CIS6930 Fall 2017: Introduction to Data Mining
Project I: Classification

September 22, 2017

Project Description

This project aims to make you familiar with the classification packages, available in R to do some easy data mining analysis on real-world problems. You need to apply several different classification methods on the given datasets and submit:

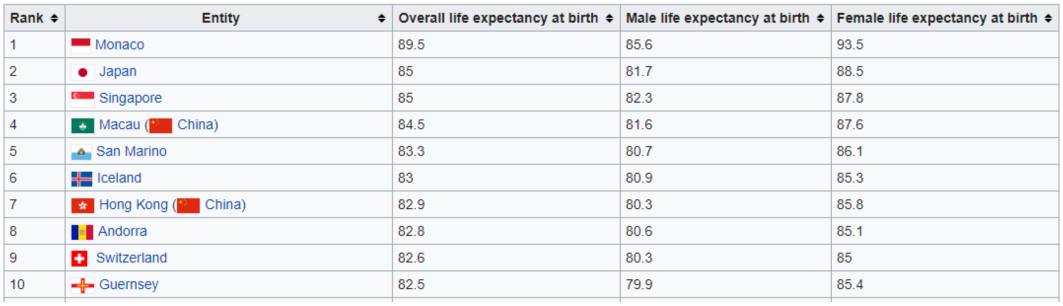
- A detailed report showing:
  - Description of the dataset preparation (*e.g.* data pre-processing, random training / test set creation, *etc.*).
  - Description of the classification methods you use and the choice of the parameters.
  - Classification results and analysis (e.g. confusion matrices, accuracies, precision-recall curves, F-Measure, *etc.*)
  - Your conclustion and reference list
- A *Readme.txt* file, explaining how to run your script.
- A .R script, containing every function definition and necessary comment. By running this script, we should be able to get your reported results.

Submit your files as a .zip file to Canvas, with the name format as: **Firstname_Lastname_UFID.zip**.

Dataset

Life Expectancy Table is the main dataset you are going to use for this project. It can be obtained from the Wikipedia Website (see the section "List by the CIA (2016)"). Figure 1 shows the top 10 rows of the table provided on the page.

To use this dataset, you need to add a *Continent* column. It is not provided in the given link, but you can easily find the continent of every country by a simple search. This additional column plays as



Rank	Entity	Overall life expectancy at birth	Male life expectancy at birth	Female life expectancy at birth
1	Monaco	89.5	85.6	93.5
2	Japan	85	81.7	88.5
3	Singapore	85	82.3	87.8
4	Macau (China)	84.5	81.6	87.6
5	San Marino	83.3	80.7	86.1
6	Iceland	83	80.9	85.3
7	Hong Kong (China)	82.9	80.3	85.8
8	Andorra	82.8	80.6	85.1
9	Switzerland	82.6	80.3	85
10	Guernsey	82.5	79.9	85.4

Figure 1: A partial screenshot of the Life Expectancy Dataset. Totally, there are 223 rows and 5 columns with the rank, entity name, overall life expectancy at birth, male life expectancy at birth and female life expectancy at birth.

the class label role. Preserve the other columns as the features (except the column of *Entity*). Note that when you copy the data from the table, you need to add this resource to the *reference* on your report.

Before classification, you need to divide your dataset into training set (80%) and test set (20%). Randomly select samples for training set and test set and independently repeat to create multiple groups are highly recommended (*e.g.* 5 different groups of training set and test set). Standard deviation and average value can be used when you evaluate the classification accuracies from different groups.

Classification Methods

For any mentioned classification method, try to find the right package and method. If you cannot find it, you have to write your own method. You may want to tune the classification method by providing the right arguments, based on your understanding on the dataset in use. You should clearly mention why you choose such parameters (for instance, the number of nearest neighbors in *kNN* method, *i.e.* the *k*) using the below classification methods:

- k-Nearest Neighbor (kNN)
- Decision Tree: either two algorithms of the RIPPER, C4.5 and Oblique
- Support Vector Machine

Note that the final reported results should be about your classification performance on the *test set*.

R Script

Along with your report, you should submit your script (.R file) too. Your script should contain every line of code to be executed to get the final results. Here is an example of your function prototype:

```
myC45 <- function(...) {
        # learns a fit based on the given training set
        # returns a fit
}
myC45Predict <- function(...) {
        # for each sample in test set, it predicts the label.
        # returns the required values (accuracy, precision, ...)
}
# Other function definitions

# Reading datasets and dividing into training and test sets
divideDataset <- function(...){
        set.seed(2017)
        # returns:
        # dataset$training
        # and dataset$test
}
```

If we are not able to run your script (by following your *Readme.txt* file), you will lose most of the points. Your results should be consistent with what you report. If there is a specific variable that holds the final values, you should mention it in your Readme.txt so we know how to check your results.