Course: COMPSCI 260
Name: Divya Koyyalagunta
NetID: dk160
Problem: 1
Problem Set: 5
Due: Fri 7 Nov 2016, 5pm
Using free extension (yes/no): no

Statement of collaboration and resources used (put None if you worked entirely without collaboration or resources; otherwise cite carefully): None
My solutions and comments for this problem are below.

PROBLEM 1

a)

Name	Gene expression profile	k = 1	k = 3
Alice	(2,7)	responder	responder
Bobby	(10,7)	responder	non-responder
Cindy	(8,4)	non-responder	non-responder
Donny	(4,9)	non-responder	responder
Ellen	(8,8.5)	1:2 (responder : non-responder)	non-responder

For Alice, I can make a more confident prognosis since all three closest neighbors are of the same type (responder). For Bobby, the inconsistency between k being 1 and k being 3 makes me doubt the validity of the initial prognosis (responder). The distance disparity between 1 neighbor and 3 neighbors also raises doubts, since distance on the Cartesian plot could impact these results. In Cindy's case, many neighbors in close proximity (further than even k = 3) are all non-responders, and this thus increases the confidence for her prognosis of being a non-responder. Similar to Bobby's case, Donny has an inconsistency between the prognosis from k = 1 and k = 3, which reduces the confidence with which we can say that he really is closest to being a non-responder, since that is the prognosis from the first neighbor. In Ellen's case, we see that there is a three way tie between the nearest neighbor, and so we cannot make any prognosis, but can say that there is a 1:2 ratio between responder to non-responder, which matches of course with the k = 3 prognosis of being a non-responder, and increasing our confidence in this prognosis although nothing definitive could be said in the k = 1 case.

b) Using $k = 5$, the following is the prognosis for the ten patients:

$\{1{:}\,N, 2{:}\,R, 3{:}\,N, 4{:}\,R, 5{:}\,N, 6{:}\,N, 7{:}\,\mathrm{R}, 8{:}\,\mathrm{R}, 9{:}\,\mathrm{N}, 10{:}\,\mathrm{N}\}$

I included the 5 closest neighbors along with their distances and labels for each patient. Upon manual inspection, having the distances from the patient to the data points along with labels for these data points increases the confidence I have for the prognosis for each patient. We can see for some patients that the 5 closest neighbors are all of the same label as their prognosis. For some patients where the closest neighbor does not share their same prognosis label, the rest of the neighbors that are still a very close distance away to the patient share the same label.