Exercise 1 (Modelling inputs/outputs) - Solution

① a) Somerville Happines Survey Data Set

143 instances

This data set is taken from a survey conducted in Somerville, Massachusetts that was used to measure the overall happines of its residents. It contains information about the quality level of 6 aspects of the life in Somerville:

$x_1 \rightarrow$ the availability of information about city services

$x_2 \rightarrow$ the cost of housing

$x_3 \rightarrow$ the overall quality of public schools

$x_4 \rightarrow$ your trust in local police

$x_5 \rightarrow$ the maintenance of streets and sidewalks

$x_6 \rightarrow$ the availability of social community events

All of these attributes take values from 1 to 5 (whole numbers), so from low to high quality.

Each instance is accompanied by a decision attribute (D) with values 0 meaning unhappy and 1 for happy.

b) Modelling the dataset via input/output RVs

$$S = \{(\underbrace{x_i}_{\text{input}}, \underbrace{y_i}_{\text{output}})\}_{i=1}^{143} \quad \text{where}$$

$$x_i = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \qquad \text{and } y_i \in \{0, 1\}$$

$x_1, x_2, x_3, x_4, x_5, x_6 \in \{1, 2, 3, 4, 5\}$

c) A question that can be asked is:

Given the answers of a resident on the quality of the 6 aspects of life in Somerville, can we predict if he is happy or not?

This question can be solved using a supervised (output is included in data set) ML algorithm for classification. (output takes only 2 values)

2) a) Wine Quality Data Set

1599 instances of red wine
4898 instances of white wine } 2 data sets

These two data sets contain information about physiochemical characteristics of (red and white respectively) variations of the Portugese "Vinho Verde" wine. Each instance of wine has been given a grade from 0 to 10 from a wine expert. In total there are 11 input variables and 1 output variable (the quality).

Input	Output
1 - fixed acidity	12 - quality
2 - volatile acidity	
3 - citric acid	
4 - residual sugar	
5 - chlorides	
6 - free sulfur dioxide	
7 - total sulfur dioxide	
8 - density	
9 - pH	
10 - sulphates	
11 - alcohol	

There is a problem with the data set however. It is unbalanced, because it contains a lot more normal quality wines than excellent or poor ones.

b) Modeling the dataset via input/output RVs

Red wine

$$S_r = \{(x_i, y_i)\}_{i=1}^{1599}$$ (input, output)

White wine

$$S_w = \{(x_i, y_i)\}_{i=1}^{4898}$$ (input, output)

For both:

$$x_i = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \end{bmatrix}$$

→ the ranges of each are unknown but can be found if the dataset is further analysed

and $y_i \in \{0, 1, 2, \ldots, 10\}$

→ Some of the input RVs may be correlated, so a dimensionality reduction method may also be used.

c) Given the physiochemical parameters of an instance of (red/white) wine, what is its quality?

This question can be answered using a supervised (given actual output) ML algorithm for regression. Although, classification can be used as well (since output is discrete), it would make more sense to use regression. This is because the wine is being rated. A wine of quality 8 is closer to one of quality 10 rather than another one of quality 2 for example. Usually in classification we assume the classes to be equidistant from each other. It is not the case here.