K Nearest Neighbors

Mengta Chung, PhD

Department of Management Sciences

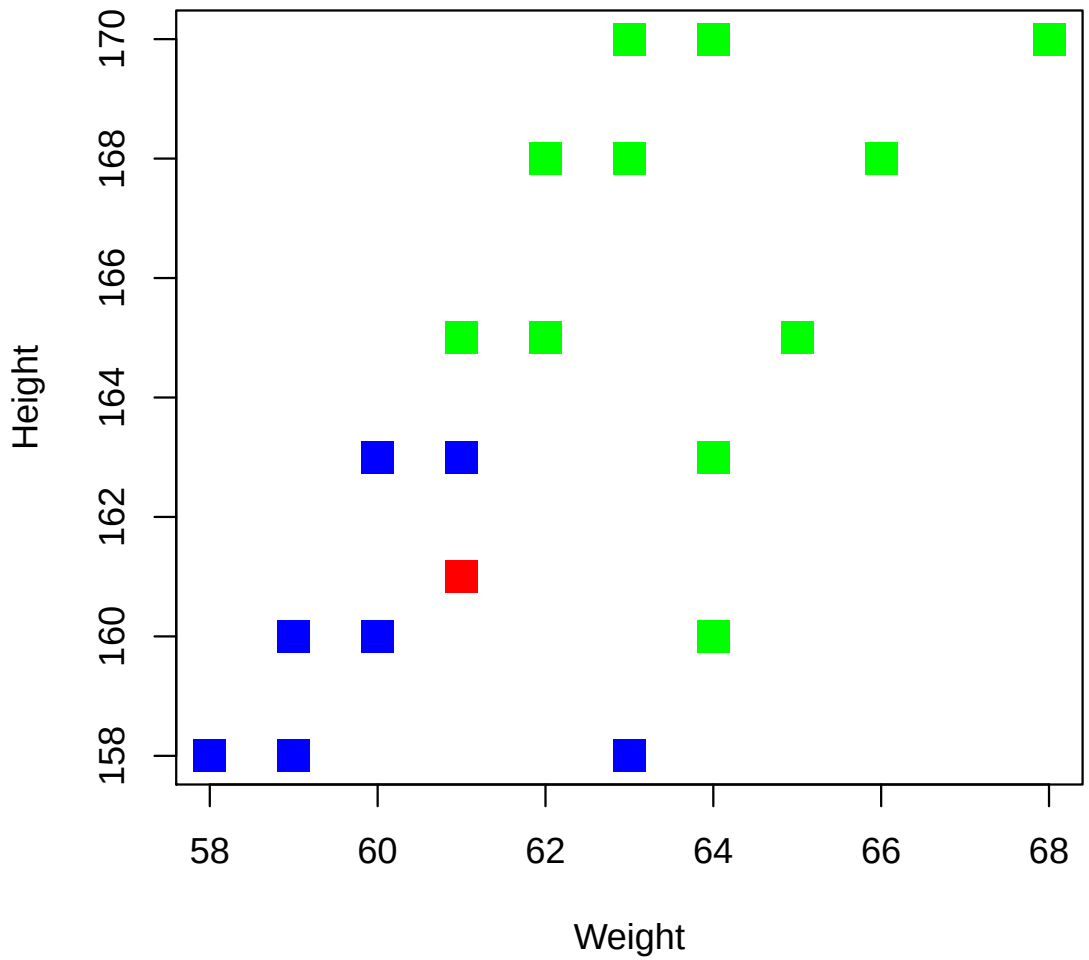
Tamkang University

Goal for Today

Monica is 161cm and 61kg. What is her size?

#	Height	Weight	Size
1	158	58	M
2	158	59	M
3	158	63	M
4	160	59	M
5	160	60	M
6	163	60	M
7	163	61	M
8	160	64	L
9	163	64	L
10	165	61	L
11	165	62	L
12	165	65	L
13	168	62	L
14	168	63	L
15	168	66	L
16	170	63	L
17	170	64	L
18	170	68	L

Plot



Introduction

- K Nearest Neighbors (KNN)
- KNN was developed by Evelyn Fix and Joseph Hodges Jr. in 1951 (C was developed in 1972).
- KNN can be used for classification (supervised learning) and regression (continuous dependent variable).

KNN

- KNN has been used as a non-parametric technique in statistics since the beginning of 1970's (Multivariate Data Analysis).
- KNN is an instance-based learning algorithm.

Algorithm

In classifying a point,

- calculate distances between the point and others
- find K nearest neighbors
- assign the class label to the point by majority vote

Results

#	Height	Weight	Size	Distance	Rank
1	158	58	M	4.243	
2	158	59	M	3.606	
3	158	63	M	3.606	
4	160	59	**M**	2.236	3
5	160	60	**M**	1.414	1
6	163	60	**M**	2.236	3
7	163	61	**M**	2	2
8	160	64	**L**	3.162	5
9	163	64	L	3.606	
10	165	61	L	4	
11	165	62	L	4.123	
12	165	65	L	5.657	
13	168	62	L	7.071	
14	168	63	L	7.28	
15	168	66	L	8.602	
16	170	63	L	9.22	
17	170	64	L	9.487	
18	170	68	L	11.402	

3-fold Cross-validation for finding Optimal K

K	Training	Test	Accuracy	Mean Accuracy
	d1 d2	d3	a_3	
$K=1$	d3 d1	d2	a_2	$a_{k=1} = \frac{a_1+a_2+a_3}{3}$
	d2 d3	d1	a_1	
	d1 d2	d3	a_3	
$K=2$	d3 d1	d2	a_2	$a_{k=2} = \frac{a_1+a_2+a_3}{3}$
	d2 d3	d1	a_1	
	d1 d2	d3	a_3	
$K=3$	d3 d1	d2	a_2	$a_{k=3} = \frac{a_1+a_2+a_3}{3}$
	d2 d3	d1	a_1	

Lab

Analyze knn_c.csv in Python