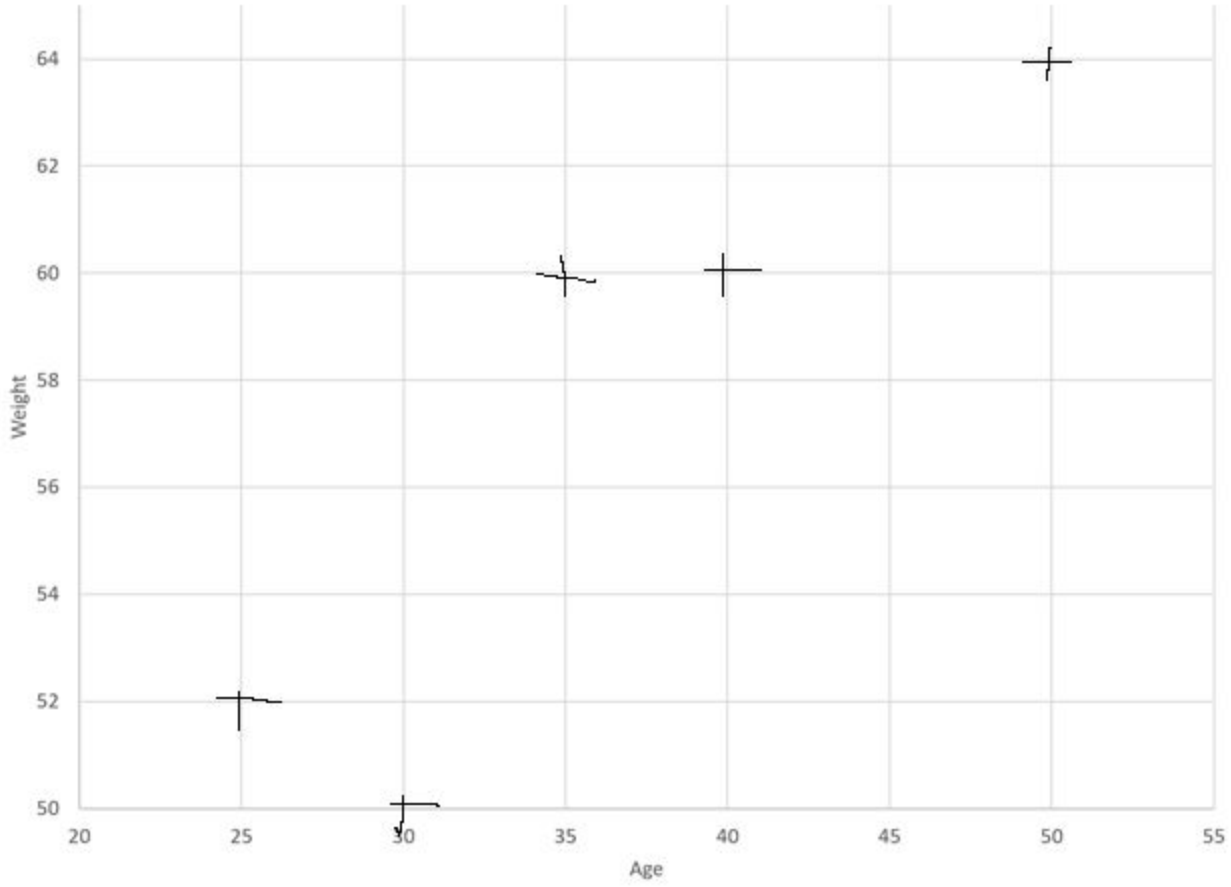
Alexander Rasho

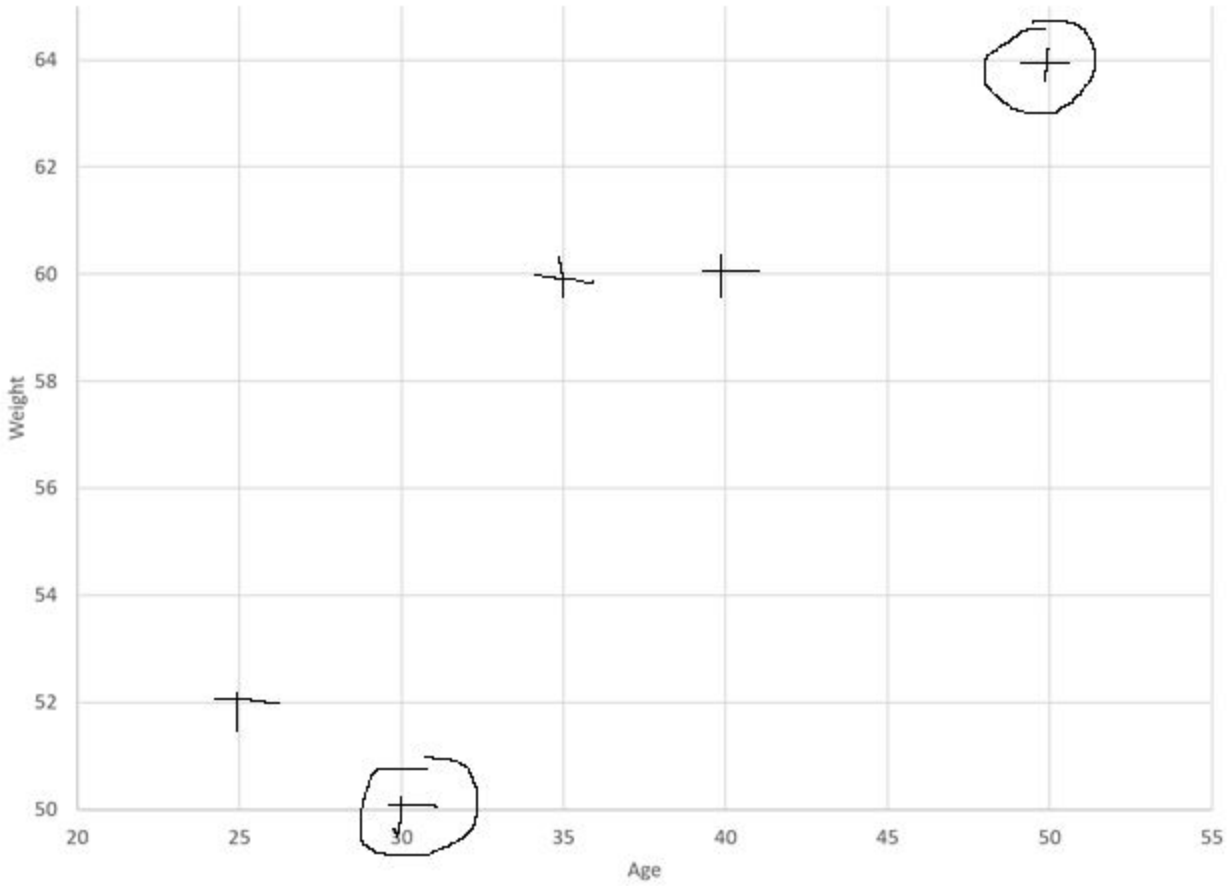
1. Consider the following dataset.

Age	25	30	35	40	50
Weight	52	50	60	60	64

a) Mark the data points on the graph below (use '+' to indicate each point).



b) Let k=2. Let one of the two initial centers be (Age=30, Weight=50). Select the second center using the Farthest Distance Heuristic. Indicate the two centers on the graph (circle the centers).

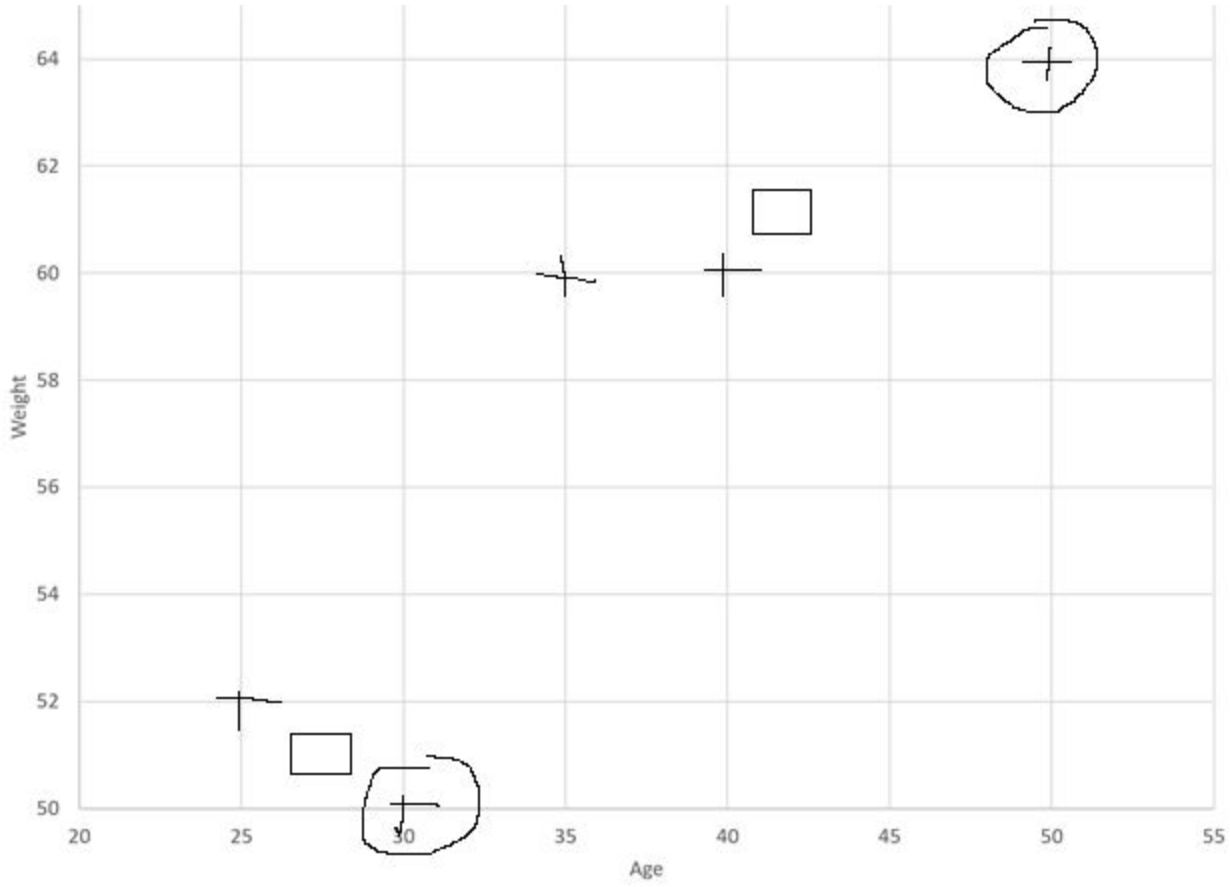


c) Recompute the centers after the first iteration of the k-means algorithm.

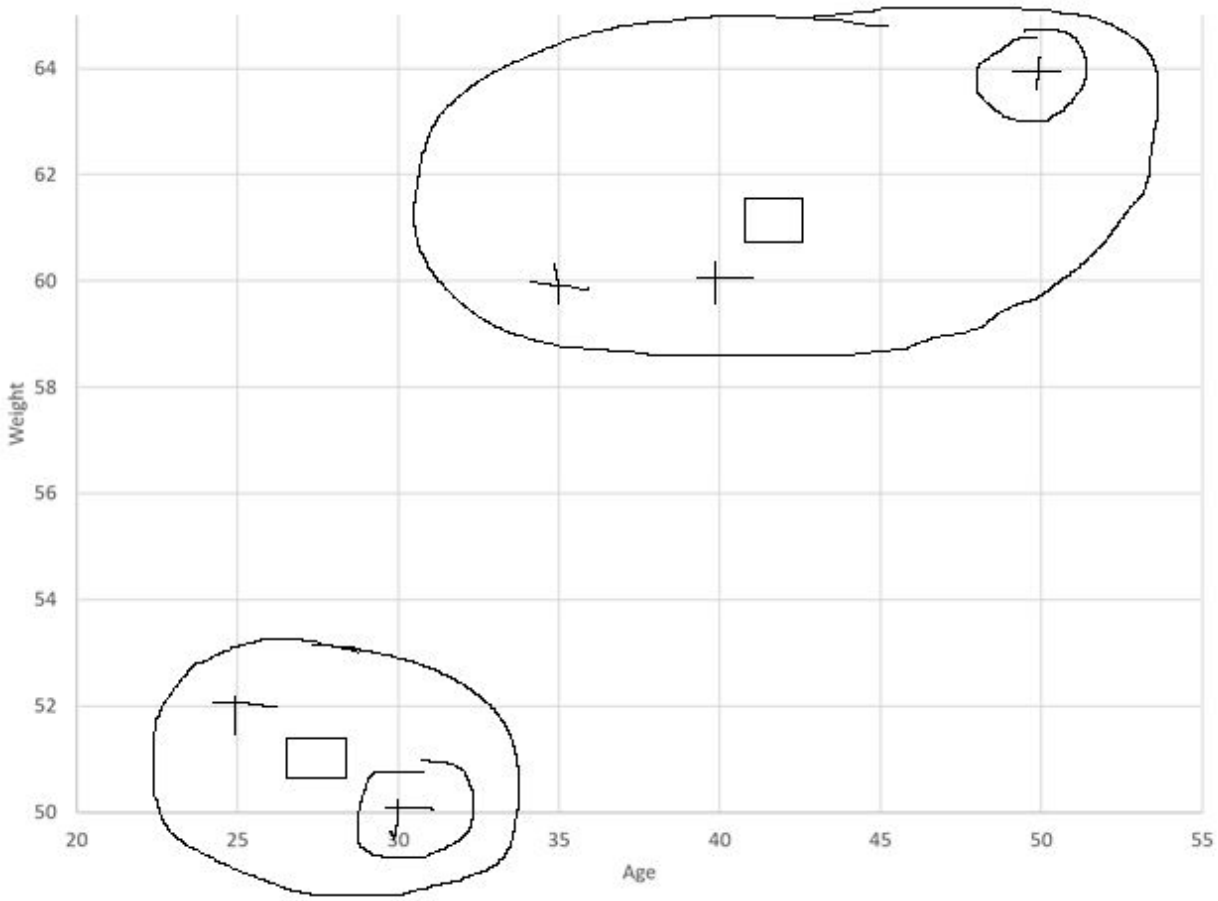
New center 1: (27.5, 51)

New center 2: (41.6, 61.3)

Indicate the two new centers on the graph (mark new centers with squares).



d) What are the two clusters after this first iteration? Draw two ovals, each containing all the points in one cluster in the graph above.



e) Will the k-means algorithm terminate after this first iteration or will it continue? Answer in 1-2 sentences.

1. The algorithm can continue on forever if it is allowed to because it's trying to optimize the center position of the cluster. It can be terminated by using a stopping condition.

f) If a new point (Age=32, Weight=55) is given, to which cluster will it belong?

1. The new point would belong to the bottom cluster.

2) Evaluate the performance of the 1-NN algorithm (with Euclidean distance) on the following data set. The goal is to predict credit risk from credit score and weekly income. **Use the first 3 rows for training and last 3 rows for testing**.

Credit score	Weekly income ($)	Risk
600	500	High
650	600	Low
800	550	Low
550	550	Low
660	500	High
750	580	Low

Testset[1]

Testset[1] and Trainset[1]
sqrt((600-550)^2 + (500-550)^2) = sqrt(2500 + 2500) = 70.71
Testset[1] and Trainset[2]
sqrt((550-650)^2 + (600-550)^2) = sqrt(10000 + 2500) = 111.80
Testset[1] and Trainset[3]
sqrt((800-550)^2 + (550-550)^2) = sqrt(22500 + 0) = 150

Predicted: High
Actual: Low

Testset[2]

Testset[2] and Trainset[1]
sqrt((660-600)^2 + (500-500)^2) = sqrt(3600 + 0) = 60
Testset[2] and Trainset[2]
sqrt((660-650)^2 + (600-500)^2) = sqrt(100 + 10000) = 100.49
Testset[2] and Trainset[3]
sqrt((800-660)^2 + (550-500)^2) = sqrt(19600 + 2500) = 148.66

Predicted: High
Actual: High

Testset[3]

Testset[3] and Trainset[1]
sqrt((750-600)^2 + (500-580)^2) = sqrt(22500 + 6400) = 170
Testset[3] and Trainset[2]
sqrt((750-650)^2 + (600-580)^2) = sqrt(10000 + 400) = 102
Testset[3] and Trainset[3]
sqrt((750-800)^2 + (550-580)^2) = sqrt(2500 + 900) = 58.31

Predicted: Low
Actual: Low

	Predicted		
Risk	High	Low	Total
High	1	0	1
Low	1	1	2
Total	2	1	

TP = 1
FP = 1
TN = 1
FN = 0

Calculate the following metrics for the classifiers.

1. Accuracy: (1+1)/3 = ⅔ = .667
2. Error rate: (0+1)/3 = ⅓ = .330
3. Precision of identifying high risk: 1/(1+1) = .5
4. Recall of identifying high risk: 1/(1+0) = 1 = 1
5. F1-score of identifying high risk: 2*0.5*1 / 1+.5 = 1/1.5 = .667