Train In Data

Discretization methods

Discretization task

Discretization is the process of sorting continuous variables into contiguous intervals.

- How do we determine the interval limits?
- How do we determine the number of intervals?

Discretization methods

Unsupervised

- Equal-width
- Equal-frequency
- Arbitrary
- Binarization
- K means

Supervised

- Decision Trees
- Chi-Merge
- CAIM

Discretization methods

Unsupervised

- Equal-width
- Equal-frequency
- Arbitrary
- Binarization
- K means

Given the number of intervals, they find the interval limits.

Supervised

- Decision Trees
- Chi-Merge
- CAIM

Find the optimal number of bins and their limits.

Basic discretization methods

Unsupervised

- **Equal-width**
- **Equal-frequency**
- **Arbitrary**
- Binarization
- K means

Given the number of intervals, they find the interval limits.

Supervised

- Decision Trees
- Chi-Merge
- CAIM

Find the optimal number of bins and their limits.

Alternative discretization methods

Unsupervised

- Equal-width
- Equal-frequency
- Arbitrary
- **Binarization**
- **K means**

Given the number of intervals, they find the interval limits.

Supervised

- **Decision Trees**
- Chi-Merge
- CAIM

Find the optimal number of bins and their limits.

Advanced discretization methods

Unsupervised

- Equal-width
- Equal-frequency
- Arbitrary
- Binarization
- K means

Given the number of intervals, they find the interval limits.

Supervised

- Decision Trees
- **Chi-Merge**
- **CAIM**

Find the optimal number of bins and their limits.

Accompanying Jupyter Notebook

- How to perform discretization:
  - Pandas
  - Scikit-learn
  - Feature-engine

Train In Data

THANK YOU

www.trainindata.com