Data Normalization

Attribute Normalization

- In the context of machine learning, it is termed as feature normalization
- An attribute is normalised by scaling its value so that they fall within a small specified range (for example 0.0 to 1.0)
- Normalization is particularly useful for classification algorithms involving distance measurements and clustering
- For distance based approaches, normalization helps prevent attributes with large ranges from overweighting attributes with smaller ranges

Data Standardization (z-score Normalization)

- The process of rescaling one or more attributes so that the transformed data have 0 mean and unit variance i.e. standard deviation of 1
- Standardization assumes that data has a Guassian distribution
  - This assumption does not strictly have to be true, but this technique is more effective if your attribute distribution is Gaussian
- In this process, values of an attribute, A, are normalised based on the mean and standard deviation of A
- A value, x, of attribute A is normalised to $\hat{x}$ by computing

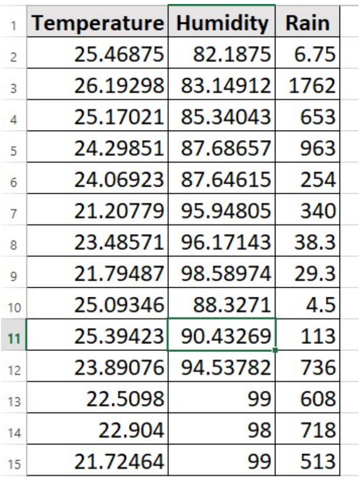
$$\widehat{x} = \frac{x - \mu_{\mathrm{A}}}{\sigma_{\mathrm{A}}}$$

- μ_{A}: mean of attribute A
- σ_{A}: standard deviation of attribute A

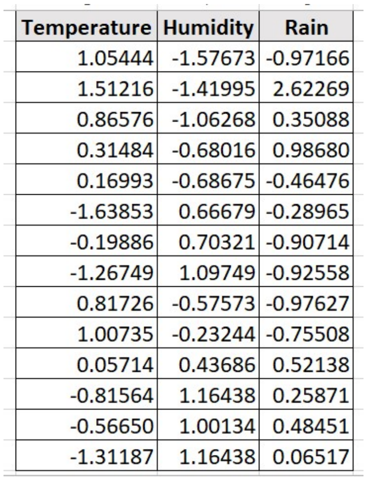
Data Standardization (z-score Normalization)

- This method of normalization is useful
  - when the actual minimum and maximum of attribute A are unknown
  - when there are outliers that dominates the Min-Max normalization
  - when data has Gaussian distribution (symmetric distribution)
- This method of normalization is useful when the ML algorithms make any assumptions of Gaussian distribution

Illustration of Data Standardization (z-score Normalization)



1	Temperature	Humidity	Rain
2	25.46875	82.1875	6.75
3	26.19298	83.14912	1762
4	25.17021	85.34043	653
5	24.29851	87.68657	963
6	24.06923	87.64615	254
7	21.20779	95.94805	340
8	23.48571	96.17143	38.3
9	21.79487	98.58974	29.3
10	25.09346	88.3271	4.5
11	25.39423	90.43269	113
12	23.89076	94.53782	736
13	22.5098	99	608
14	22.904	98	718
15	21.72464	99	513



Temperature	Humidity	Rain
1.05444	-1.57673	-0.97166
1.51216	-1.41995	2.62269
0.86576	-1.06268	0.35088
0.31484	-0.68016	0.98680
0.16993	-0.68675	-0.46476
-1.63853	0.66679	-0.28965
-0.19886	0.70321	-0.90714
-1.26749	1.09749	-0.92558
0.81726	-0.57573	-0.97627
1.00735	-0.23244	-0.75508
0.05714	0.43686	0.52138
-0.81564	1.16438	0.25871
-0.56650	1.00134	0.48451
-1.31187	1.16438	0.06517

	Temperature	Humidity	Rain	Temperature	Humidity	Rain
μ:	23.80035	91.86	481	0.000	0.000	0.000
σ:	1.58225	6.13	488	1	1	1