Data Transformation

Normalisation

Data Normalisation

- Sometimes the attributes selected are *raw attributes.*
  - They have a meaning in the original domain from where they were obtained
  - They are designed to work with the operational system in which they are being currently used

- Usually these original attributes are not good enough to obtain accurate predictive models

Data Normalisation

- It is common to perform a series of manipulation steps to transform the original attributes or to generate new attributes
- Data are scaled to fall within a small, specified range
  - They will show better properties that will help the predictive power of the model

- The new attributes are usually named *modeling variables or analytic variables.*

Data Transformation: Normalisation

- min-max Normalisation

$$v' = \frac{v - min_A}{max_A - min_A}(new - max_A - new - min_A) + new - min_A$$

- z-score Normalisation

$$v' = \frac{v - \overline{A}}{\sigma_A}.$$

- Normalisation by decimal scaling

$$v' = \frac{v}{10^j},$$

Where j is the smallest integer such that Max(| v' |)<1

Data Normalisation

Min-Max Normalisation

- The min-max Normalisation aims to scale all the numerical values *v of a numerical* attribute *A to a specified range denoted by [new − minA, new − maxA].*
- The following expression transforms *v to* the new value *v':*

$$v' = \frac{v - min_A}{max_A - min_A}(new - max_A - new - min_A) + new - min_A$$

Data Normalisation

Z-score Normalisation

- If minimum or maximum values of attribute *A* are not known, or the data is noisy, the *min-max* Normalisation is infeasible
- Alternative: normalize the data of attribute A to obtain a new distribution with mean 0 and std. deviation equal to 1

$$v' = \frac{v - \overline{A}}{\sigma_A}.$$

Data Normalisation

Decimal-scaling Normalisation

- A simple way to reduce the absolute values of a numerical attribute

$$v' = \frac{v}{10^j},$$

- where *j* is the smallest integer such that *new* – $max_A < 1$.