

Condensed Nearest Neighbor (CNN)

```
def cnn(data, y, samp_method = "balance", drop_na_col = True, drop_na_row = True,
n_seed = 1, rel_thres = 0.5, rel_method = "auto", rel_xtrm_type = "both", rel_coef = 1.5,
rel_ctrl_pts_rg = None, k=1, n_jobs=1)
```

Function designed to help solve the problem of imbalanced data for regression, CNN under-samples the majority class.

Parameters:

main arguments / inputs:

data: pandas dataframe, the training set.

y: string, response variable y by name. It should be a column name found in the dataframe data.

samp_method: {'balance', 'extreme'}, default = 'balance', specified method to determine over / under sampling percentage.

drop_na_col: bool, default = 'True', if 'True', auto drop columns with nan's.

drop_na_row: bool, default = 'True', if 'True', auto drop rows with nan's.

n_seed: int, default = '1', represent the number of seed samples moved from the majority set to STORE at the beginning.

phi relevance function arguments / inputs:

rel_thres: positive real, default = 0.5, define the relevance threshold considered rare in phi relevance function.

rel_method: {'auto', 'manual'}, default = 'auto', the relevance method in phi relevance function.

rel_xtrm_type: {'low', 'high', 'both'}, default = 'both', distribution focus on high, low or both.

rel_coef: positive real, default = 1.5, coefficient for box plot in phi relevance function to consider rare.

rel_ctrl_pts_rg: 2d array, default = None, when rel_method = 'manual', it inputs for "manual" rel method.

KNeighborsClassifier attribute:

k: positive integer, default = 1, number of neighbors for k-NN.

n_jobs: positive integer, default = 1, the number of parallel jobs to run for neighbors' search.

k_neighbors_classifier: KNeighborsClassifier, default = None, allow users to customize the attributes of the classifier.

References

Branco, P., Torgo, L., Ribeiro, R. (2017). SMOGN: A Pre-Processing Approach for Imbalanced Regression. Proceedings of Machine Learning Research, 74:36-50.

<http://proceedings.mlr.press/v74/branco17a/branco17a.pdf>.

Hart, P. (1968). The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory, 14(3), 515-516. <https://ieeexplore.ieee.org/document/1054155>.

Kunz, N., (2019). SMOGN. <https://github.com/nickkunz/smogn>