Introduction of Gaussian Noise (GN)

def gn(data, y, pert = 0.02,samp_method = "balance",under_samp = True, drop_na_col = True, drop_na_row = True, replace = False, manual_perc = False, perc_u = -1, perc_o = -1, rel_thres = 0.5, rel_method ="auto", rel_xtrm_type = "both", rel_coef = 1.5, rel_ctrl_pts_rg = None)

GN is designed to help solve the problem of imbalanced data for regression; This function (gn) applies over-sampling of the minority class.

## 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.

replace: bool, default = 'True', if 'True', do the sampling replacement.

Manual_perc: bool,default = 'False', when false user are not allowed to defines percentage of under-sampling and over-sampling.

Perc_o: negative int, default = '-1', represent the percentage of over-sampling.

## 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.

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.

Branco, P., Torgo, L., & Ribeiro, R. P. (2019). Pre-processing approaches for imbalanced distributions in regression. Neurocomputing, 343, 76-99. https://www.sciencedirect.com/science/article/abs/pii/S0925231219301638.

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