Input Encoding

Continuous Attributes

- Attribute values can be encoded in a standardized manner, taking values between 0 and 1, even for categorical variables.
  - Also called “feature scaling”

For continuous variables, we can apply one of the following:

$$X^* = [X - \min(X)]/[\max(X)-\min(X)]$$

$$X^* = [X - \text{mean}(X)]/[\max(X)-\min(X)]$$

$$X^* = [X - \text{mean}(X)]/\text{stdev}$$

Categorical Attributes – One-hot-encoding

- Use *indicator (flag) variables.*
  - E.g. *marital status attribute*, containing values *single, married, divorced*.
  - Records for *single* would have

    1 for single, and 0 for the rest, i.e. (1,0,0)
  - Records for *married* would have

    1 for *married*, and 0 for the rest, i.e. (0,1,0)
  - Records for *divorced* would have

    1 for *divorced*, and 0 for the rest, i.e. (0,0,1)
  - Records for *unknown* would have

    0 for all, i.e. (0,0,0)
- In general, categorical attributes with *k* values can be translated into *k* - 1 indicator attributes.