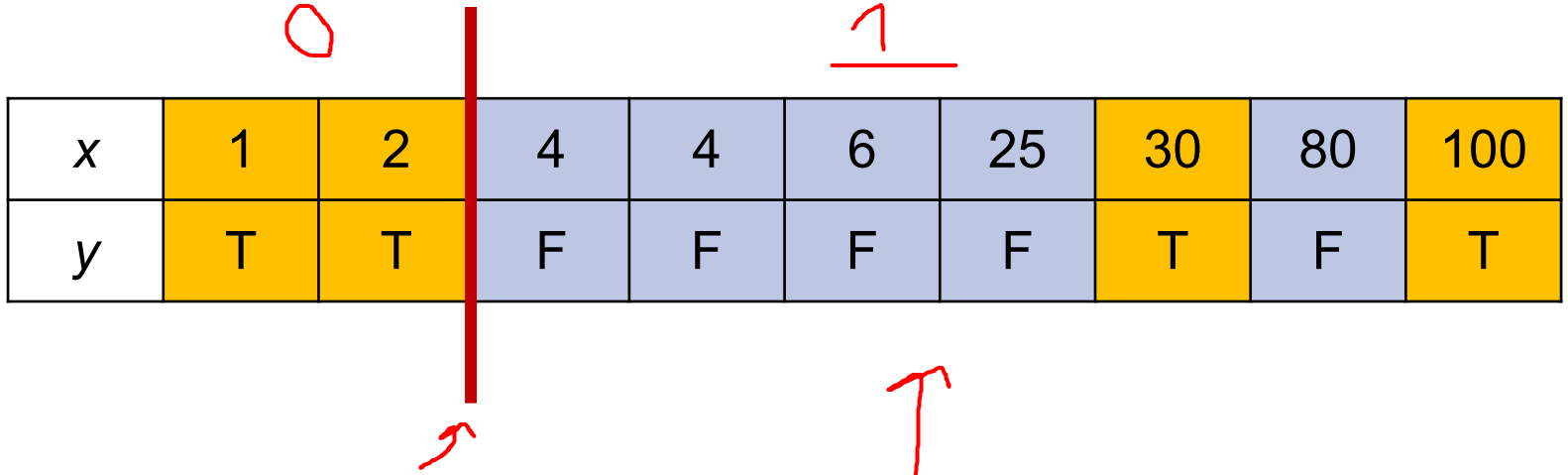
Split 1

Left Bin (< 27): 2T/4F → 66% F

Right Bin (>= 27): 2T/1F → 66% T

Low Purity in both bins

Supervised Discretization: Example



x	1	2	4	4	6	25	30	80	100
y	T	T	F	F	F	F	T	F	T

Split 2

Left Bin (< 3): 2T/0F → 100% T

Right Bin (>= 3): 2T/5F → 71% F

Medium-High Purity in both bins

Discretization: Supervised

Two approaches

- Bisect into two intervals that produce the highest purity split
- Bisect the interval with lowest purity into two

The repeat until stopping criterion is reached

- E.g. Number of intervals, sufficient purity, etc.

Measuring Purity: Wait for Decision Trees!

Learning Objectives: Feature Creation and Discretization

You now should be able to:

Compare different methods for converting continuous attributes into discrete attributes

AI Academy
NC STATE

Feature Creation & Discretization

Exercises

AI Academy

Discussion

Why might we want to do discretize an attribute?