Train In Data

Discretization

Basic discretization methods

Unsupervised

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

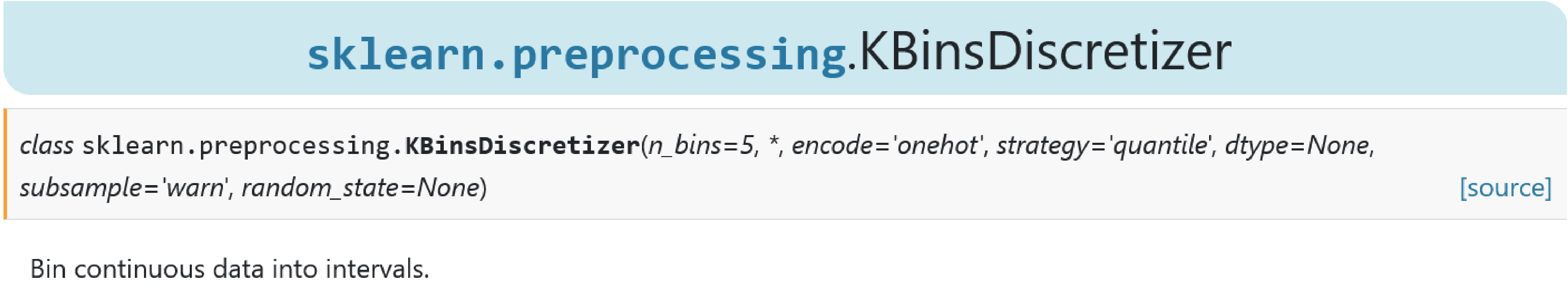
Given the number of intervals, they find the interval limits.

Supervised

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

Find the optimal number of bins and their limits.

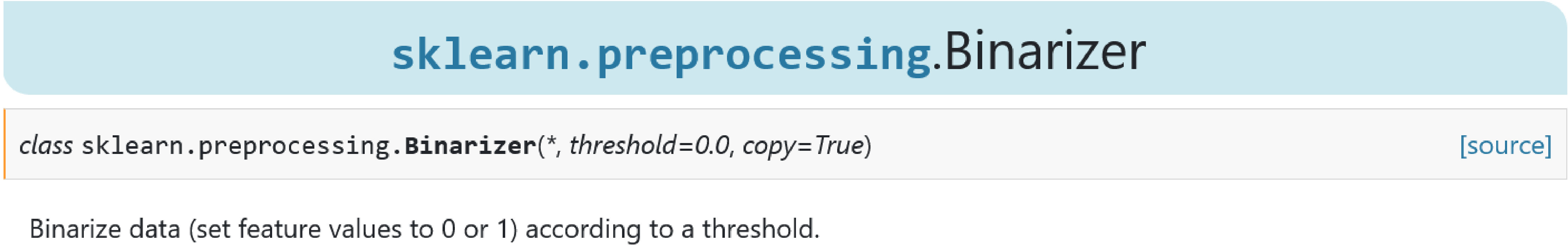
Discretization with sklearn



sklearn.preprocessing.KBinsDiscretizer

class sklearn.preprocessing.**KBinsDiscretizer**(*n_bins=5*, ***, *encode='onehot'*, *strategy='quantile'*, *dtype=None*, *subsample='warn'*, *random_state=None*) [source]

Bin continuous data into intervals.

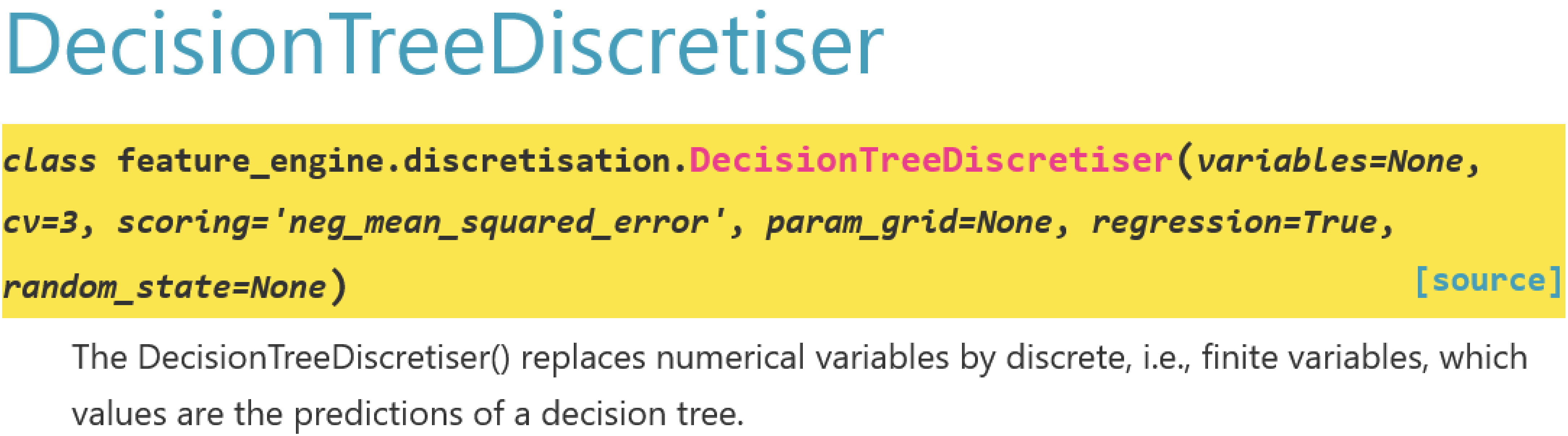


sklearn.preprocessing.Binarizer

class sklearn.preprocessing.**Binarizer**(***, *threshold=0.0*, *copy=True*) [source]

Binarize data (set feature values to 0 or 1) according to a threshold.

Discretization with Feature-engine



DecisionTreeDiscretiser

```
class feature_engine.discretisation.DecisionTreeDiscretiser(variables=None,
cv=3, scoring='neg_mean_squared_error', param_grid=None, regression=True,
random_state=None)
```

[source]

The DecisionTreeDiscretiser() replaces numerical variables by discrete, i.e., finite variables, which values are the predictions of a decision tree.

Accompanying Jupyter Notebook

- How to perform discretization:
  - Scikit-learn
    - All methods
  - Feature-engine
    - Decision tree discretization

Train In Data

THANK YOU

www.trainindata.com