Quiz 3

Subject: Functions and collections
TAs: Bahar GEZİCİ, Nebi YILMAZ, Ahmet Selman BOZKIR

Due Date: 04.12.2018 23:00
Accept your *3rd Quiz*.

Introduction

Within the context of this quiz, you are going to gain experience on using different collections (dictionary, list and their functions) as well as on using functions. The scenario of this work is to analyze sample taken from an imaginary patient to find out if he/she shows symptoms of breast cancer.

To do that you will make use of a well known breast cancer database, *Wisconsin Breast Cancer Database*, or WBC Database, containing 10 attributes including a class label (`benign` and `malignant`) of 699 instances [**?**]. Not all databases in the real world are ready to use and they often require a pre-process so that you use them on your projects. WBC database is not exception and possessing missing values. We intentionally left missing values in WBC as they are because we are asked you to tackle with them, which constructs the 1st phase of this quiz work.

The attribute and domain information of WBC database is given below:

ID	Description	Domain
attr1	Clump Thickness	1 - 10
attr2	Uniformity of Cell Size	1 - 10
attr3	Uniformity of Cell Shape	1 - 10
attr4	Marginal Adhesion	1 - 10
attr5	Single Epithelial Cell Size	1 - 10
attr6	Bare Nuclei	1 - 10
attr7	Bland Chromatin	1 - 10
attr8	Normal Nucleoli	1 - 10
attr9	Mitoses	1 - 10
Class[1]		benign & malignant

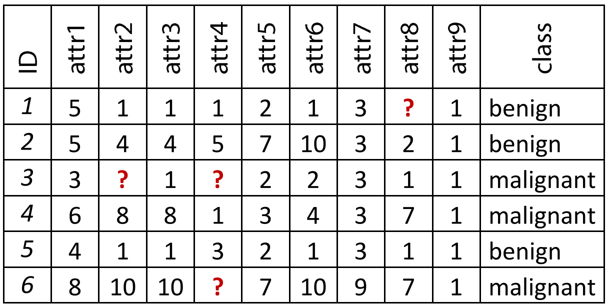
[1] Class distribution: Benign: 458 (65.5%) Malignant: 241 (34.5%)

Cleaning WBC Database

As mentioned, there are 21 missing attribute values (denoted by '?') in modified WBC database. In this phase you are expected to remove and replace them with the appropriate values.

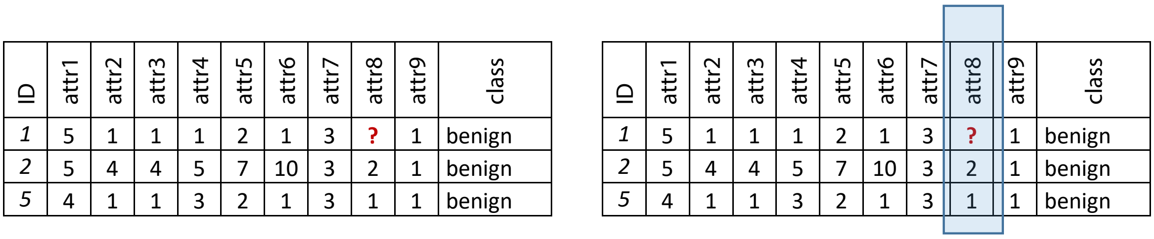
The calculation of appropriate values is as follows:

i) Consider a database shown below. There exists 4 missing attribute values; 1 and 3 of them belong to benign and malignant classes, respectively.



ID	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	class
1	5	1	1	1	2	1	3	?	1	benign
2	5	4	4	5	7	10	3	2	1	benign
3	3	?	1	?	2	2	3	1	1	malignant
4	6	8	8	1	3	4	3	7	1	malignant
5	4	1	1	3	2	1	3	1	1	benign
6	8	10	10	?	7	10	9	7	1	malignant

ii) To find appropriate value for the 8th attribute of the 1st record, take only the records belonging to the same class (benign in this example) into consideration. So you have only records at left panel of the figure below for calculation.



ID	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	class
1	5	1	1	1	2	1	3	?	1	benign
2	5	4	4	5	7	10	3	2	1	benign
5	4	1	1	3	2	1	3	1	1	benign

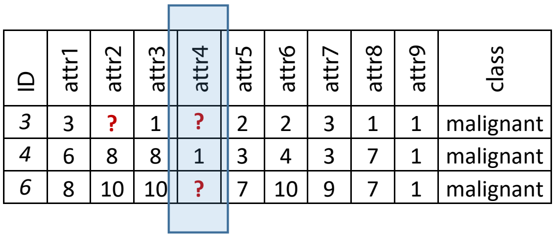
ID	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	class
1	5	1	1	1	2	1	3	?	1	benign
2	5	4	4	5	7	10	3	2	1	benign
5	4	1	1	3	2	1	3	1	1	benign

iii) After ignoring the records of opposite class (malignant), consider only the attribute values where missing value belongs to (8th attribute for this example, see right panel of figure above). The calculation of Appropriate Value (AV) of **attr8** is as follow:

$$AV = round(\overline{attr8}) \quad | \quad i \neq '?' \;\; and \;\; i \in attr8$$

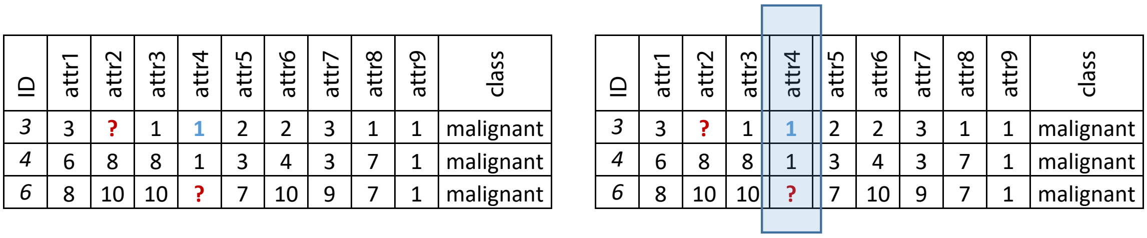
Concretely speaking, AV is a value obtained from the calculation of average values of **attr8** of each record belonging to same class whose **attr8** value is not missing. As only integer values present in WBC dataset, you need to *round* the resulting average value to the nearest integer value.

To elaborate further, the calculation of missing attribute values in **attr4** is also illustrated below:



ID	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	class
3	3	?	1	?	2	2	3	1	1	malignant
4	6	8	8	1	3	4	3	7	1	malignant
6	8	10	10	?	7	10	9	7	1	malignant

In this example the average value of attr4 of 3rd record ($\overline{attr4}$) is equal to 1 (see left panel of the figure below). As the mean value of attr4 remains same, the AV value for the missing value in 6th record is also equal to 1 (see right panel of the figure above).



ID	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	class
3	3	?	1	1	2	2	3	1	1	malignant
4	6	8	8	1	3	4	3	7	1	malignant
6	8	10	10	?	7	10	9	7	1	malignant

ID	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	class
3	3	?	1	1	2	2	3	1	1	malignant
4	6	8	8	1	3	4	3	7	1	malignant
6	8	10	10	?	7	10	9	7	1	malignant

Within the context of this phase, you should define a function `funDataClean` that performs data cleaning process described here.

After cleaning process it should print out a message containing the average of all calculated appropriate values in the following format (for `WBC.data` file):

```
The average of all missing values is  : 3.0476
```

Retrieving knowledge from WBC dataset

This phase is constructed on a scenario where you take sample from an imaginary patient that is necessary for the diagnosis of breast cancer. Instead of having exact values, we have only *conditions* containing i) *relational operators* and ii) *numeric values* to calculate the probability of him/her to show symptoms of breast cancer.

The relational operators and their descriptions are provided below:

#	Operator	Description
1	<	less than
2	<=	less than or equal to
3	>	greater than
4	>=	greater than or equal to
5	!=	not equal to
6	=	equal to
7	?	any value

In order of your program to calculate the probability, one should provide the samples as **command-line argument** within the following format:

```
python 1.py <:11,!=:0,>=:1,>:5,?,?,<=:9,>:2,?
```

As you see from the input format, there are 9 conditions each of which is separated by commas (,) and 1st and 2nd conditions correspond to the 1st and 2nd attributes (attr1, attr2), respectively; and so on. Relational operator in each condition (except the one having '?') is also separated from its numeric value by colons (:).

Once such argument is provided to your program, only the records whose

```
attr1 is less than 11
attr2 is not equal to 0
attr3 is greater than or equal to 1
attr4 is greater than 5
attr5 is any value
attr6 is any value
attr7 is less than or equal to 9
attr8 is greater than 2
attr9 is any value
```

are considered for calculation.

Finally your program should display the findings (for the conditions above) in the following format:

```
Test Results:
----------------------------------------------
Positive (malignant) cases            : 82
Negative (benign) cases               : 3
The probability of being positive     : 0.9647
----------------------------------------------
```

Within the context of this phase, you should define a function `performStepWiseSearch` that performs step-wise search, as the name suggests.

Another example:

Given that you have taken a sample within the following conditions where:

```
attr2 is not equal to 9
attr6 is greater than or equal to 3
attr8 is less than 8
attr9 is equal to 1
```

then the order of command-line arguments should be provided as follows:

```
python 1.py ?,!=:9,?,?,?,>=:3,?,<:8,=:1
```

The output of your program should exactly match with:

```
The average of all missing values is  : 3.0476

Test Results:
----------------------------------------------
Positive (malignant) cases            : 84
Negative (benign) cases               : 31
The probability of being positive     : 0.7304
----------------------------------------------
```

Notes specific to this quiz work

- As highlighted before, every record in WBC database comprises of **record ID**, **attributes** from *attr1* to *attr9*, and a **class label**. The content of any record in WBC is as below:

  `Xth record -> X,3,7,7,4,4,9,4,8,1,malignant`

  For this quiz you do not need to think how to read a file as we have already did it for you. (see the content of starter code file `1.py`). In this file you can access the every record through a dictionary named `dataDic`. The key and value pair of `dataDic` in the file is as given below (exemplified with respect to the record above):

  `{'X'} -> {['3','7','7','4','4','9','4','8','1','malignant']}`

  The left and right sides of arrow are key and value of `dataDic`, respectively.
- We will take **print format** into consideration while evaluating your works, so your output should exactly match the print format provided here including the dashed lines and white spaces.
- Be aware that we will modify the content of WBC dataset to validate your program is not input dependent.
- Your program should print out total benign and malignant cases in WBC dataset in the case of all conditions are provided as '?'.

General Notes

- Even if we intentionally say `python` to indicate how to run your program with the argument, you should test your work on our department's dev server by typing/calling `python3`.
- Do not miss the submission deadline.
- Save all your work until the quiz is graded.

- You can ask your questions via Piazza and you are supposed to be aware of everything discussed on Piazza.
- You must submit your work with the file hierarchy as stated below:

→ <1.py>