# installation   
#install.packages("daltoolbox")  
  
# loading DAL  
library(daltoolbox)

About the technique - smoothing\_freq: discretization/smoothing by frequency (quantiles), producing bins with similar counts.

Sample data and general idea.

# Discretization and smoothing  
# Discretization: transform continuous functions, models, variables, and equations into discrete versions.   
  
# Smoothing: create an approximating function to capture important patterns, reducing noise and high-frequency variation.  
  
# Defining bin intervals is essential to enable the approximation/discretization.  
  
# General function to evaluate different smoothing techniques  
  
iris <- datasets::iris  
head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

Apply frequency-based discretization and inspect intervals.

# smoothing using regular frequency  
  
obj <- smoothing\_freq(n = 2)   
obj <- fit(obj, iris$Sepal.Length)  
sl.bi <- transform(obj, iris$Sepal.Length)  
print(table(sl.bi))

## sl.bi  
## 5.19875 6.58   
## 80 70

obj$interval

## [1] 4.3 5.8 7.9

References - Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques. (Discretization)

Evaluate conditional entropy.

entro <- evaluate(obj, as.factor(names(sl.bi)), iris$Species)  
print(entro$entropy)

## [1] 1.097573

Optimize the number of bins and apply again.

# Optimizing the number of binnings  
  
opt\_obj <- smoothing\_freq(n=1:20)  
obj <- fit(opt\_obj, iris$Sepal.Length)  
obj$n

## [1] 8

obj <- fit(obj, iris$Sepal.Length)  
sl.bi <- transform(obj, iris$Sepal.Length)  
print(table(sl.bi))

## sl.bi  
## 4.69090909090909 5.04736842105263 5.38888888888889 5.7047619047619 6.02 6.315 6.65 7.31176470588235   
## 22 19 18 21 15 20 18 17