Waveform Dataset

1 Introduction

This synthetic dataset consists of 21 features with continuous values and a variable representing the three classes (33% for each). Each class is created by combining two of three "base" waves.

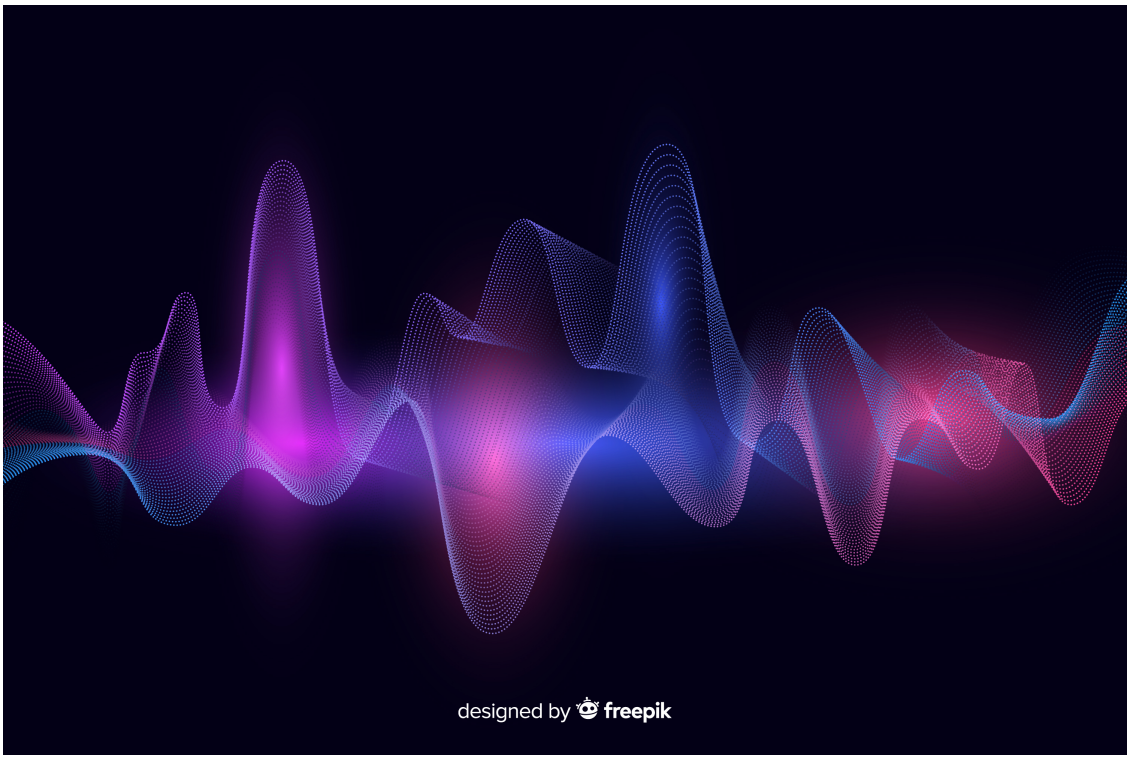

Figure 1: Source: pikisuperstar (link)

To generate the dataset, we need to define `n` - number of patterns to create.

```
# load the dataset from mlbench
waveform <- mlbench.waveform(n = 300) %>% as_tibble()
print(waveform, width = Inf)
```

```
## # A tibble: 300 x 22
##       x.1    x.2    x.3    x.4    x.5     x.6    x.7   x.8   x.9  x.10  x.11
##     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  1.19  -0.507  0.372  1.16  -1.96   0.0970  0.349 2.75  2.55  3.90   4.72
##  2 -0.247  0.671  2.15  -0.757  2.91   2.66    4.81  3.90  3.70  3.46   4.85
##  3  0.722  1.14   0.403 -1.46   0.899  3.52    3.35  4.85  2.32  2.65   6.67
##  4  0.773  0.409  2.31   0.905  2.50   0.305   2.01  0.475 2.98  0.724  2.21
##  5  1.89   0.197  1.47   3.72   2.68   3.33    7.08  3.89  1.67  1.07   2.01
##  6 -1.51  -0.840  4.26   1.14   4.82   3.98    6.31  5.47  4.17  2.39   2.89
##  7 -2.03   2.12  -0.818  1.17  -0.171 -0.893  -1.45  1.77  0.441 3.02   3.85
##  8 -1.81  -1.16  -2.17  -0.421 -0.966 -0.232  -0.163 2.35  2.19  1.30   2.82
##  9  0.398  1.02   1.42   0.893  4.17   5.42    5.58  4.23  5.57  3.45   1.39
## 10  1.02   0.319 -1.01   0.697 -0.952 -1.37    0.273 2.52  2.79  4.13   6.19
##       x.12   x.13   x.14    x.15    x.16    x.17    x.18   x.19   x.20     x.21
##      <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>    <dbl>
##  1  3.81   4.18    5.43   1.52    1.63    2.72    0.417  -1.64  -1.05  -0.434
##  2  2.21   2.53    3.65  -0.106   0.0657 -0.545   0.788   0.469  1.35  -0.210
##  3  3.39   2.17    0.595  2.74    0.398   0.761   0.318  -0.586  0.758  0.314
##  4  2.23   3.63    4.43   4.30    4.25   -0.342   2.85    2.26   0.708  0.550
##  5  1.17   0.768   1.06   0.790   1.58    0.860   1.81    0.771  0.758 -0.516
##  6  1.63   0.299   1.31   0.0818 -1.49   -0.0449 -1.11    0.685  0.108  3.17
##  7  2.40   4.97    5.17   5.44    3.41    3.55    2.36    1.90   1.08  -0.00339
##  8  5.52   3.57    4.71   3.45    4.81    3.93    2.49   -0.488 -0.224  0.213
##  9 -0.0220 0.0543 -0.550  0.882  -0.244   1.38   -2.52    0.340 -0.414  0.0245
## 10  4.88   4.74    4.72   2.92    2.07    1.38    0.0347  1.94   1.33  -0.620
##    classes
##    <fct>
##  1 3
##  2 2
##  3 2
##  4 1
##  5 1
##  6 1
##  7 3
##  8 3
##  9 1
## 10 3
## # ... with 290 more rows
```

2 Dataset Generation Mechanism

The dataset is generated based on the three base waveforms $h_1(t)$ (Figure 2), $h_2(t)$ (Figure 3), $h_3(t)$ (Figure 4). Each class is defined as a random convex combination of two base waveforms with added standard Gaussian noise.

The procedure for generating a data point $\mathbf{x} = (x_1, ..., x_{21})$ (vector of 21 features) is as follows:

- Independently sample a uniform random number u and 21 standard Gaussian distributed random numbers $\epsilon_1, ..., \epsilon_{21}$.
- Choose a class for the data point and obtain the data point based on the class:
  - Class 1: $x_m = u \times h_1(m) + (1-u) \times h_2(m) + \epsilon_m, m = 1, ..., 21$
  - Class 2: $x_m = u \times h_1(m) + (1-u) \times h_3(m) + \epsilon_m, m = 1, ..., 21$
  - Class 3: $x_m = u \times h_2(m) + (1-u) \times h_3(m) + \epsilon_m, m = 1, ..., 21$

For more details regarding the dataset and the source code for generating the dataset, please refer to Leo Breiman (1984) and Dua and Graff (2017).

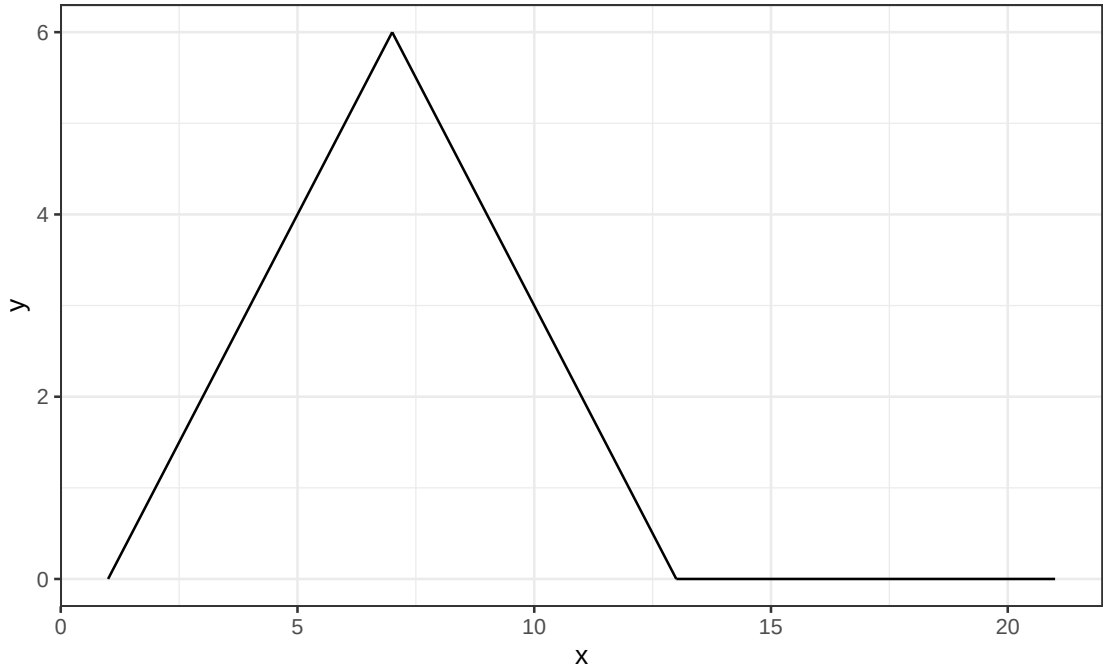

Figure 2: Base waveform $h_1(t)$

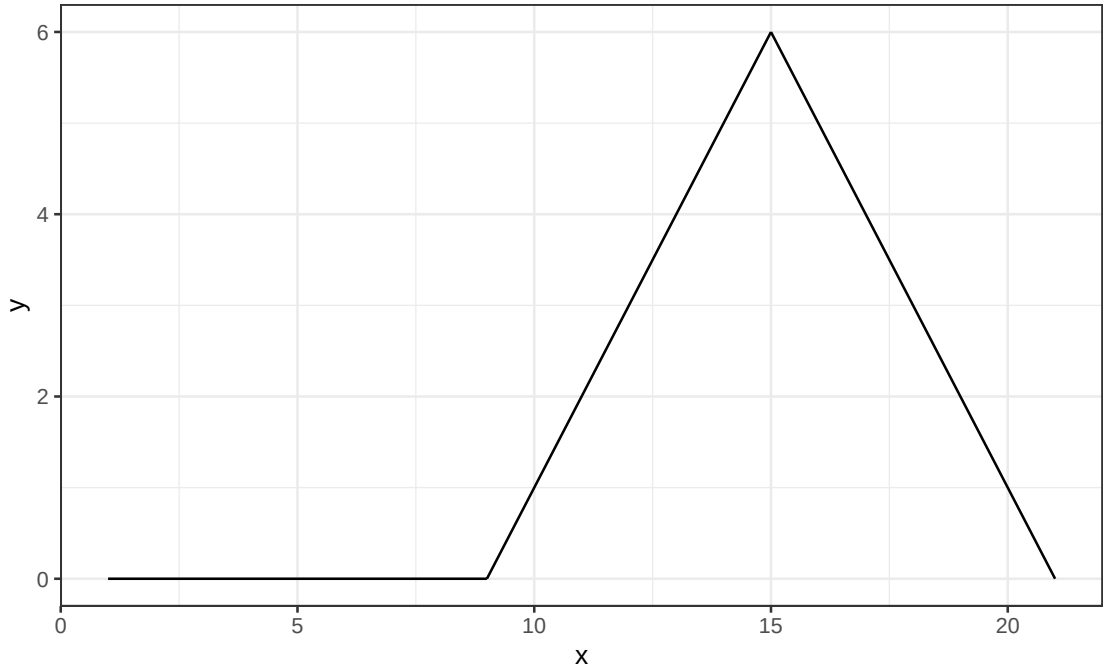

Figure 3: Base waveform $h_2(t)$

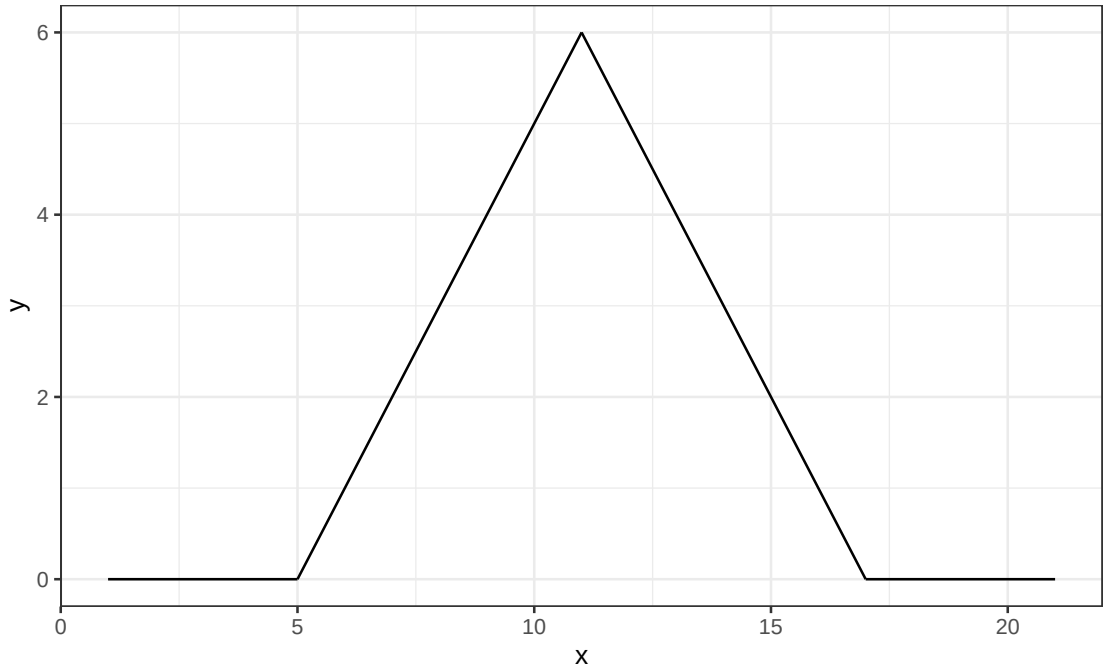

Figure 4: Base waveform $h_3(t)$

References

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

Leo Breiman, Charles J. Stone, Jerome Friedman. 1984. *Classification and Regression Trees.* Chapman; Hall/CRC.