# Classification Over Small Datasets with Increasing Cardinality

#### Brook Queree 48423384

## Data Exploration

TODO

## Data Cleaning

There is a singular

## Problem Description

The “83\_Loeschcke\_et\_al\_2000\_Thorax\_&\_wing\_traits\_lab pops” dataset contains three main classification columns:

* Species (D.\_aldrichi or D.\_buzzatii)
* Population (Binjour, Gogango\_Creek, Grandchester, Oxford\_Downs, Wahruna)
* Sex (Male or Female)

With a total of 1731 data points measuring various parts of the fruit fly anatomy. These anatomical

In order to simulate a high-cardinality classification these classification columns were combined into an appended “class” column with a total of 20 possible class vales.

## Decision Tree Classification

## Softmax Classification

## Neural Network Classification

## Lower Cardinality Classification

Single feature classification:

* Species (cardinality of 2)
* Sex (cardinality of 2)
* Population (cardinality of 5)

Double-feature classification:

* Species, sex (cardinality of 4)
* Species, population (cardinality of 10)
* Population, sex (cardinality of 10)

Decision tree classification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class Definition | Cardinality | Test Set Accuracy (min-leaf) | Class Medians | Class Standard Deviations | Class Population |
| Species | 2 | 0.657 | D.\_aldrichi:  D.\_buzzatii: | D.\_aldrichi:  D.\_buzzatii: |  |
| Sex | 2 | 0.811 |  |  |  |
| Species, Sex | 4 | 0.509 |  |  |  |
| Population | 5 | 0.233 |  |  |  |
| Species, Population | 10 | 0.164 |  |  |  |
| Population, Sex | 10 | 0.204 |  |  |  |
| Species, Population, Sex | 20 | 0.133 |  |  |  |

Softmax classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Definition | Cardinality | Test Set Accuracy (Standardised) | Class Medians | Class Standard Deviations |
| Species | 2 | 0.719 |  |  |
| Sex | 2 | 0.830 |  |  |
| Species, Sex | 4 | 0.566 |  |  |
| Population | 5 | 0.260 |  |  |
| Species, Population | 10 | 0.204 |  |  |
| Population, Sex | 10 | 0.225 |  |  |
| Species, Population, Sex | 20 | 0.15 |  |  |

Neural Network classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Definition | Cardinality | Test Set Accuracy after 1200 epochs | Class Medians | Class Standard Deviations |
| Species | 2 | 128: 0.686  256: 0.680  512: 0.678  1024: 0.8660 |  |  |
| Sex | 2 | 128: 0.825  256: 0.830  512: 0.819  1024: 0.834 |  |  |
| Species, Sex | 4 | 128: 0.557  256: 0.553  512: 0.549  1024: 0.530 |  |  |
| Population | 5 | 128: 0.256  256: 0.237  512: 0.191  1024: 0.177 |  |  |
| Species, Population | 10 | 128: 0.164  256: 0.187  512: 0.190  1024: 0.164 |  |  |
| Population, Sex | 10 | 128: 0.229  256: 0.225  512: 0.214  1024: 0.224 |  |  |
| Species, Population, Sex | 20 | 128: 0.150  256: 0.135  512: 0.123  1024: 0.135 |  |  |