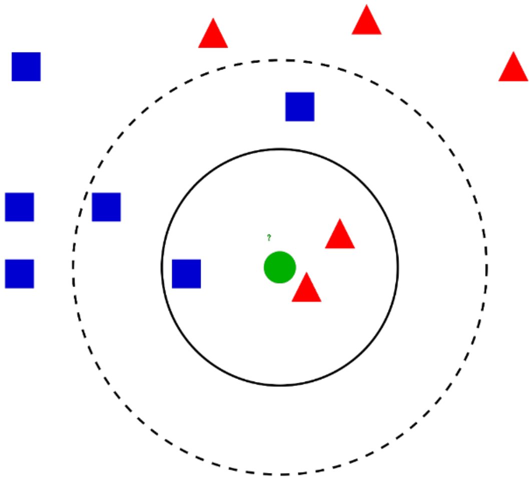
COSC 1306
Spring 2020

Assignment 5: Lists

[1] **Objectives**: The primary purpose of this assignment is to make sure everyone is familiar with the list data structure. In particular, we will be using a list of lists as our main data structure to hold points in a 2-D plane. This may not be the best choice. A list of tuples may actually be better, but we have not learned tuples yet. For a similar reason, we will hard-wire the input data into the program at this time. You are required to use functions to make your main function simple and easy to understand.

[2] **Requirements**: We are going to implement a VERY simple kNN (k-nearest neighbors) algorithm in this assignment. KNN is a supervised classification machine learning algorithm. (Did I scare you? It's actually fairly simple.) Given some training data with labels, kNN allows us to make a prediction on which label should we assign to test data.



Our data are points on the plane similar to the one we used in Assignment 5. You may restrict the coordinates to integer numbers. Each point is labeled with color (blue or red in the above example). Again, we simplify the problem to only two classes (Label 0 and Label 1). The six blue squares form a cluster and the five red triangles form another one. The question: Given an unlabeled green point, should we classify it as red or blue?

Here is how kNN decides the class:

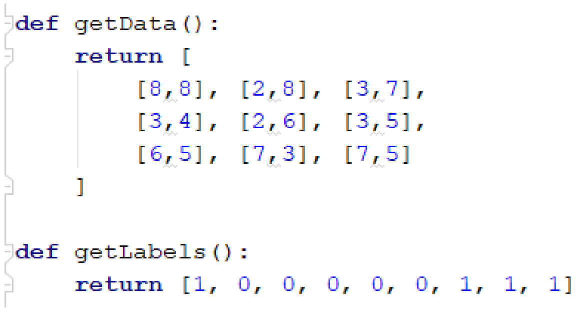
- Calculate the distance between test data and all training data. Here we will use Euclidean distance as our distance metric. Store the distance in another list to be used later.
- Sort the calculated distances in ascending order based on distance values. **Keep the original distance list unchanged.**
- Find the k (= 3) nearest neighbors (any color).
- The predicted class of the test point is determined by a vote of the three nearest neighbors.

Since there are two red and only one blue, the test point will get the red color. Now you know why k is set to 3, not 2 or 3.

Reminder:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} .$$

All the training points are stored in a list such as [[1,2], [2,3] …]. Their labels are in a separate list [0, 1, 0, … 0]. In this assignment, I will give you the values. Your answer should be 0 or 1 for a given test data.



```
def getData():
    return [
        [8,8], [2,8], [3,7],
        [3,4], [2,6], [3,5],
        [6,5], [7,3], [7,5]
    ]

def getLabels():
    return [1, 0, 0, 0, 0, 0, 1, 1, 1]
```

To test multiple points, please add a loop (see sample output).

Here is a simplified algorithm.

- Get the data and labels
- Print the data and labels
- Get the k (even though it is 3)
- Repeat the following until the user wants to quit.
    - Get the test point
    - Compute and store the distance from test to all points
    - Sort the distance list and select a threshold to separate the three smallest distance out. A good choice will be the average of the 3rd and 4th distance in the sorted list.
    - Print the distance list (both)
    - Let them vote.
    - Ask if the user want to quit.

Here are the additional **suggested** functions to use. You are free to use more functions.

- `dist(p1, p2)` returns distance between the two points
- `getDist(dataList, test)` returns a distance list
- `getCut(dlist, k)` returns the cut off value for the distance
- `vote(dlist, labels, cut)` returns the vote result

[3] **Output**: See sample output below and the demo in class.

```
[x, y] Label
[8, 8]   1
[2, 8]   0
[3, 7]   0
[3, 4]   0
[2, 6]   0
[3, 5]   0
[6, 5]   1
[7, 3]   1
[7, 5]   1

Enter k: 3

X-coordinate: 3
Y-coordinate: 3
```

```
Distance from test to all points:
  7.07,   5.10,   4.00,   1.00,   3.16,   2.00,   3.61,   4.00,   4.47,
Sorted list:
  1.00,   2.00,   3.16,   3.61,   4.00,   4.00,   4.47,   5.10,   7.07,

***  0
***  0
***  0
Test [3, 3] belongs to Group  0

Do you want to test  more cases (Y/N): y

X-coordinate: 4
Y-coordinate: 4

Distance from test to all points:
  5.66,   4.47,   3.16,   1.00,   2.83,   1.41,   2.24,   3.16,   3.16,
Sorted list:
  1.00,   1.41,   2.24,   2.83,   3.16,   3.16,   3.16,   4.47,   5.66,

***  0
***  0
***  1
Test [4, 4] belongs to Group  0

Do you want to test  more cases (Y/N): Y

X-coordinate: 5
Y-coordinate: 5

Distance from test to all points:
  4.24,   4.24,   2.83,   2.24,   3.16,   2.00,   1.00,   2.83,   2.00,
Sorted list:
  1.00,   2.00,   2.00,   2.24,   2.83,   2.83,   3.16,   4.24,   4.24,

***  0
***  1
***  1
Test [5, 5] belongs to Group  1

Do you want to test  more cases (Y/N): n
```

[4] **Deadline**: Midnight, Monday, March 16, 2020.