

UNIVERSITY OF ST ANDREWS

MACHINE LEARNING

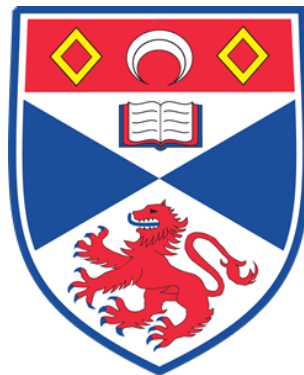
CS5014

Classification

Author:

150008022

March 28, 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

I	Loading Data	1
II	Cleaning Data	2
III	Data Visualisation and Analysis	2
1	Distributions	3
2	Relationships	3
IV	Feature Selection	3
V	Model Selection and Training	3
3	Model 1	3
4	Model 2	3
VI	Evaluation and Comparison	3
VII	Discussion	3

Part I

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, using stratification to ensure that all classes were represented in the training data.

Part II

Cleaning Data

.

Part III

Data Visualisation and Analysis

1 Distributions

2 Relationships

Part IV

Feature Selection

Part V

Model Selection and Training

3 Model 1

4 Model 2

Part VI

Evaluation and Comparison

Part VII

Discussion

References