Final Project

Programming for Informatics

BIN500, Fall 2019 – 2020

Important Dates

Final project - Python codes submission: 23:59 on Jan 6, 2020
Final project - Demo session: Jan 7, 2020 Reserve your spot at doodle link (attendance required, https://doodle.com/poll/dswuuwb6fin94ffz)

Your final submission must cover all or a subset of the topics covered in the lecture. Your project must include exception handling and programmer-defined functions.
The final code must be modular with functions.
Your program must be commented properly.

Deliverables

You must submit your code at ODTUClass with the .py extension– this is your source code solution; be sure to include the date, the project number and comments describing your code.

Project Specification

Given a data file containing hundreds of records with values describing cancer tumors and whether or not each tumor is malignant or benign, develop a simple rule-based 'classifier' that can be used to predict the class (malignant or benign) of a set of unknown records.
Here is the general idea: Malignant tumors are different than benign tumors. Malignant tumors tend to have larger radii, to be smoother, to be more symmetric, etc. Measurements have been taken on many tumors whose status (malignant or benign) is known.
The code you are going to write will get the average score across all the malignant tumors for an attribute (e.g. 'area') as well as the average score for that attribute for benign tumors. Let's say that the average area for malignant tumors is 100, and it is 50 for benign tumors. We can then use that information to try to predict whether a given tumor is malignant or benign. Imagine I presented you with a new tumor and told you the area was 99. All else being equal, we would have reason to think this tumor is more likely to be malignant than had its area been 51. While real classifiers use more complicated rules, we are going to create a very simple classification scheme. We will calculate the mid-point between the malignant average and the benign average (75 in

our hypothetical example), and simply say that for each new tumor, if its value for that attribute is greater than or equal to the midpoint value for that attribute, that is one vote for the tumor being malignant. Each attribute that we are using produces a vote, and at the end of counting votes for each attribute, if the malignant votes are greater than or equal to the benign votes, we predict that the tumor is malignant.

To ‘learn’ what the midpoint values are we need access to attribute values of many tumors whose status is known. You can find these in the training set file. Once we have the classifier (which is the midpoint values), we will then predict the status of the tumors in the test data file. It turns out that we also know the status of the tumors in the test data file. As such, a final step will be to compare the prediction your classifier makes to the real value, to see how accurate your classifier is.

In this project, you will write a program to predict whether or not a cancer tumor is malignant or benign when given the mean for each of 10 attributes describing the tumor. Each tumor is described by the mean for 10 values for this project. Each tumor also has data for the identification number and the class (malignant or benign). That makes 10 total values in the file per tumor. Look at the top of the data files (e.g. cancerTraining.txt) to see all 10 values.

Here are the 10 tumor attributes:

1. radius
2. texture
3. perimeter
4. area
5. smoothness
6. compactness
7. concavity
8. concave
9. symmetry
10. fractal

The mean values are described with the suffix ‘_mean’ in the data files. For the class label, the M stands for malignant and the B for Benign. We’ll use these attributes to predict the value of the class (malignant or bening) for a new tumor. Note the ID number has no bearing on the tumor’s class and is used only to differentiate tumors. A single set of attributes for a single tumor is known as a record. We don’t need to know what any of the attributes’ names mean (what does ‘fractal’ mean in this context?). All we need to know is that they are measurements of the tumors, and that benign and malignant tumors tend to have different attribute values. For these 10 attributes when comparing means, higher numbers indicate malignancy.

There are several tasks associated with this project.

(60 pts to complete tasks 1 to 4)

1. **Train a simple classifier.**

A classifier is a model of the problem such that when we're given a new record we can compare the new record to the model in order to predict the class of the new record. We use the training set to build up this model. Our model is very simple. For all malignant records, for each attribute, we calculate the average value of each attribute. For all benign records, for each attribute, we calculate the average value of each attribute.
To create the model, we then calculate the midpoint of these averages for each attribute. Then to classify new records, if the majority of the new record's attributes are above their respective midpoints, then the new record is predicted to be malignant, otherwise, benign.
Definition of the majority can be multiple:

- **a.** 5 or more of the new record's attributes are above their respective midpoints, then the new record is predicted to be malignant, otherwise benign
- **b.** 6 or more of the new record's attributes are above their respective midpoints, then the new record is predicted to be malignant, otherwise benign
- **c.** 7 or more of the new record's attributes are above their respective midpoints, then the new record is predicted to be malignant, otherwise benign.

Desgin your code accordingly.

1. **Apply the classifier to test set and report accuracy of classifier.**

For each record in the test set, test all the majority definitions. For each record in the test set, compare the predicted class to the actual class and then print out the accuracy of the classifier both as number correct / total number and as a percentage. Find the best predictor based on the accuracy that can be used in the coming parts.

2. **Output Results**

You should provide two other kinds of output besides the accuracy. First, report the statistics you gathered (the malignant and benign averages, as well as the classifier midpoint value). Second, you should provide some feedback on an individual patient. The system should prompt for a patient ID from the test set, and then provide the patient values from the test set, the classifier cutoff for that value and the diagnosis for that particular value, as well as the patient's overall predicted diagnosis (see example output).

3. **Additional Tasks**

a. **(20 pts)** Configure your code so that a user can type a record for a patient and your code must predict the class and inform the user. Here, if the given record is not correct you must ask for another input.

b. **(20 pts)** Given the model constructed in your code, generate a new data set composed of 250 patients with random patient ID of which 50% must be benign and 50% must be malignant. Select 10 tumor attributes such as texture, concavity etc. randomly for each patient. This random selection must use the average values you have calculated for each attribute.

Notes:

- **a.** Most importantly, demo output is provided below. You can look at the example output to see how the program should run on the provided data files. Make your output look exactly like the output in the demo!
- **b.** The tasks have to be completed in order. You obviously can't use a classifier before you've trained it.
- **c.** Don't try to tackle this project all at once. Complete one function (or part of a function) and test it out.

Deliverables

You must submit your code at ODTUClass with the .py extension– this is your source code solution; be sure to include the date, the project number and comments describing your code.

Example Output

Note: user input is in bold (just the patients' ids to check and 'quit' to quit), everything else is output by the program.

```
Classifier, benign and malignant stats
======================================================================
                         Key    Malignant   Classifier       Benign
                                  Average     Midpoint      Average
                      radius       17.075       14.545       12.016
                     texture       21.385       19.279       17.174
                   perimeter      112.687       94.919       77.152
                        area      934.017      693.338      452.659
                  smoothness        0.103        0.098        0.093
                 compactness        0.144        0.110        0.077
                   concavity        0.153        0.100        0.046
                     concave        0.084        0.055        0.025
                    symmetry        0.194        0.185        0.175
                     fractal        0.063        0.063        0.063

Reading in test data...
Done reading test data.

Classifying records...
Done classifying.

The classifier correctly predicted the class (malignant/benign) of 213
records out of 231 records.
The classifier achieved an accuracy of 92.21 percent.

Type an ID to check a patient ('quit' to stop):897880
Checking ID:897880's classification
                         Key      Patient   Classifier        Class
                                    Value       Cutoff
                   perimeter       64.410       94.919       Benign
                    symmetry        0.189        0.185    Malignant
                        area      310.800      693.338       Benign
                     concave        0.018        0.055       Benign
                     texture       17.530       19.279       Benign
                   concavity        0.025        0.100       Benign
                      radius       10.050       14.545       Benign
                 compactness        0.073        0.110       Benign
                     fractal        0.063        0.063    Malignant
                  smoothness        0.101        0.098    Malignant

Overall Diagnosis for patient 897880: Benign

Type an ID to check a patient ('quit' to stop):89812
Checking ID:89812's classification
                         Key      Patient   Classifier        Class
                                    Value       Cutoff
                   perimeter      155.100       94.919    Malignant
```

```
                    symmetry        0.180        0.185       Benign
                        area     1747.000      693.338    Malignant
                     concave        0.141        0.055    Malignant
                     texture       24.270       19.279    Malignant
                   concavity        0.231        0.100    Malignant
                      radius       23.510       14.545    Malignant
                 compactness        0.128        0.110    Malignant
                     fractal        0.055        0.063       Benign
                  smoothness        0.107        0.098    Malignant

Overall Diagnosis for patient 89812: Malignant

Type an ID to check a patient ('quit' to stop):quit
Program finished.
```