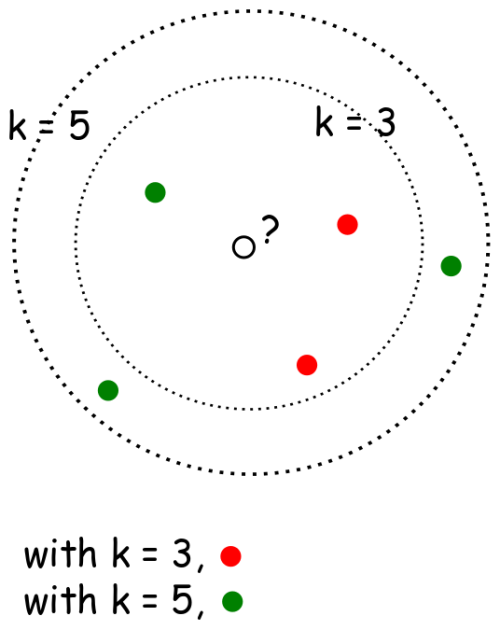
User Documentation

K-nearest neighbor’s (KNN) classifier

KNN classifier is a non-parametric supervised learning method for classification problems. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).



**Schema**: how KNN classifier works

The KNN classifier use case is as follows:

* The dependent variable (or target) of a dataset is categorical, i.e. classification problem. There are over 75% of the dataset being labeled, and the rest (less than 25% of the dataset) are missing labels. KNN classifier can be used to predict the missing labels for the dataset.
* The attributes/features should be numeric values instead of categorical values. No missing values in the attributes/features. In case of missing values in attributes/features, using an imputation method to impute missing values first.



**Schema**: KNN classifier to predict the missing labels.

## Algorithm details

KNN classifier is a classification algorithm, which classifies unknown instances by relating the unknown to the known according to some distance/similarity function. KNN is an non parametric lazy learning algorithm, which means that it does not make any assumptions on the underlying data distribution. The KNN algorithm works by finding the k nearest neighbor and do a majority voting.

# **Input Description**

The label field is “Target”. Missing label is “null”.

{

"data":{

"SepalLn":[5.1,4.9,4.7,6.5,6.2,5.9],

"SepalWd":[3.5,3,3.2,3,3.4,3],

"PetalLn":[1.4,1.4,1.3,5.2,5.4,5.1],

"PetalWd":[0.2,0.2,0.2,2,2.3,1.8],

"Target":["Iris-setosa","Iris-setosa","Iris-setosa","Iris-virginica","Iris-virginica",null]

}

}

**data**

The main part of the input contains the features and labels. The features here are: “SepalLn”, “SepalWd”, “PetalLn” and “PetalWd”. The label is “Target”. **Note, the label is always expected to be named “Target” in the JSON input file.** The missing label is always ***null***.

# **Outputs Description**

The output of the KNN classifier give the original dataset in addition with another column as “Predict”:

{

"data":{

"SepalLn":[5.1,4.9,4.7,6.5,6.2,5.9],

"SepalWd":[3.5,3,3.2,3,3.4,3],

"PetalLn":[1.4,1.4,1.3,5.2,5.4,5.1],

"PetalWd":[0.2,0.2,0.2,2,2.3,1.8],

"Target":["Iris-setosa","Iris-setosa","Iris-setosa","Iris-virginica","Iris-virginica",null]

"Predict":["Iris-setosa","Iris-setosa","Iris-setosa","Iris-virginica","Iris-virginica","Iris-virginica"]

},

“KofModel”: 3,

“testscore”: 0.985

}

“KofModel” is the best choice of K decided by cross-validation.

“testscore” is the expected accuracy of the KNN classifier on the given dataset.

# **Usage samples**

Refer to input and output descriptions.

©2016 General Electric Company – All rights reserved.

GE, the GE Monogram and Predix are trademarks of General Electric Company.

No part of this document may be distributed, reproduced or posted without the express written permission of General Electric Company.

THIS DOCUMENT AND ITS CONTENTS ARE PROVIDED “AS IS,” WITH NO REPRESENTATION OR WARRANTIES OF ANY KIND, WHETHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO WARRANTIES OF DESIGN,

MERCHANTABILITY, OR FITNESS FOR A PARTICULAR PURPOSE. ALL OTHER LIABILITY ARISING FROM RELIANCE UPON ANY INFORMATION CONTAINED HEREIN IS EXPRESSLY DISCLAIMED.