INFS4203/7203 Assignment

Semester 2/2018

Marks:	100 marks (20%)
Due Date:	23rd October 2018, 11:55PM
What to Submit:	See deliverables part
Where to Submit:	Electronic submission via blackboard

The goal of this project is to gain practical experience in applying clustering and classification to real data. You must work on this project **individually**. The standard academic honesty rules apply. You must use R for this project.

There are three main tasks: Data Preparation, Clustering, and Classification. Please read the report carefully until the end. Since some parts of your code require a random seed, you need to pre-set the seed for reproducible results. Your seed value is the last four digits of your student ID.

Dataset:

You will be using the "**Breast Cancer Wisconsin (Original)**" data[1]. The data has 699 observations (rows) and 11 attributes (columns). One of the attributes is a binary class variable with value 2 for "benign" and 4 for "malignant". You may find and learn more detailed information about the data in the data description page[2]. Please note that some of the observations can be missing (usually filled with "?").

---

[1] Data Folder: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ . Download file breast-cancer-wisconsin.data .

[2] Data Description: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

Task 1 - Data Preparation:

Write the necessary code to pre-process the data. That pre-processing stage includes the following tasks:

1.1. Extract the data into an R data frame.

1.2. Assign the following names to the 11 different columns in your dataset

```
1. Sample code number
2. Clump Thickness
3. Uniformity of Cell Size
4. Uniformity of Cell Shape
5. Marginal Adhesion
6. Single Epithelial Cell Size
7. Bare Nuclei
8. Bland Chromatin
9. Normal Nucleoli
10. Mitoses
11. Class
```

When applicable, change any space character to dot (e.g. "`Normal Nucleoli`" to "`Normal.Nucleoli`").

1.3. Remove all rows with missing values. Notice that R might define a column with missing data as "*factor*". If such column is supposed to be integer, then your need to change the column type from factor to integer.

1.4. Remove the first column (`Sample code number`) as it is not useful for our next tasks.

1.5. Notice that R might define the "*class*" column as integer. In that case, change its type from integer to factor.

1.6. Save the dataframe into a file with filename `bcw_processed.Rda` .

Task 2 - Clustering:

Apply K-Means and Hierarchical clustering to cluster the data as follows.

2.1. Load the preprocessed data file from Task 1 into a data frame. Please note that for this set of clustering tasks, you should not include the `Class` column.

2.2. Cluster the data into 2 clusters using K-Means clustering, using the default parameters for the `kmeans` function. Plot the results of the clusters as a 2D plot where the x-axis is `Clump Thickness` and the y-axis is `Uniformity of Cell Size`.

2.3. Plot another 2D plot with the same dimensions above, but color the points according to the `Class` column.

2.4. Compare the 2 plots obtained in the previous two tasks – do the clusters visually represent the `benign` vs `malignant` classes?.

2.5. Cluster the data into more than 2 clusters (i.e., k = 3, 4, 5) using K-Means clustering and plot all the clustering results.

2.6. Compare the plots and SSEs obtained in the previous task, and provide your comments on the quality of clustering.

1. all sse (suppose i have two clusters) 2.sse in one cluster and sse in another cluster respectively 3.sum of sse+sse

2.7. Apply hierarchical clustering to the data using the `hclust` function with default parameters and plot the corresponding dendrogram. Particularly, cluster the dendrogram into 2, 3, 4, and 5 clusters and plot all of them.

4. 1-3

2.8. Compare the plots obtained in the previous task and provide your observations on the achieved clusters - should we have a new subtype of diseases?.

2.9. Try different agglomeration methods in hierarchical clustering (i.e., "`single`", "`complete`", and "`average`"). Plot the resulting dendrograms and provide your comments on the quality of clustering - is the data sensitive to the used agglomeration method? Based on your results, what do you think is the default agglomeration method used in Task 2.7?

Task 3 - Classification:

Apply binary classification using decision tree and K-NN techniques.

3.1. Load the preprocessed data file from Task 1 into data frame. Divide the dataset into “training” and “test” subsets randomly (70% and 30% respectively).

3.2. Learn a classification tree from the training data using the default parameters of the `ctree` function from the “`party`” library. Plot that classification tree and provide your comments on its structure (e.g., what are the important/unimportant variables? Is there any knowledge we can infer from the tree representation that helps in differentiating between the classes?). Using the learned tree, predict the class labels of the test data. Calculate the accuracy, precision, and recall.

3.3. Try building you classification tree again via the `ctree` function but using parameters that are different from the default settings. Can you achieve better accuracy or more meaningful representation by tuning some parameters? (Note that in the `ctree` function from “`party`” library, you can modifiy `ctree_control` parameters. Execute `?ctree` form RStudio Console for the detailed documentation.)

3.4 Apply K-NN classification to predict the labels in the test subset and calculate the accuracy, precision and recall. Particularly, try different values of K (e.g. K = 1, 2, 3, 4, 5), and report your observations on the achieved classification.

Deliverables: R project with your student number as the project's name (e.g. `12345678.Rproj`). The project should have the following folders:

1) **Code**:
    - `myPreparation.r` : Code to complete Task 1.
    - `myClustering.r` : Code to complete Task 2.
    - `myClassification.r`: Code to complete Task 3.

    Provide the appropriate header in each file (your identity and file description) and give meaningful comments in the script.
2) **Data**:
    - `breast-cancer-wisconsin.data` : orgininal dataset.
    - `bcw_processed.Rda` : preprocessed data output from Task 1.
3) **Plot**:
    - All plots generated in Tasks 2.2, 2.3. 2.4, 2.5, 2.7, 2.9, 3.2, and 3.3.
4) **Report**: your report should include the following:
    - Brief description for the main functions in your source code and any assumptions or special settings of those functions.
    - Plots, evaluations, and your comments on the observed results.

Please note that your preprocessed data, plots, and results should be reproducible. That is, we can delete them and be able to generate them by running your code. Hence, remember to set the seed before any function that requires a random value. That seed is the last 4 digits of your student number.

Marking: Your total mark earned for this assignment is based on:

- Report: accurate statistics and clear presentation;
- Code and reproducible results.
- Demo: one-on-one demo presentation, if needed.

Submit one archive file with your student number as the file name (e.g. `12345678.zip`) with all the files and folders mentioned above. The project is due **11:55PM, 23rd October 2018**. No late submission is allowed.