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.

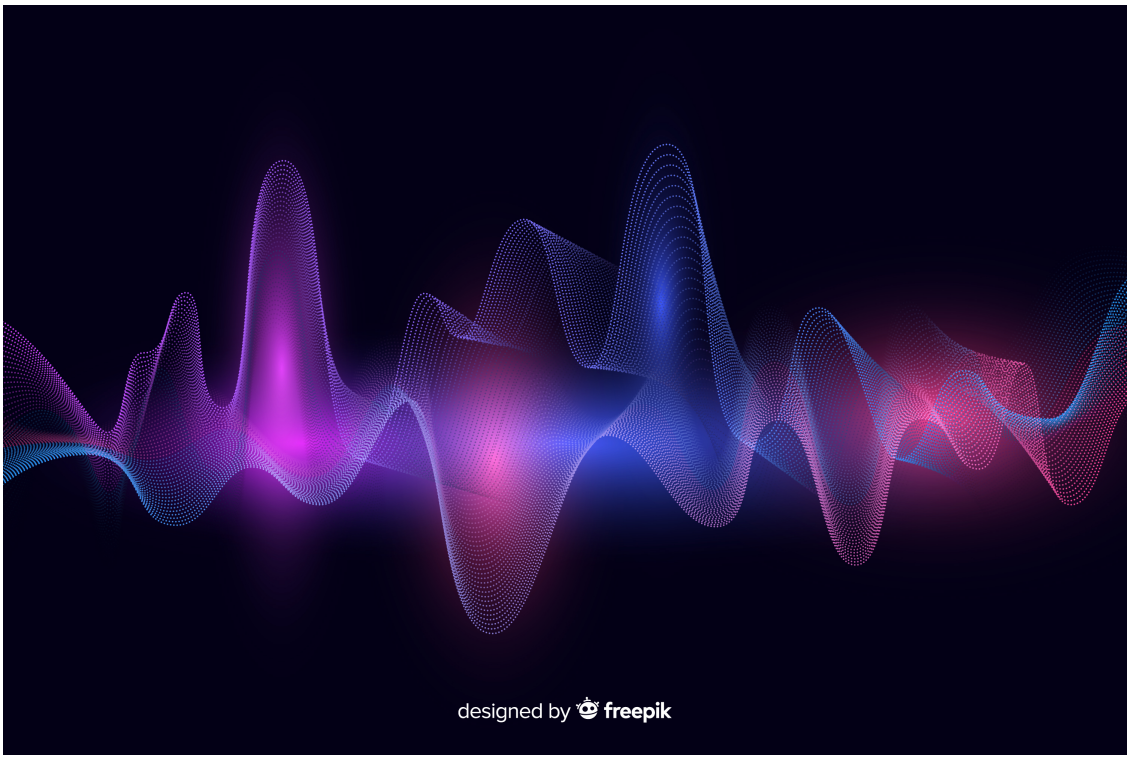

Figure 1: Source: pikisuperstar (link)

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
##     <dbl>   <dbl>  <dbl>    <dbl>  <dbl>  <dbl> <dbl>  <dbl>   <dbl>  <dbl>
##  1 -0.321  0.121   0.290 -0.00554 -0.749  1.18  1.77   3.26   0.930  3.73
##  2 -2.24   2.42    1.98   0.201    2.56   2.62  5.25   6.22   5.94   4.80
##  3  0.724 -0.303   1.79   2.45     1.49   2.87  3.75   1.34   2.86   5.34
##  4 -1.78  -0.0711  1.41   0.885   -1.72  -1.32  0.570  1.23  -2.36   0.0872
##  5 -0.176  1.17    1.86   0.00447  2.30   0.483 0.463 -0.109  0.716  2.85
##  6 -0.703 -0.861   0.562 -0.843    0.452  0.203 1.03  -0.109  0.0254 0.478
##  7  0.278  0.557  -0.486 -0.219    0.697  0.695 1.68   0.828  0.783  0.997
##  8 -0.464 -2.38   -0.501  0.0146  -0.912  0.305 1.83   1.98   2.99   3.82
##  9 -1.30   0.732   0.386  1.78     1.95   2.67  6.27   4.71   3.53   3.29
## 10 -0.499  1.33   -0.739  2.55     0.698  0.836 1.24   0.222 -0.451  0.497
##      x.11  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>   <dbl>
##  1  4.72  4.00  5.81   3.21   3.30  2.29   0.424  0.155  1.23    0.882 -0.0101
##  2  3.64  2.31 -0.0893 0.574  1.64  0.585 -0.499 -0.794 -0.167  -0.195  0.133
##  3  4.22  3.60  3.53   3.11   2.41  0.128 -0.105 -1.38   0.0110 -0.459  0.210
##  4  4.07  3.73  4.04   5.26   6.31  5.69   3.80   2.04   1.55   -1.18  -0.582
##  5  1.90  3.46  1.02   3.82   5.12  3.83   4.37   2.04   2.12    2.35  -0.0655
##  6  3.91  4.55  3.92   3.38   5.09  3.50   4.64   1.29   0.732   0.945  1.07
##  7  2.74  3.09  3.80   5.21   4.87  3.37   1.84   1.73   2.82   -0.466 -0.949
##  8  5.03  5.41  3.88   3.68   2.32  1.16   0.327  0.664  0.316  -1.33   0.662
##  9  4.08  1.67  0.866  0.508 -0.182 0.992  0.874 -0.385 -0.968  -0.676  1.53
## 10  3.14  3.15  3.84   2.18   3.77  4.28   3.32   3.72   1.92    0.300 -0.179
##    classes
##    <fct>
##  1 3
##  2 2
##  3 2
##  4 3
##  5 1
##  6 3
##  7 1
##  8 3
##  9 2
## 10 1
## # ... 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).

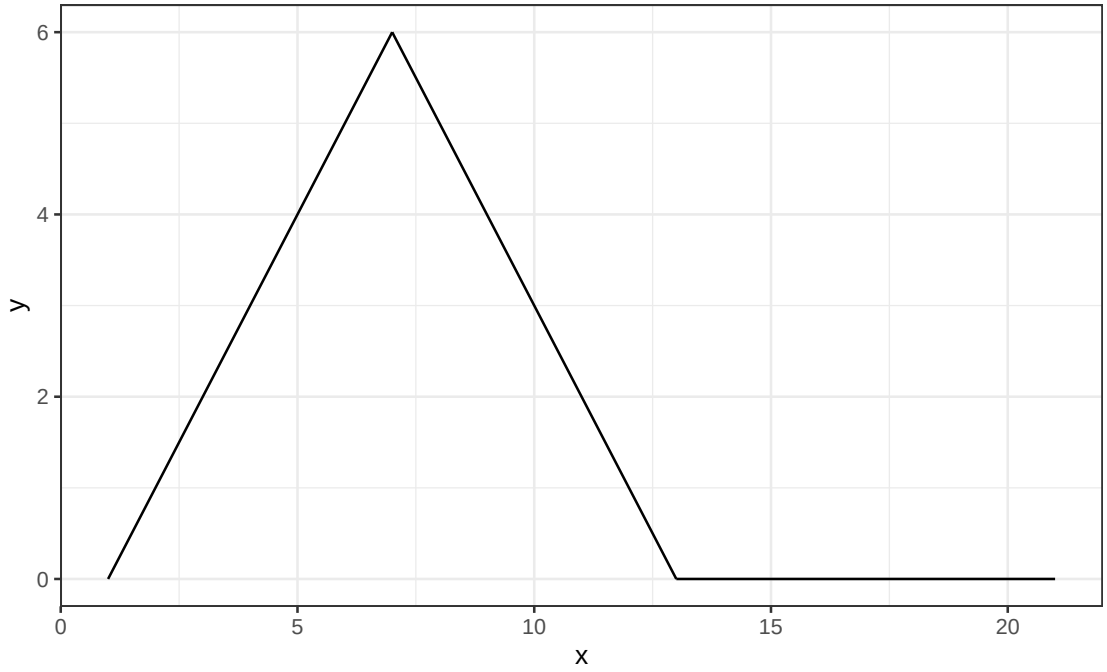

Figure 2: Base waveform $h_1(t)$

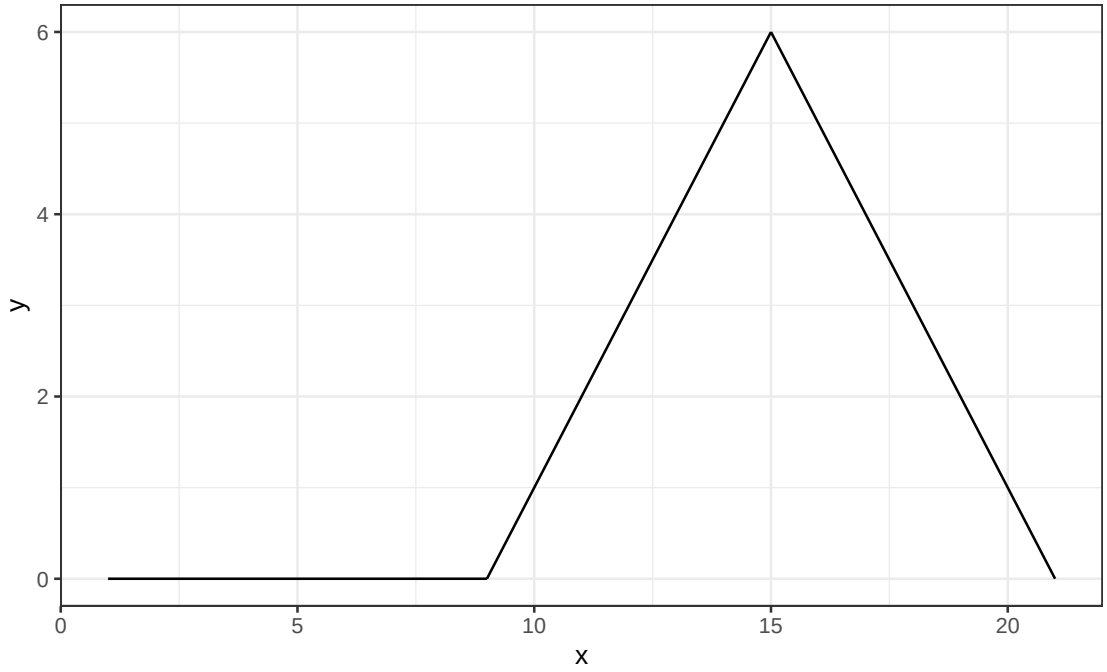

Figure 3: Base waveform $h_2(t)$

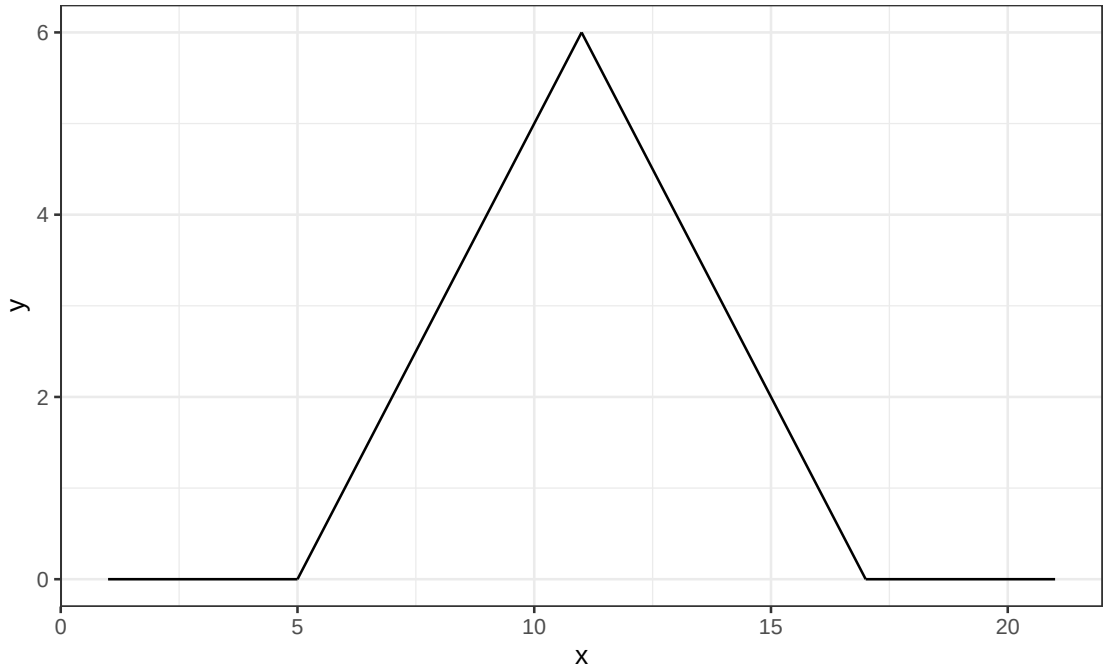

Figure 4: Base waveform $h_3(t)$

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.