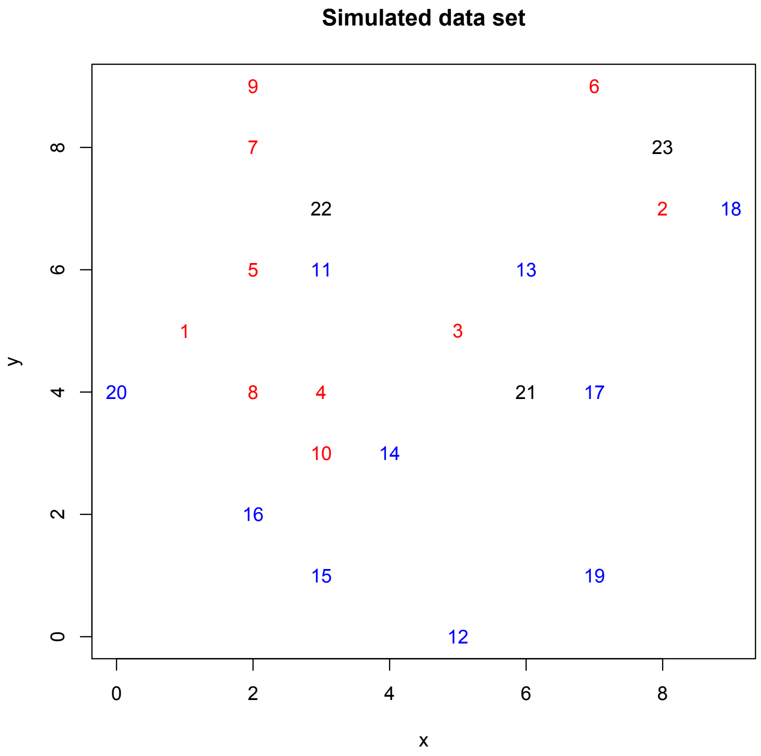
IMSA Data Mining Intersession

Day 2 Worksheet

k-Nearest Neighbors

We simulated a toy data set of 20 points, plotted below. Each data point has an observation in two variables: x and y. The red data points $(1 - 10)$ where generated differently than the blue data points $(11 - 20)$. Your task in this exercise is to implement k-Nearest Neighbors by hand to classify new data points $21 - 23$ (unlabeled, colored black) as either red or blue.



	x	y
1	1	5
2	8	7
3	5	5
4	3	4
5	2	6
6	7	9
7	2	8
8	2	4
9	2	9
10	3	3
11	3	6
12	5	0
13	6	6
14	4	3
15	3	1
16	2	2
17	7	4
18	9	7
19	7	1
20	0	4
21	6	4
22	3	7
23	8	8

The closer two points are to each other, the more similar they are. Use squared Euclidean distance: The distance between two points (x_1, y_1) and (x_2, y_2) is $(x_1 - x_2)^2 + (y_1 - y_2)^2$.

1. Classify the new data point 21 using 1-, 3-, 7- and 19-Nearest Neighbors.

2. Classify the new data point 22 using 1-, 3-, 7- and 19-Nearest Neighbors.

3. Classify the new data point 23 using 1-, 3-, 7- and 19-Nearest Neighbors.

4. Why did we only consider odd values of k in the above exercises?

5. What is a benefit of using a large value for k? What is a benefit of using a small value for k?