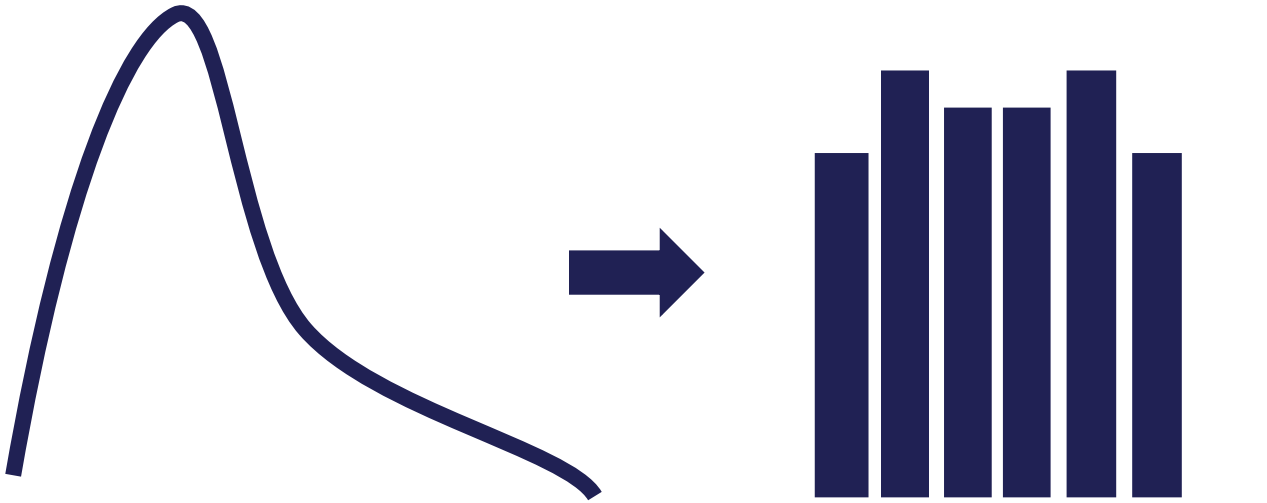
Train In Data

Discretization

Discretization



Continuous variable

Data sorted in intervals

Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of the variable's values.

Discretization is also called **binning**, where bin is an alternative name for interval.

Discretization: why use it?

- Improve performance.
- Reduce training time.
- Mitigate the effect of outliers.
- Create simpler features (for us humans).

Performance & training time

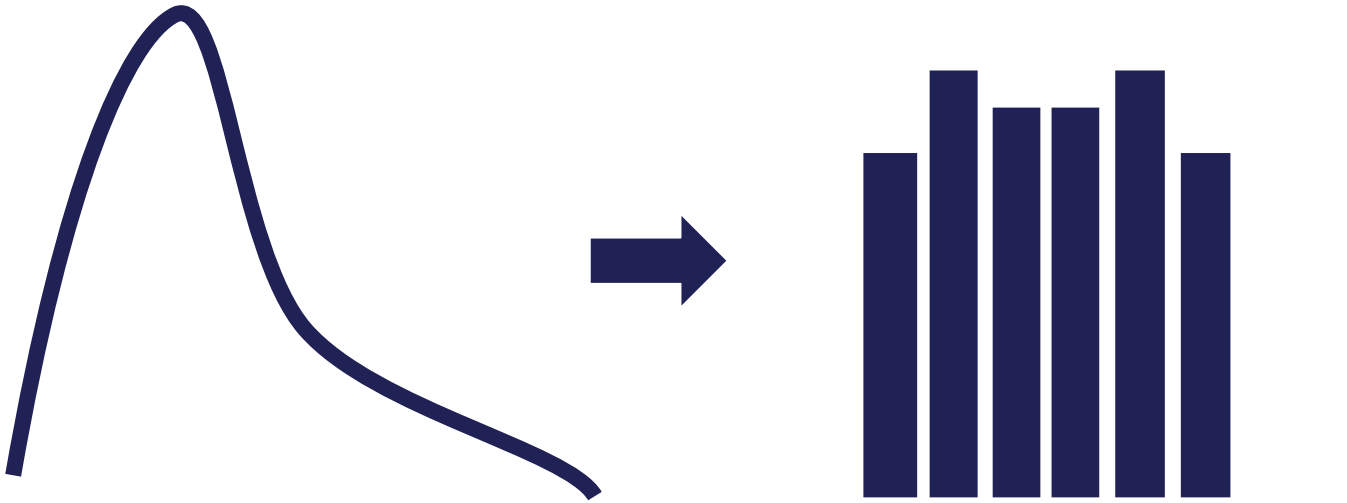
- Many machine learning models, like decision trees and Naïve Bayes, work better with discrete attributes.

Decision trees make decisions based on discrete partitions over the attributes.

During training, a decision tree evaluates all possible feature values to find the best cut-point. Thus, the more values the feature has, the longer the training time.

- Discretization reduces the time it takes to train the models.

Improved value spread

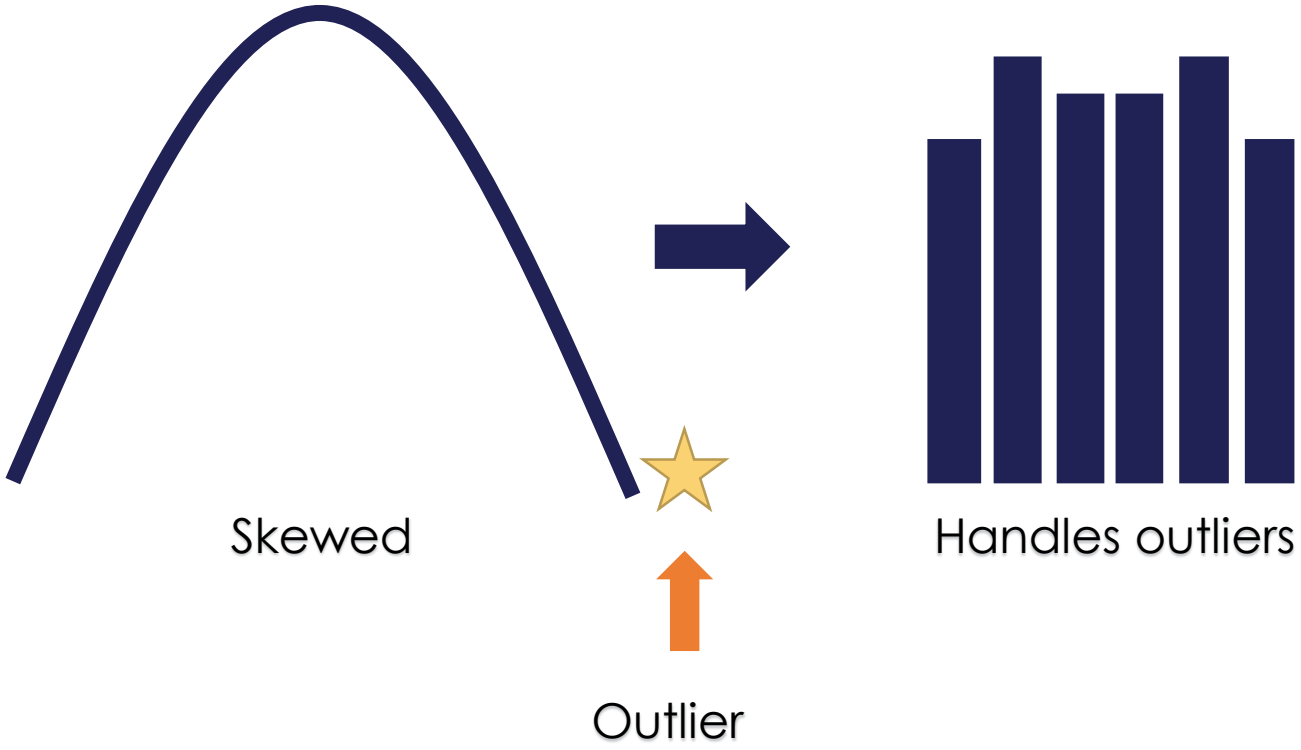


Skewed variable

Improved value spread

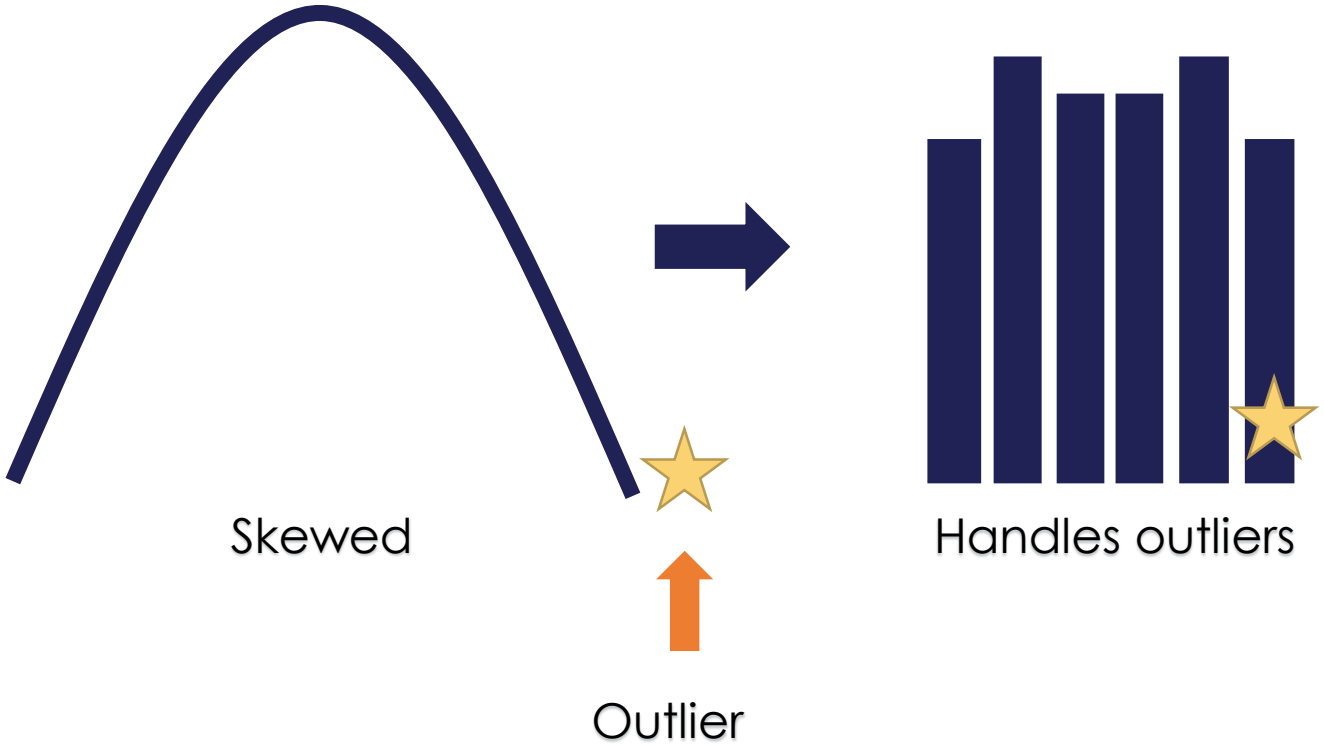
By creating intervals with similar number of observations, we can improve the spread of the values across the value range.

Mitigate effect of outliers



Outliers are placed into the lower or upper intervals, together with the remaining inlier values at the ends of the distribution.

Mitigate effect of outliers



Outliers are placed into the lower or upper intervals, together with the remaining inlier values at the ends of the distribution.

Limitations of discretization

- **Discretization can also lead to a loss of information.**
- For example by combining values that are strongly associated with different classes (target values) into the same bin.
- The aim of a discretization algorithm is to find the minimal number of intervals without a significant loss of information.

Discretization in practice

- In practice, many discretization procedures require the user to input the number of intervals into which the values will be sorted.

- The job of the algorithm is then to find the cut-points for those intervals.

Train In Data

THANK YOU

www.trainindata.com