University of St Andrews

Machine Learning

CS5014

Classification

Author:
150008022

April 14, 2019

Goal

The goal of this practical is to analyse a dataset in order to produce a classification model that can make predictions based on a set of inputs.

Contents

1 Loading Data

To load the data, the paths to the relevant files are supplied as arguments to the *__main__.py* script. The *pandas* module was used to load the file contents into *DataFrames*.

A test set was isolated from the original data using an 80%-20% split. Stratification was used to ensure that all classes were represented in the training data. Since the dataset was originally grouped by output class, the order of the samples were shuffled. This would avoid the later model being trained on several similiar instances in a row, which can have an affect on some algorithms performance.

2 Cleaning Data

When originally loading the CSV files the parameter to raise an exception on missing or extra columns was included, and so it could be assumed that all rows had the same number of columns. The dtype=**float** argument was also passed when loading the data to ensure that each column contained the expected numerical data. Any rows containing empty or NaN values were dropped from the dataset.

3 Data Visualisation and Analysis

The input CSV was understood to have the structure shown in figure 1. Each value is either the mean, minimum, or maximum reading from 100 radar pulses, and these values are referred to as components.

The mean, min, and max values were plotted for each channel for each sensor. The plots of the means of each channel for the book and plastic case objects are shown in figures 2 and 3 respectively. The difference between the resulting signals from the two objects are very clear.

In the binary dataset, the minimum and maximum components observed all followed a similiar shape as the average, but the book class did contain one severe outlier in two plots. The full plots are included in the submission under `plots/binarybook.png` and `plots/binaryplasticcase`, in which the plot of the minimum components in channel one and the maximum components in channel three both include one row of outliers. Since the average did not

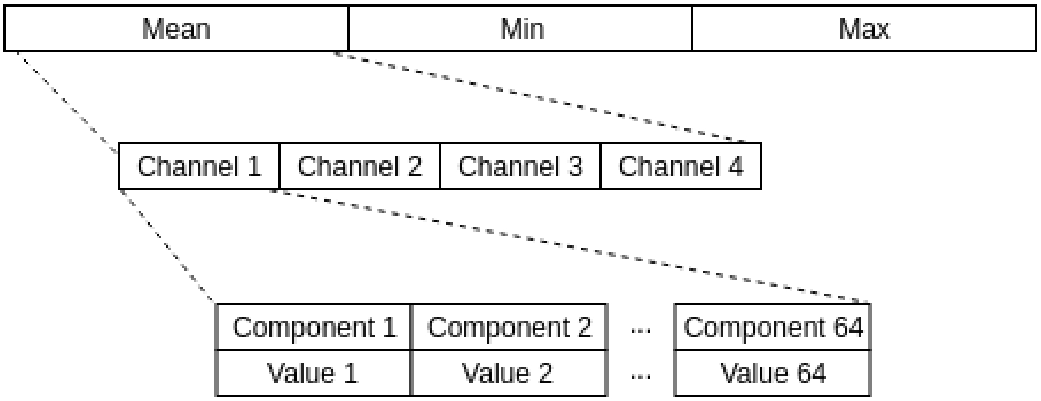


Figure 1: The structure of each row of the CSV file which is repeated for minimum and maximum values.

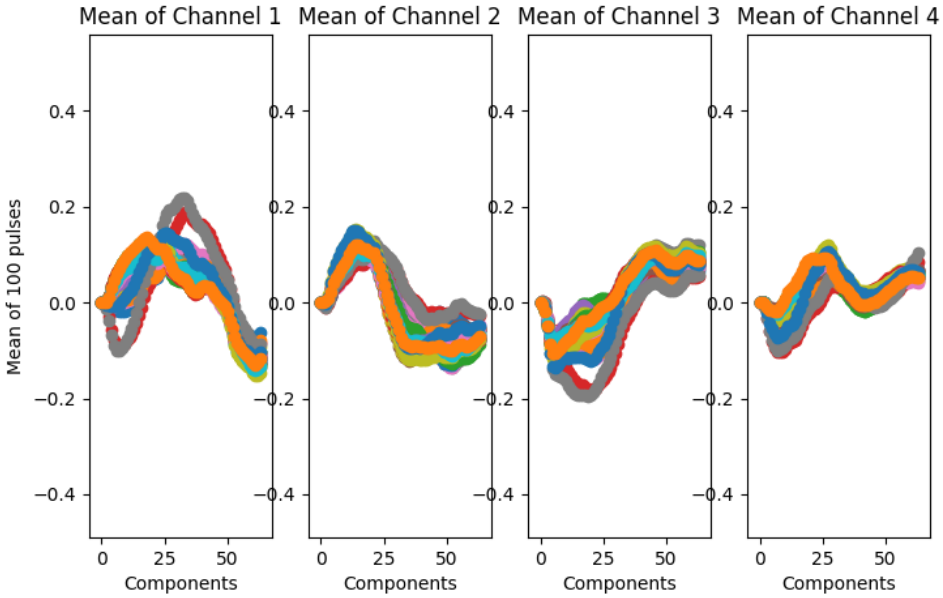


Figure 2: Mean of each channel measured for the book

deviate from other components for that class, it seemed fair to say that these maximum and minimum readings were outliers. Instead of removing them and risking producing a biased model, the row was left in the data set. Since in a real world application of such a sensor would likely involve noise, it made

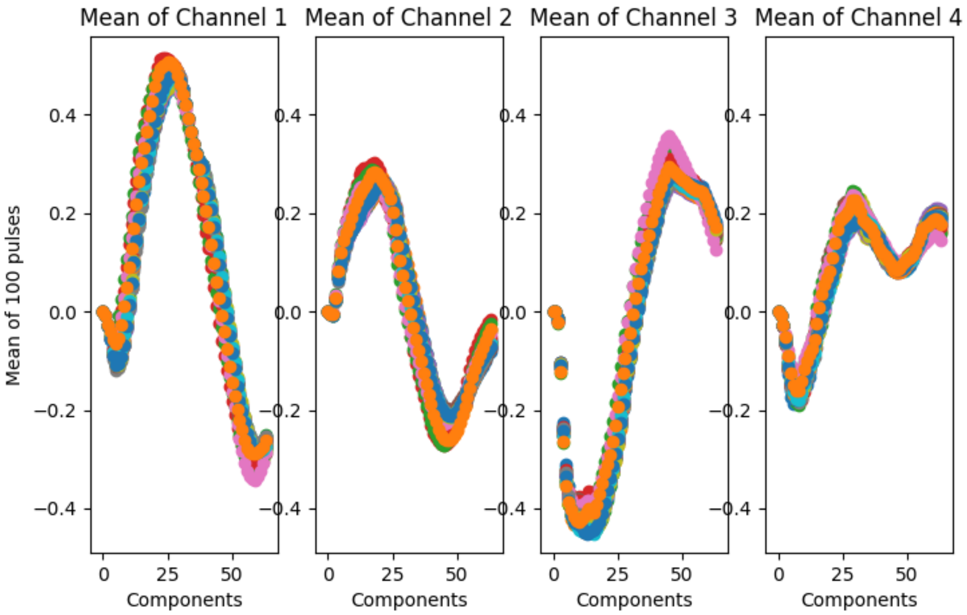


Figure 3: Mean of each channel measured for the plastic case

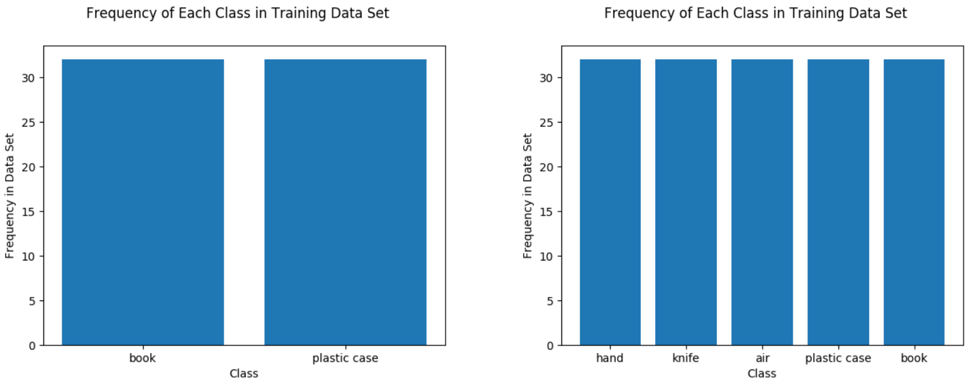
sense to train the model to be able to handle such anomalies. The existence of these outliers was however noted when choosing a cost function however in order to try and minimise their affect.

The same plots were made for the multiclass dataset, and from this it was clear that each material produced very different results, with varying levels of consistency. For example, the data aquired when the radar sensor was applied to a human hand varied wildly, whilst the readings for the plastic cover were a very clear sinusoidal shape. These plots can be found in `plots/multiclass*`.

The training data set was shown not be skewed by plotting the distribution of each class (figure4). The equal distribution of each class meant that cross validation of a classifier that always guesses the same class will have a ratio of correct predictions inversely proportional to number of classes in the training data set.

4 Feature Selection

From the plots of samples ov



(a) Binary

(b) Multiclass

Figure 4: Frequencies of each class in the binary and multiclass training data sets

5 Model Selection and Training

Since each dataset contains an equal number of samples for each classes, the probability of a random classifier guessing correctly is equal to .

5.1 Model 1

5.2 Model 2

6 Evaluation and Comparison

7 Discussion

References