Train In Data

KNN Imputation

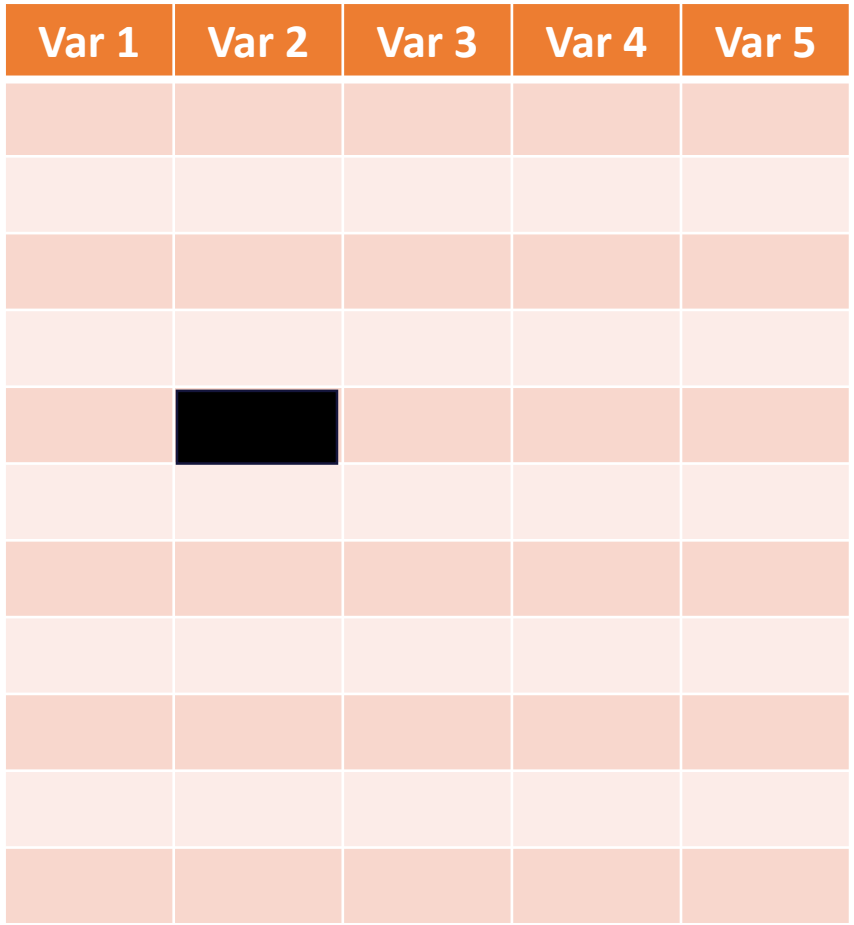
KNN Imputation

Determines the missing data point value, as the weighted average of the values of its K nearest neighbours.

The logic

If an observation looks very similar to other observations in the data set, most likely, the missing value would be similar to the value shown in those similar observations.

KNN Imputation: procedure

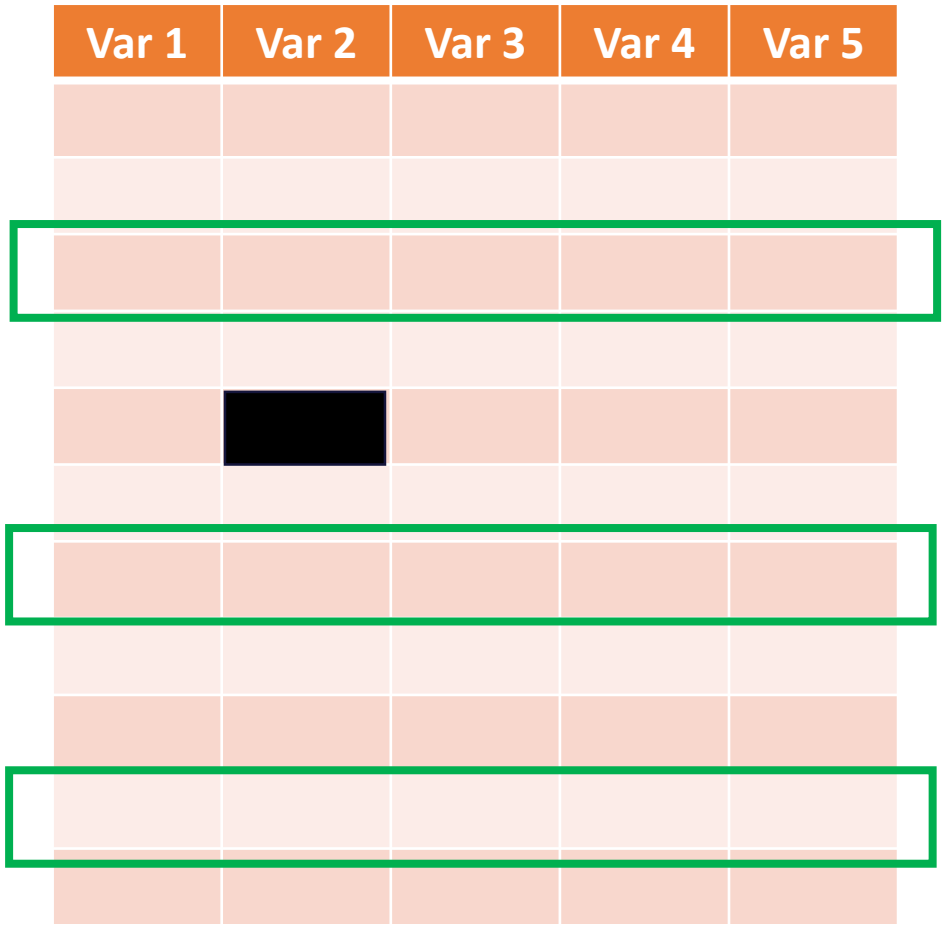


For each variable with missing value:

1- train a KNN using the other variables

2- Find the K closest neighbours

KNN Imputation: procedure

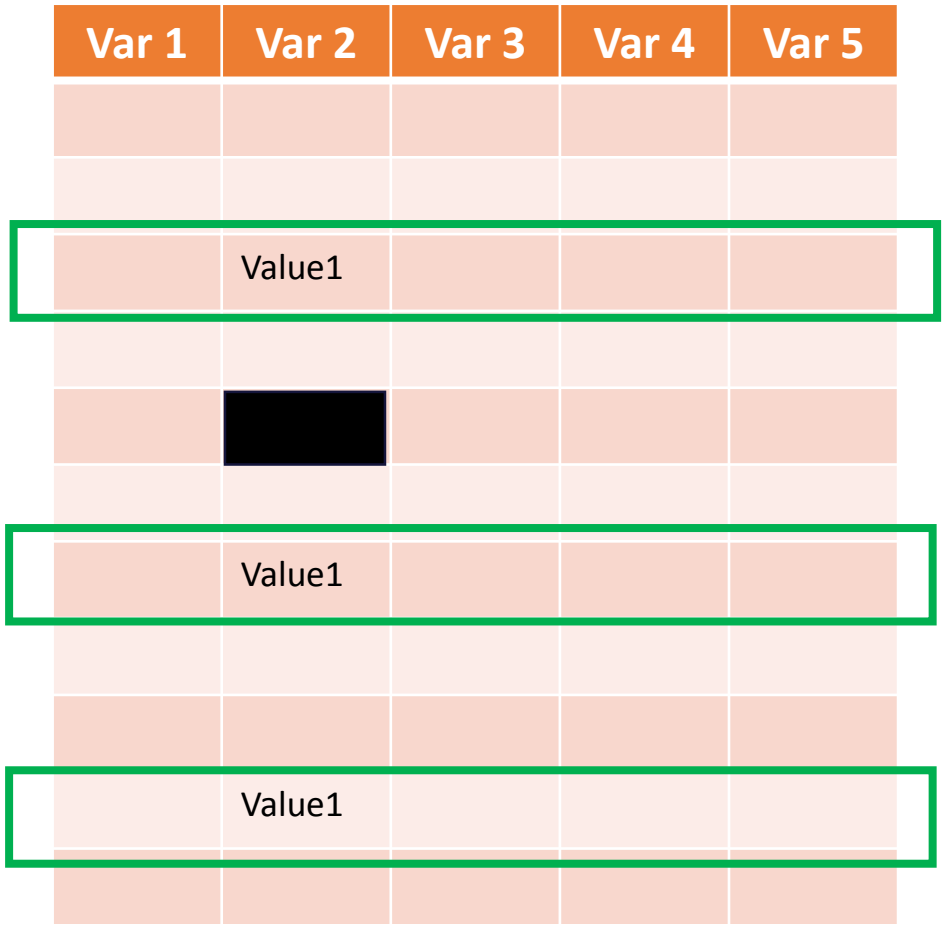


For each variable with missing value:

1- train a KNN using the other variables

2- Find the K closest neighbours

KNN Imputation: procedure



For each variable with missing value:

3 – Determine the weighted average of the K neighbours

$$NA\ replacement = \frac{w1\ \times Value1 + w2\ \times Value2 + w3\ \times Value3}{k}$$

KNN imputation: weight

Uniform: all neighbours count equally ($w_i = 1$)

Distance: $$Wi = \frac{1}{Euclidean\ distance(neighbour - observation)}$$

KNN imputation: challenge

Find the best number of neighbours ➔ K

If you have time in your hands, this turns into a regression problem, where we need a training set to train the KNN and determine the optimal K

KNN general guidelines

The authors from the algorithm claim:

- A small percentage of missing data makes the imputation more precise (up to 20% missing data)
- The method is relative insensitive to the value of K, for K between 10 and 20

These parameters were developed to fill in missing values in gene sequencing data and may not extend to other problems.

Train In Data

THANK YOU

www.trainindata.com