An automatic procedure for selecting weights in kNN imputation
No solutions, just some ideas . . .
kNN

```r
kNN(data, variable = colnames(data), metric = NULL, k = 5, dist_var = colnames(data), weights = NULL, numFun = median, catFun = maxCat, makeNA = NULL, NAcond = NULL, impNA = TRUE, donorcond = NULL, mixed = vector(), mixed.constant = NULL, trace = FALSE, imp_var = TRUE, imp_suffix = "imp", addRandom = FALSE, useImputedDist = TRUE, weightDist = FALSE)
```

- `dist_var` - variables to be used for distance computation
- `weights` - weights for the distance computation
- `numFun, catFun` - function for aggregating (`sampleCat, maxCat`).
- `addRandom` - add a random value to the distance computation (with very low weight) to break ties
- `weightDist` - use the distances of the knns in the aggregation function
Distance computation

- Extension of the Gower distance

\[ d_{i,j} = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k}, \]  

where

- \( w_k \) is the weight for the contribution of variable \( k \) to the total distance.
- \( w_k \) should represent the “importance” or the “dependency” of variable \( k \) to the target variable.
- If the number of target variables and/or (possible) distance variables is large \( \Rightarrow \) automatic choice of weights.
Method 1 - Random forest

- Random forest is applied to the data set (or a subsample) and the importance is used as weight variable.
- $V_1$: Directly apply RF to the data sets
- $V_2$: Create binomial dummy variables for categorical variables and apply RF and then kNN to it
Method 2 - Lasso shrinkage

- Lasso regression excludes variables, which are presumably not important for estimating the model.
- Can be applied to a subsample
- $V_1$: within 1 standard deviation of minimal cross-validated error
- $V_2$: minimum cross-validated error
Method 3 - Single feature accuracy

- Apply kNN to the observed data points (or a subsample) with only one distance variable at a time.
- Use the inverse of the error (classification error or mean absolute error depending on the type of target variable) as weight for the kNN.
(small) Simulation study

- Simulated data set from a normal distribution with varying correlation
- 2 numerical variables
- 2 mixed variables
- 2 categorical variables
- and a varying number of uncorrelated random numerical variables
Results - numeric target variable

![Graph showing numeric - number of random variables 0]
Results - numeric target variable

Numeric – Number of Random Variables 5

MAE

Correlation

variable

knn
knnw
knnw2
knnl.l1se
knnl.lmin
knnw3
Results - numeric target variable

Numeric - Number of Random Variables 10

- Variable: knn, knnw, knnw2, knnl.l1se, knnl.lmin, knnw3
- Correlation vs. MAE

Kowarik, Meraner (www.statistik.at)
Results - numeric target variable

Numeric – Number of Random Variables 15

Correlation

MAE

variable

knn
knnw
knnw2
knnl.l1se
knnl.lmin
knnw3

Numeric − Number of Random Variables 15

0.4
0.6
0.8
1.0
0.25 0.50 0.75

Kowarik, Meraner (www.statistik.at)
Results - numeric target variable

Numeric – Number of Random Variables 20

MAE

Correlation

variable
- knn
- knnw
- knnw2
- knnl.l1se
- knnl.lmin
- knnw3

Kowarik, Meraner (www.statistik.at)
Results - categorical target variable

Categorical – Number of Random Variables 0

<table>
<thead>
<tr>
<th>MER</th>
<th>knn</th>
<th>knnw</th>
<th>knnw2</th>
<th>knnl.l1se</th>
<th>knnl.lmin</th>
<th>knnw3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correlation

Kowarik, Meraner (www.statistik.at)
Results - categorical target variable

Categorical – Number of Random Variables 5

Correlation

MER

variable
knn
knnw
knnw2
knnl.l1se
knnl.lmin
knnw3

Kowarik, Meraner (www.statistik.at)
Results - categorical target variable

Categorical – Number of Random Variables 10

- knn
- knnw
- knnw2
- knnl.l1se
- knnl.lmin
- knnw3

Correlation

MER

Kowarik, Meraner (www.statistik.at)
Results - categorical target variable

Categorical – Number of Random Variables 15

<table>
<thead>
<tr>
<th>Variable</th>
<th>knn</th>
<th>knnw</th>
<th>knnw2</th>
<th>knnl.l1se</th>
<th>knnl.lmin</th>
<th>knnw3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MER</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Correlation
Results - categorical target variable

![Graph](attachment:graph.png)

- MER
- Correlation
- Categorical - Number of Random Variables 20