×

-

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

(http://play.google.com/store/apps/details?id=com.analyticsvidhya.android)

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

(https://www.facebook.com/AnalyticsVidhya) (https://twitter.com/analyticsvidhya)

(https://plus.google.com/+Analyticsvidhya/posts)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

(https://in.linkedin.com/company/analytics-vidhya)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

LOGIN / REGISTER (HTTPS://ID.ANALYTICSVIDHYA.COM/ACCOUNTS/LOGIN/?NEXT=HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2018/03/INTRODUCTION-K-NEIGHBOURS-ALGORITHM-CLUSTERING/)

HOME (HTTPS://WWW.ANALYTICSVIDHYA.COM) BLOG ARCHIVE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG-ARCHIVE/)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

DISCUSS (HTTPS://DISCUSS.ANALYTICSVIDHYA.COM) CORPORATE (HTTPS://WWW.ANALYTICSVIDHYA.COM/CORPORATE/)

Analytics Vidhya
Learn everything about analytics

(https://www.analyticsvidhya.com/myfeed/?utm-source=blog&utm-medium=top-icon/)

Home (https://www.analyticsvidhya.com/) › Algorithm (https://www.analyticsvidhya.com/blog/category/algorithm/) › Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R) (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/)

ALGORITHM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/ALGORITHM/)

BIG DATA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BIG-DATA/)

BUSINESS ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/)

CLASSIFICATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/CLASSIFICATION/)

INTERMEDIATE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/INTERMEDIATE/)

MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/)

PYTHON (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/PYTHON-2/)
BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)
R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/R/)
STRUCTURED DATA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/STRUCTURED-DATA/)
COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)
SUPERVISED (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/SUPERVISED/)
HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)
DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)
BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)
CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R)

TAVISH SRIVASTAVA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/TAVISH1/), M...

Note: This article was originally published on Oct 10, 2014 and updated on Mar 27th, 2018

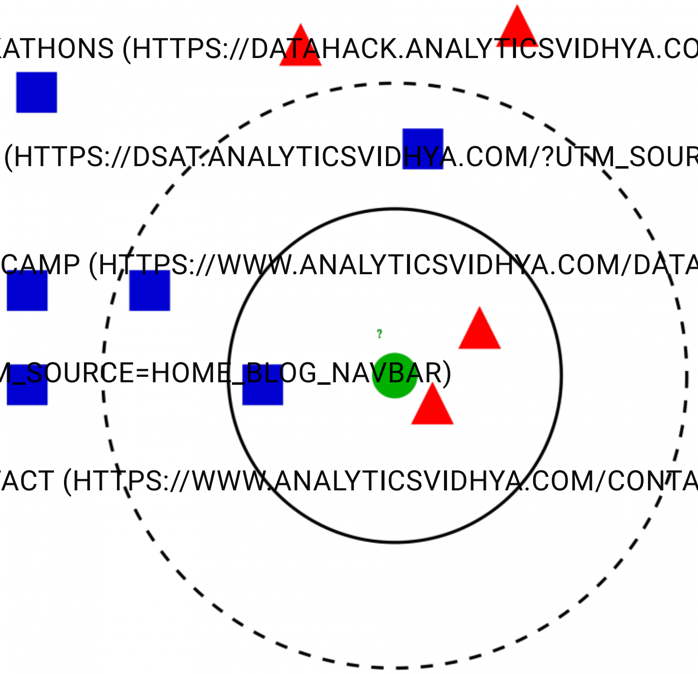
Overview

- Understand k nearest neighbor (KNN) – one of the most popular machine learning (https://www.analyticsvidhya.com/machine-learning/?utm_source=blog&utm_medium=k-nearest-neighbors) algorithms
- Learn the working of kNN in python
- Choose the right value of *k* in simple terms

Introduction

In the four years of my data science career (https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2/?utm_source=blog&utm_medium=introknearestneighborarticle), I have built more than 80% classification models and just 15-20% regression models. These ratios can be more or less generalized throughout the industry. The reason behind this bias towards classification models (https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2/?utm_source=blog&utm_medium=introknearestneighborarticle) is that most analytical problems involve making a decision.

For instance, will a customer attrite or not, should we target customer X for digital campaigns, whether customer has a high potential or not etc. These analysis are more insightful and directly linked to an implementation roadmap.



(https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/03/knn3.png)

In this article, we will talk about another widely used machine learning (https://www.analyticsvidhya.com/machine-learning/?utm_source=blog&utm_medium=k-nearest-neighbors) classification techniqu (https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2/?utm_source=blog&utm_medium=introknearestneighborarticle)e called K-nearest neighbors (KNN) . Our focus will be primarily on how does the algorithm work and how does the input parameter affect the output/prediction.

Table of Contents

- How do we choose the factor K?
- Breaking it Down – Pseudo Code of KNN
- Implementation in Python from scratch
- Comparing our model with scikit-learn

When do we use KNN algorithm?

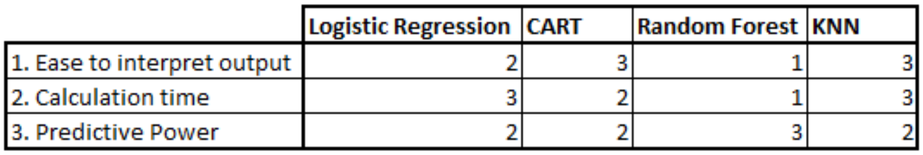
KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output

2. Calculation time

3. Predictive Power

Let us take a few examples to place KNN in the scale :

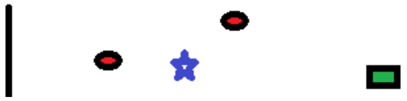


	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

(https://www.analyticsvidhya.com/wp-content/uploads/2014/10/Model-comparison.png)KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time.

How does the KNN algorithm work?

Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS) :



BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

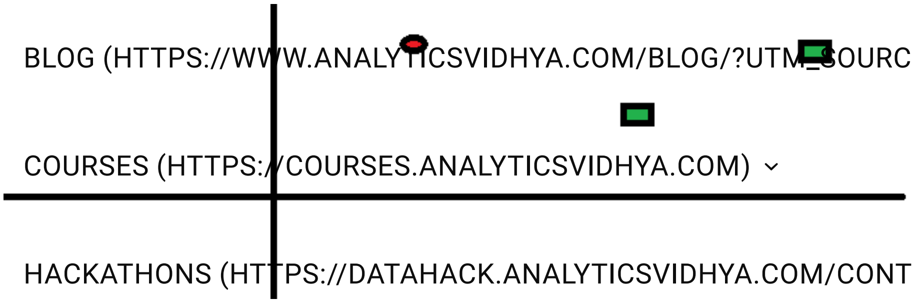
COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

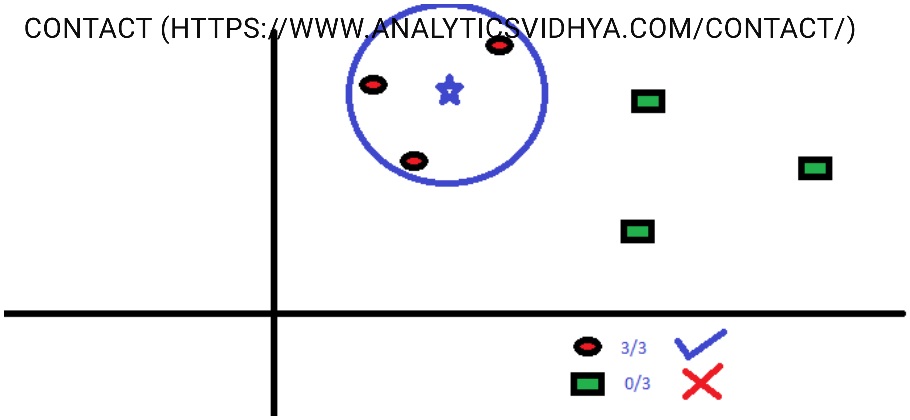
DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)



(https://www.analyticsvidhya.com/wp-content/uploads/2014/10/scenario1.png)You intend to find out the class of the blue star (BS) . BS can either be RC or GS and nothing else. The “K” is KNN algorithm is the nearest neighbors we wish to take vote from. Let’s say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three datapoints on the plane. Refer to following diagram for more details:



(https://www.analyticsvidhya.com/wp-content/uploads/2014/10/scenario2.png) The three closest points to BS is all RC. Hence, with good confidence level we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm. Next we will understand what are the factors to be considered to conclude the best K.

How do we choose the factor K?

First let us try to understand what exactly does K influence in the algorithm. If we see the last example, given that all the 6 training observation remain constant, with a given K value we can make boundaries of each class. These boundaries will segregate RC from GS. The same way, let's try to see the effect of value "K" on the class boundaries. Following are the different boundaries separating the two classes with different values of K.

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

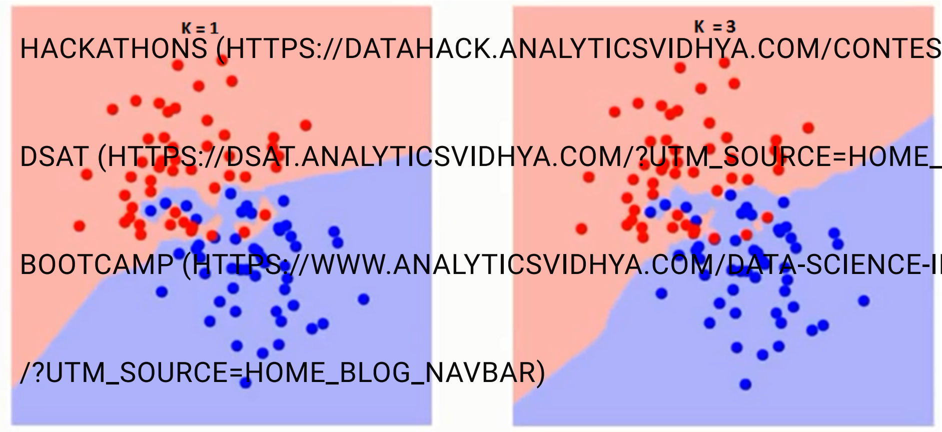
COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

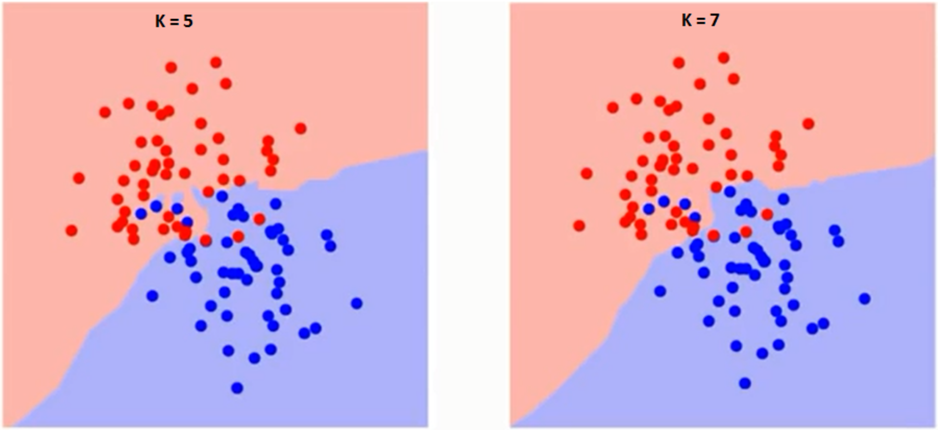
DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

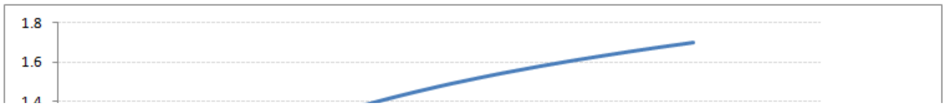


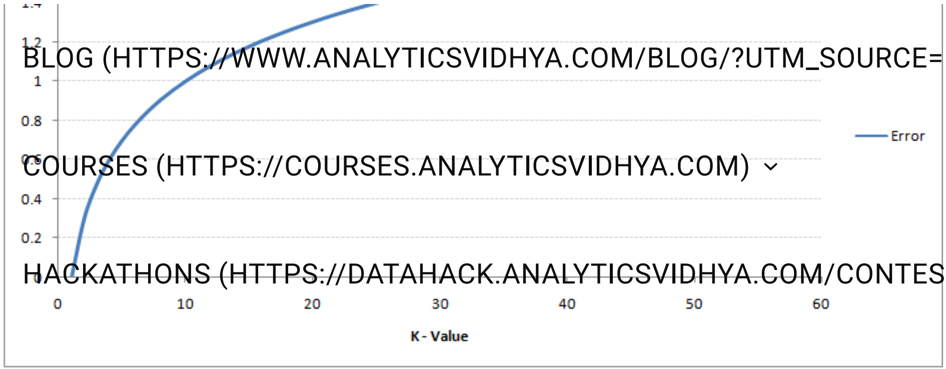
(https://www.analyticsvidhya.com/wp-content/uploads/2014/10/K-judgement.png)



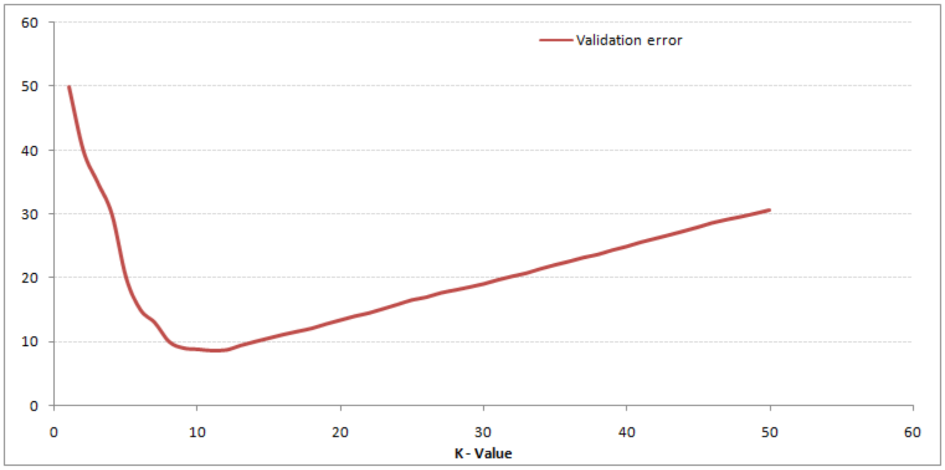
(https://www.analyticsvidhya.com/wp-content/uploads/2014/10/K-judgement2.png)

If you watch carefully, you can see that the boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority. The training error rate and the validation error rate are two parameters we need to access on different K-value. Following is the curve for the training error rate with varying value of K :





(https://www.analyticsvidhya.com/wp-content/uploads/2014/10/training-error.png)As you can see, the error rate at K=1 is always zero for the training sample. This is because the closest point to any training data point is itself.Hence the prediction is always accurate with K=1. If validation error curve would have been similar, our choice of K would have been 1. Following is the validation error curve with varying value of K:



(https://www.analyticsvidhya.com/wp-content/uploads/2014/10/training-error_11.png)This makes the story more clear. At K=1, we were overfitting the boundaries. Hence, error rate initially decreases and reaches a minima. After the minima point, it then increase with increasing K. To get the optimal value of K, you can segregate the training and validation from the initial dataset. Now plot the validation error curve to get the optimal value of K. This value of K should be used for all predictions.

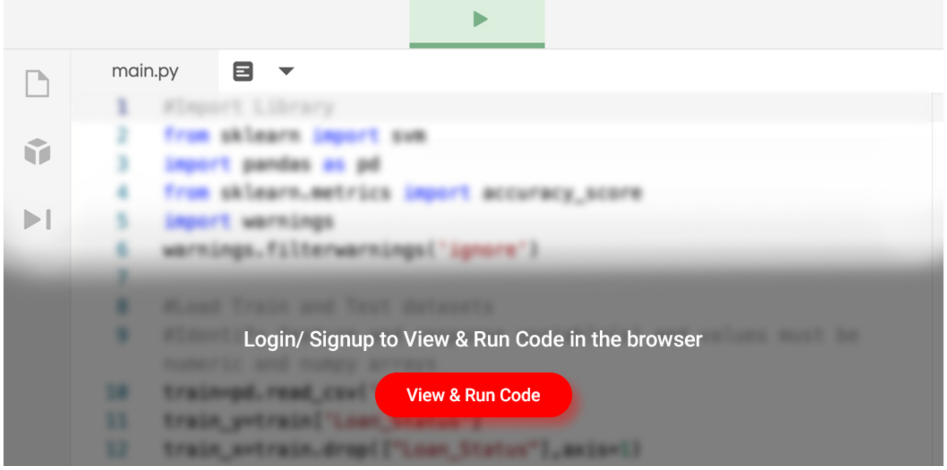
Breaking it Down – Pseudo Code of KNN

We can implement a KNN model by following the below steps:

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
    1. Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
    2. Sort the calculated distances in ascending order based on distance values
    3. Get top k rows from the sorted array
    4. Get the most frequent class of these rows
    5. Return the predicted class

Implementation in Python from scratch

We will be using the popular Iris dataset for building our KNN model. You can download it from here (https://gist.githubusercontent.com/gurchetan1000/ec90a0a8004927e57c24b20a6f8c8d35/raw/fcd83b35021a4c1d7f1f1d5dc83c07c8ffc0d3e2/iris.csv).



```
#Import Library
from sklearn import svm
import pandas as pd
from sklearn.metrics import accuracy_score
import warnings
warnings.filterwarnings('ignore')

#Load Train and Test datasets
```

(https://id.analyticsvidhya.com/accounts/login/?next=https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/?&utm_source=coding-window-blog&source=coding-window-blog)

Comparing our model with scikit-learn

```
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=3)
neigh.fit(data.iloc[:,0:4], data['Name'])


# Predicted class
print(neigh.predict(test))

-> ['Iris-virginica']

# 3 nearest neighbors
print(neigh.kneighbors(test)[1])
-> [[141 139 120]]
```

We can see that both the models predicted the same class ('Iris-virginica') and the same nearest neighbors ([141 139 120]). Hence we can conclude that our model runs as expected.

Implementation of kNN in R

Step 1: Importing the data

Step 2: Checking the data and calculating the data summary

```
data<-read.table(file.choose(), header = T, sep = ",", dec = ".")
head(data)  #Top observations present in the data
dim(data)   #Check the dimensions of the data
summary(data) #Summarise the data
```

view raw (https://gist.github.com/Harshit1694/bfc073fa1c57767c99a0b8e49fdf7ef0/raw/698f09e0156405053d62270dd0e989a6e16a6478/import_knn.R) import_knn.R (https://gist.github.com/Harshit1694/bfc073fa1c57767c99a0b8e49fdf7ef0#file-import_knn-r) hosted with ❤ by GitHub (https://github.com)

Output

```
#Top observations present in the data
SepalLength SepalWidth PetalLength PetalWidth Name
1 5.1 3.5 1.4 0.2 Iris-setosa
2 4.9 3.0 1.4 0.2 Iris-setosa
3 4.7 3.2 1.3 0.2 Iris-setosa
4 4.6 3.1 1.5 0.2 Iris-setosa
5 5.0 3.6 1.4 0.2 Iris-setosa
6 5.4 3.9 1.7 0.4 Iris-setosa

#Check the dimensions of the data
[1] 150 5

#Summarise the data
SepalLength SepalWidth PetalLength PetalWidth Name
Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 Iris-setosa :50
1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 Iris-versic
olor:50
Median :5.800 Median :3.000 Median :4.350 Median :1.300 Iris-virgin
ica :50
Mean :5.843 Mean :3.054 Mean :3.759 Mean :1.199
3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
```

Step 3: Splitting the Data

```
#Splitting the data set into train and test
set.seed(2)

part <- sample(2, nrow(data), replace = TRUE, prob = c(0.7, 0.3))
```

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

```

train<- data[part == 1,]

test <- data[part == 2,]
```

view raw (https://gist.github.com/Harshit1694/449771ddaa71af871bb85f14d9c5c743/raw/b21c3822d146600bd20ff111bcd149cfc2d40e0d/split_kNN.R) split_kNN.R (https://gist.github.com/Harshit1694/449771ddaa71af871bb85f14d9c5c743#file-split_knn-r) hosted with ❤ by GitHub (https://github.com)

Step 4: Calculating the Euclidean Distance

```
#Calculating the euclidean distance

ED<-function(data1,data2){
  distance=0
  for (i in (1:(length(data1)-1))){
    distance=distance+(data1[i]-data2[i])^2
  }
  return(sqrt(distance))
}
```

view raw (https://gist.github.com/Harshit1694/e7f45054499e015ff75034958ff7d40c/raw/d517bfa117cc6838303f0ee0df3de57fe9cdc53c/euc_kNN.R) euc_kNN.R (https://gist.github.com/Harshit1694/e7f45054499e015ff75034958ff7d40c#file-euc_knn-r) hosted with ❤ by GitHub (https://github.com)

Step 5: Writing the function to predict kNN

Step 6: Calculating the label(Name) for K=1

```
#Writing the function to predict kNN
knn_predict <- function(test, train, k_value){
  pred <- c()
  #LOOP-1
  for(i in c(1:nrow(test))){
    dist = c()
    char = c()
    setosa =0
    versicolor = 0
    virginica = 0
  }

    #LOOP-2-looping over train data
    for(j in c(1:nrow(train))){}

      dist <- c(dist, ED(test[i,], train[j,]))
      char <- c(char, as.character(train[j,][[5]]))


    df <- data.frame(char, dist$SepalLength)
    df <- df[order(df$dist.SepalLength),]       #sorting datafra
    df <- df[1:k_value,]


    #Loop 3: loops over df and counts classes of neibhors.
    for(k in c(1:nrow(df))){
      if(as.character(df[k,"char"]) == "setosa"){
        setosa = setosa + 1
      }else if(as.character(df[k,"char"]) == "versicolor"){
        versicolor = versicolor + 1
      }else
        virginica = virginica + 1
    }


    n<-table(df$char)
    pred=names(n)[which(n==max(n))]

  return(pred) #return prediction vector
}

#Predicting the value for K=1
K=1
predictions <- knn_predict(test, train, K)
```

view raw (https://gist.github.com/Harshit1694/e784618fd626ea3161523fbfdae47d19/raw/5f9395c6b227d1bf1950d7159e8afe1a94ec25e7/kNN_func.R) kNN_func.R (https://gist.github.com/Harshit1694/e784618fd626ea3161523fbfdae47d19#file-knn_func-r) hosted with ❤ by GitHub

(https://github.com)

Output

```
For K=1

[1] "Iris-virginica"
```

In the same way, you can compute for other values of K.

Comparing our kNN predictor function with “Class” library

```
install.packages("class")
library(class)

#Normalization
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x))) }
norm <- as.data.frame(lapply(data[,1:4], normalize))

set.seed(123)
data_spl <- sample(1:nrow(norm),size=nrow(norm)*0.7,replace = FA

train2 <- data[data_spl,] # 70% training data
test2 <- data[-data_spl,] # remaining 30% test data

train_labels <- data[data_spl,5]
test_labels <-data[-data_spl,5]
knn_pred <- knn(train=train2, test=test2, cl=train_labels, k=1)
```

view raw (https://gist.github.com/Harshit1694/614dc7641f42ddc83ac0b36bc83bd9dd/raw/982347e2f92886585f8157e1b4d6c0ee33dcaef9/classlib_kNN.R) classlib_kNN.R (https://gist.github.com/Harshit1694/614dc7641f42ddc83ac0b36bc83bd9dd#file-classlib_knn-r) hosted with ❤ by GitHub (https://github.com)

Output

```
For K=1
[1] "Iris-virginica"
```

We can see that both models predicted the same class ('Iris-virginica').

End Notes

KNN algorithm is one of the simplest classification algorithm. Even with such simplicity, it can give highly competitive results. KNN algorithm can also be used for regression problems. The only difference from the discussed methodology will be using averages of nearest neighbors rather than voting from nearest neighbors. KNN can be coded in a single line on R. I am yet to explore how can we use KNN algorithm on SAS.

Did you find the article useful? Have you used any other machine learning tool recently? Do you plan to use KNN in any of your business problems? If yes, share with us how you plan to go about it.

If you like what you just read & want to continue your analytics learning, subscribe to our emails (http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya), follow us on twitter (http://twitter.com/analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

utm_source=blog_article&utm_campaign=blog&pcampaignid=MKT-Other-global-all-co-prtnr-py-PartBadge-Mar2515-1)

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

Share this:

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

(https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/?share=linkedin&nb=1)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

(https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/?share=facebook&nb=1)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

(https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/?share=twitter&nb=1)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

(https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/?share=pocket&nb=1)

(https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/?share=reddit&nb=1)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

Related Articles

(https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)
Commonly used Machine Learning Algorithms (with Python and R Codes) (https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)
September 9, 2017
In "Algorithm"

(https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/)
Stock Prices Prediction Using Machine Learning and Deep Learning Techniques (with Python codes) (https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-

(https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)
Understanding Support Vector Machine algorithm from examples (along with code) (https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-

price-machine-learning-and-deep-learning-techniques-python/)
October 25, 2018
In "Deep Learning"

code/)
September 13, 2017
In "Algorithm"

TAGS : K NEAREST (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/K-NEAREST/), KNN (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/KNN/), KNN FROM SCRATCH (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/KNN-FROM-SCRATCH/), LIVE CODING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LIVE-CODING/), MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MACHINE-LEARNING/), SIMPLIED SERIES (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/SIMPLIED-SERIES/)

PREVIOUS ARTICLE

DeepMind is Using 'Neuron Deletion' to Understand Deep Neural Networks

(https://www.analyticsvidhya.com/blog/2018/03/deepmind-using-neuron-deleting-understand-deep-neural-networks/)

NEXT ARTICLE

AVBytes: AI & ML Developments this week – IBM's Library 46 Times Faster than TensorFlow, Baidu's Massive Self-Driving Dataset, the Technology behind AWS SageMaker, etc.

(https://www.analyticsvidhya.com/blog/2018/03/avbytes-ai-ml-developments-this-week-260318/)

Tavish Srivastava (Https://www.analyticsvidhya.com

(https://www.analyticsvidhya.com

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP /?UTM_SOURCE=HOME_BLOG_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

/blog/author /tavish1/)

/blog/author/tavish1/)

Tavish Srivastava, co-founder and Chief Strategy Officer of Analytics Vidhya, is an IIT Madras graduate and a passionate data-science professional with 8+ years of diverse experience in markets including the US, India and Singapore, domains including Digital Acquisitions, Customer Servicing and Customer Management, and industry including Retail Banking, Credit Cards and Insurance. He is fascinated by the idea of artificial intelligence inspired by human intelligence and enjoys every discussion, theory or even movie related to this idea.

This article is quite old and you might not get a prompt response from the author. We request you to post this comment on Analytics Vidhya's Discussion portal (https://discuss.analyticsvidhya.com/) to get your queries resolved

35 COMMENTS

HARSHAL Reply

October 10, 2014 at 3:29 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27697)

Useful article.
Can you share similar article for randomforest ?
What are limitations with data size for accuracy?

TAVISH SRIVASTAVA Reply

October 10, 2014 at 4:49 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27709)

Harshal,

We have already published many articles on random forest. Here is the link of the article on random forest on similar lines:http://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/ (http://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/). You can also subscribe to analyticsvidhya to get access to weekly updates on such articles.

SAURABH Reply

October 10, 2014 at 5:00 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27710)

Good one please share the value of
Red circle and green square

TAVISH SRIVASTAVA Reply

October 10, 2014 at 9:51 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27813)

Saurabh,

The first graph is for illustrating purposes. You can create a random dataset to check the algorithm.

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

DEBASHIS ROUT **Reply**

October 10, 2014 at 7:02 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27738)

I am currently doing part time MS in BI & Data Mining. I found this article is really helpful to understand in more detail and expecting to utilize in my upcoming project work. I need to know do you have any article on importance of Data quality in BI , Classification & Decision Tree.

TAVISH **Reply**

October 10, 2014 at 9:48 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27808)

Debashish,
We have published many articles on CART models before. Here is a link which will give you a kick start http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/ (http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/).

SARASWATHI Reply

October 10, 2014 at 1:44 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27943)

Hello
the article is very clear and precise. I would like some clarification on the single line
“To get the optimal value of K, you can segregate the training and validation from the initial dataset”. Do you mean that we segregate those points on the border of the boundaries for validation and keep the remaining for training. This is not very clear to me. Can you please elaborate ?

Thanks.

TAVISH SRIVASTAVA Reply

October 10, 2014 at 3:38 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27994)

Saraswathi,

Here is what I meant : Take the entire population, and randomly split it into two. Now on the training sample, score each validation observation with different k-values. The error curve will give you the best value of k.

Hope it becomes clear now.

SARASWATHI Reply

October 10, 2014 at 4:41 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-28016)

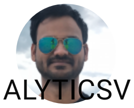
I want to make sure I understand this correctly.
Please confirm or correct.

You say, take the entire population and split into two – are these two divisions, one for training and one for validation ? (I am assuming so).

So, now I use different values of K to cluster the training samples.

I try to see where the validation samples fall in these clusters.

I draw the error curve and choose the K with the smallest error ?



TAVISH Reply

October 13, 2014 at 11:02 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-29158)

Saraswathi,
Let me make it even simpler. Say, you have 100 datapoints. Split this population into two samples of 70 and 30 observations. Use these 70 observation to predict for the other 30. Once you have the prediction for a particular value of k, check the misclassification with actual value. Repeat this exercise for different value of k. Hopefully, you will get a curve similar to that shown in the article. Now choose the k for which the misclassification is least.

Hope this makes it clear.

Tavish

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

HARVEY S **Reply**

October 10, 2014 at 1:50 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27944)

Nice tease: "KNN can be coded in a single line on R. "

Can you give an example?

TAVISH SRIVASTAVA **Reply**

October 10, 2014 at 3:38 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-27995)

We will cover this piece in our coming articles. Stay tuned.

FELIX **Reply**

October 13, 2014 at 9:02 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-29121)

Hi, great post, thanks. I would like to add, that "low calculation time" is not true for the prediction phase with big, high dimensional datasets. But it's still a good choice in many applications.

TAVISH **Reply**

October 13, 2014 at 11:03 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-29162)

Felix,

You are probably right for cases where the distance between observations comparable in the large dataset. But in general population have natural clusters which makes the calculation faster. Let me know in case you disagree.

Tavish

KABIR SINGH Reply

October 15, 2014 at 8:05 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-30127)

I am trying to figure out churn analysis, any suggestions where I am start looking?

BTW following this website for 6-8 months now, you guys are doing an amazing job

NAJMA NAAZ Reply

June 10, 2015 at 5:08 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-88260)

That was very helpful. Thank you! Can you please share a concise article on neural nets and deep learning as well?

TIAGO Reply

June 21, 2016 at 6:32 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-112504)

Thank you.

CHARLES Reply

February 12, 2017 at 6:45 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-122572)

Very useful. it was very explanatory. Thanks for that. Can you please post about adabooster algorithm?

AISHWARYA SINGH Reply

October 8, 2018 at 7:40 pm (https://www.analyticsvidhya.com

/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-155233)

You'll find it here : https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/ (https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/)

THUE XE DU LICH GIA RE Reply

June 26, 2017 at 4:24 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-131138)

Quality articles or reviews is the secret to invite the visitors to visit the website, that's what this web site is providing.

EMANUEL FAKHAR Reply

July 23, 2017 at 8:17 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-132717)

The world of DS would be so boring and exaggerated without you guys. Anything I study, I get a better perspective from this site. And you are so generous and grounded compared to idiots here in UK. God bless.

STONEHEAD PARK Reply

September 24, 2017 at 10:45 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-137835)

Excellent post, I appreciate your effort. 🙂

JUST81100 Reply

September 28, 2017 at 12:23 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-

138166)

BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM_SOURCE=HOME_BLOG_NAVBAR)

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM)

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM_SOURCE=HOME_BLOG_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP/?UTM_SOURCE=HOME_BLOG_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

KNN is fast to train but the prediction speed grows exponentially with the data set size and his complexity rather than Random forest…

SOUMYA SHREYA Reply

March 27, 2018 at 7:00 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152214)

It is a really nice and well explained article, I am a beginner in the field of data science and machine learning and I find these articles really helpful to learn and understand the algorithms. Thanks for publishing them.
Can you suggest me some datasets where I can experiment and apply KNN.

AISHWARYA SINGH Reply

March 29, 2018 at 1:32 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152256)

Hi Soumya,

You can use the Cancer dataset to practice kNN.
Refer this link (https://discuss.analyticsvidhya.com/t/practice-dataset-for-knn-algorithm/3104) for the same.

KRISHNA Reply

March 27, 2018 at 7:04 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152215)

How do we handle categorical features with kNN? Do we need to create dummies for them? Do you suggest any other distance method other than euclidean distance if we have more number of features? I feel we have to treat outliers as they may have impact on the distances, similarly missing values.. Please share your opinion.

AISHWARYA SINGH Reply

March 28, 2018 at 3:21 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152229)

Hi,

Yes you can create dummies for categorical variables in KNN.
Apart from Euclidean distance, there are other methods that can be used to find the distance such as Manhattan or Minkowski.
For outliers adn missing value treatment, you can refer this article (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/) .

AANISH SINGLA Reply

March 28, 2018 at 8:01 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152239)

IMO limitation of KNN comes into play when dimensions increase because in higher dimensions, finding neighbors which are quite close to each other in all dimensions might be tough, hence so called neighbors might be really far apart from each other which defeats the purpose of the algorithm.

Kindly share your thoughts/experiences.

AISHWARYA SINGH Reply

March 29, 2018 at 3:07 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152261)

Hi Aanish,

Thank you for sharing your thoughts.

AMLESH KANEKAR Reply

April 25, 2018 at 9:21 am (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152826)

I found it “inspiring”. Have spent last 4 months learning linear algebra, statistics, python. This learning list was culled from analyticsvidhya.com. Now just glimpsing through your article gives me the confidence to code knn from scratch. Thank you!

AISHWARYA SINGH Reply

April 25, 2018 at 4:16 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152839)

Hi Amlesh,

Glad you found this useful!

AMLESH KANEKAR Reply

May 2, 2018 at 6:46 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-152995)

I created my own dataset to experiment with KNN. When I plotted my data, the three targets/labels I have are extremely randomly distributed across the 2D plane … no clustering of the three colours is evident.
The Iris dataset shows a fairly high degree of clustering.
Should I continue with my dataset or there is the concept of “so-and-so distribution does not qualify for KNN”?
I can email a picture of my data plot if needed.

AMLESH KANEKAR Reply

May 8, 2018 at 12:53 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-153109)

I figured this out. So it is fine if you do not respond.

MAX Reply

October 6, 2018 at 12:21 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-155193)

That was very helpful. Thank you!

How to make the same visualization as in the pictures in section "How do we choose the factor K" ?

AISHWARYA SINGH Reply

October 8, 2018 at 7:46 pm (https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/#comment-155234)

Hi Max,

For this, you will have to use a for loop. For each value of k, calculate the validation error and store in a separate list. Then plot these validation error values against k values.

ANALYTICS VIDHYA

About Us (http://www.analyticsvidhya.com/about-me/)
Our Team (https://www.analyticsvidhya.com/about-me/team/)
Career (https://www.analyticsvidhya.com/career-analytics-vidhya/)
Contact Us (https://www.analyticsvidhya.com/contact/)
Write for us (https://www.analyticsvidhya.com/about-me/write/)

DATA SCIENTISTS

Blog (https://www.analyticsvidhya.com/blog)
Hackathon (https://datahack.analyticsvidhya.com/)
Discussions (https://discuss.analyticsvidhya.com/)
Apply Jobs (https://www.analyticsvidhya.com/jobs/)
Leaderboard (https://datahack.analyticsvidhya.com/users/)

COMPANIES

Post Jobs (https://www.analyticsvidhya.com/corporate/)
Trainings (https://trainings.analyticsvidhya.com)
Hiring Hackathons (https://datahack.analyticsvidhya.com/)
Advertising (https://www.analyticsvidhya.com/contact/)
Reach Us (https://www.analyticsvidhya.com/contact/)

JOIN OUR COMMUNITY :

f (https://www.facebook.com/AnalyticsVidhya/) 46336 (https://www.facebook.com/AnalyticsVidhya/) Followers (https://www.facebook.com/AnalyticsVidhya/)

(https://twitter.com/analyticsvidhya) 23018 (https://twitter.com/analyticsvidhya) Followers (https://twitter.com/analyticsvidhya)

G+ (https://plus.google.com/+Analyticsvidhya) (https://plus.google.com/+Analyticsvidhya) Followers (https://plus.google.com/+Analyticsvidhya)

in (https://in.linkedin.com/company/analytics-vidhya) 7513 (https://in.linkedin.com/company/analytics-vidhya) Followers (https://in.linkedin.com/company/analytics-vidhya)

Subscribe to emailer >

Privacy Policy (https://www.analyticsvidhya.com/privacy-policy/)
Terms of Use (https://www.analyticsvidhya.com/terms/)
Refund Policy (https://www.analyticsvidhya.com/refund-policy/)

Don't have an account? Sign up (https