Tutorial exercises - Outlier Detection

Exercise 1. Z-score, Box-plot and Scatter-plot

The doctor of a school has measured the height of pupils in a 5th grade class. The result (in cm) is as follows:

130	132	138	136	131	153	131	133	129	133	110	132	129	134	135

132	135	134	133	132	130	131	134	135	135	134	136	133	133	130

a- Which ones are outliers and why?

b- The weight of those pupils was measured in kg and the results is as follows. Draw the box-plot for weight.

37	40	39	40.5	42	51	41.5	39	41	30	40	42	40.5	39.5	41

40.5	37	39.5	40	41	38.5	39.5	40	41	39	40.5	40	38.5	39.5	41.5

c- Draw the scatter-plot for both variables height and weight.

Exercise 2. k-Nearest neighbor approach

The data from the previous exercise is organized in a table as follows. Use the k-nearest neighbor to rank the pupils by most outlier to least outlier and give the top 4 outliers. Use k=3 and the Euclidian distance.

Pupil	Height	Weight
S1	130	37
S2	132	40
S3	138	39
S4	136	40.5
S5	131	42
S6	153	51
S7	131	41.5
S8	133	39
S9	129	41
S10	133	30
S11	110	40
S12	132	42
S13	129	40.5
S14	134	39.5
S15	135	41

Pupil	Height	Weight
S16	132	40.5
S17	135	37
S18	134	39.5
S19	133	40
S20	132	41
S21	130	38.5
S22	131	39.5
S23	134	40
S24	135	41
S25	135	39
S26	134	40.5
S27	136	40
S28	133	38.5
S29	133	39.5
S30	130	41.5

Exercise 3. Density-based outliers (to be done at your own time, not in class)

Use the previous dataset and calculate the LOF for each pupil data point and give the top 4 outliers. Use k=3. Use the same distance matrix you calculated in the previous exercise.

Exercise 4: Resolution-based outliers (to be done at your own time, not in class)

Use the previous dataset and calculate the ROF for each pupil data point and give the top 4 outliers. Use the same distance matrix you calculated in the previous exercise.