



## IBM Developer SKILLS NETWORK

# K-Nearest Neighbors

Estimated time needed: **25** minutes

## Objectives

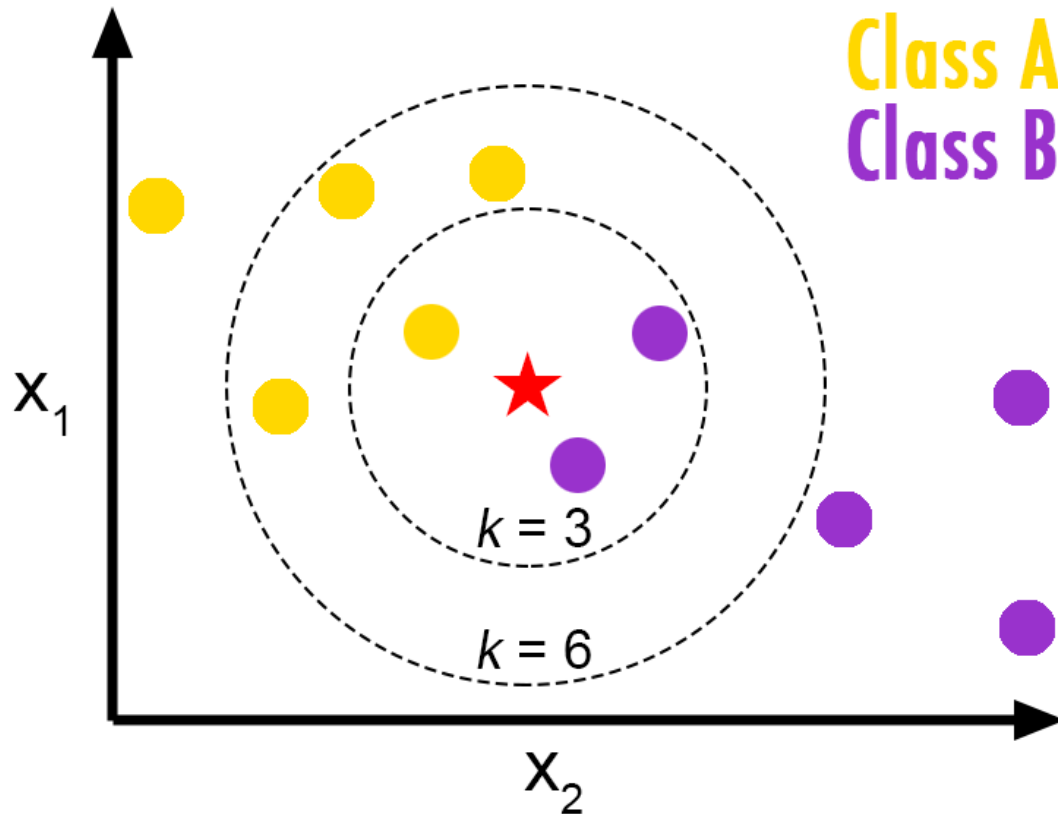
After completing this lab you will be able to:

- Use K Nearest neighbors to classify data

In this Lab you will load a customer dataset, fit the data, and use K-Nearest Neighbors to predict a data point. But what is **K-Nearest Neighbors**?

**K-Nearest Neighbors** is a supervised learning algorithm. Where the data is 'trained' with data points corresponding to their classification. To predict the class of a given data point, it takes into account the classes of the 'K' nearest data points and chooses the class in which the majority of the 'K' nearest data points belong to as the predicted class.

**Here's an visualization of the K-Nearest Neighbors algorithm.**



In this case, we have data points of Class A and B. We want to predict what the star (test data point) is. If we consider a  $k$  value of 3 (3 nearest data points), we will obtain a prediction of Class B. Yet if we consider a  $k$  value of 6, we will obtain a prediction of Class A.

In this sense, it is important to consider the value of  $k$ . Hopefully from this diagram, you should get a sense of what the K-Nearest Neighbors algorithm is. It considers the ' $K$ ' Nearest Neighbors (data points) when it predicts the classification of the test point.

## Table of contents

1. [About the dataset](#)
2. [Data Visualization and Analysis](#)
3. [Classification](#)

```
In [1]: !pip install scikit-learn==0.23.1
```

```
Collecting scikit-learn==0.23.1
  Downloading scikit_learn-0.23.1-cp38-cp38-manylinux1_x86_64.whl (6.7 MB)
```



```
a.s3.us.cloud-object-storage.appdomain.cloud)... 198.23.119.245
Connecting to cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud (cf-courses
-data.s3.us.cloud-object-storage.appdomain.cloud)|198.23.119.245|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 36047 (35K) [text/csv]
Saving to: 'teleCust1000t.csv'
```

```
teleCust1000t.csv 100%[=====>] 35.20K --.-KB/s in 0.002s
```

```
2021-11-07 00:43:42 (17.2 MB/s) - 'teleCust1000t.csv' saved [36047/36047]
```

**Did you know?** When it comes to Machine Learning, you will likely be working with large datasets. As a business, where can you host your data? IBM is offering a unique opportunity for businesses, with 10 Tb of IBM Cloud Object Storage: [Sign up now for free](#)

## Load Data From CSV File

```
In [4]: df = pd.read_csv('teleCust1000t.csv')
df.head()
```

```
Out[4]:
```

	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	13	44	1	9	64.0	4	5	0.0	0	2	1
1	3	11	33	1	7	136.0	5	5	0.0	0	6	4
2	3	68	52	1	24	116.0	1	29	0.0	1	2	3
3	2	33	33	0	12	33.0	2	0	0.0	1	1	1
4	2	23	30	1	9	30.0	1	2	0.0	0	4	3

## Data Visualization and Analysis

Let's see how many of each class is in our data set

```
In [5]: df['custcat'].value_counts()
```

```
Out[5]:
```

3	281
1	266
4	236
2	217

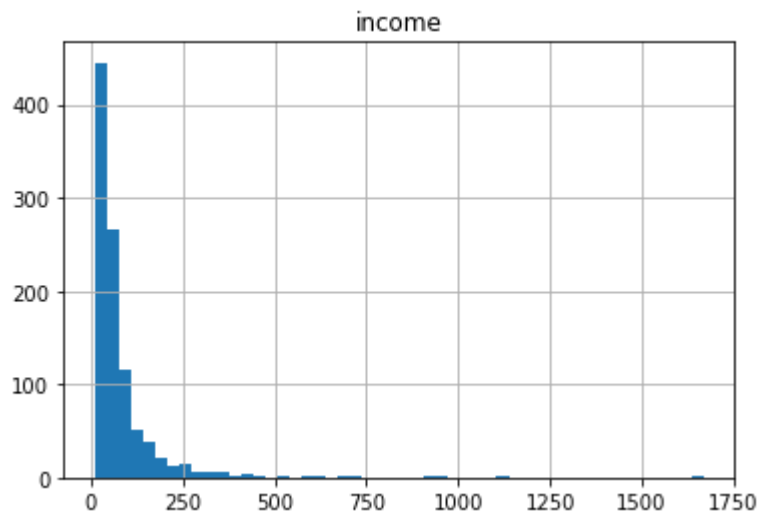
Name: custcat, dtype: int64

**281 Plus Service, 266 Basic-service, 236 Total Service, and 217 E-Service customers**

You can easily explore your data using visualization techniques:

```
In [6]: df.hist(column='income', bins=50)
```

```
Out[6]: array([[<AxesSubplot:title={'center':'income'}>]], dtype=object)
```



## Feature set

Let's define feature sets, X:

```
In [7]: df.columns
```

```
Out[7]: Index(['region', 'tenure', 'age', 'marital', 'address', 'income', 'ed',  
              'employ', 'retire', 'gender', 'reside', 'custcat'],  
             dtype='object')
```

To use scikit-learn library, we have to convert the Pandas data frame to a Numpy array:

```
In [8]: X = df[['region', 'tenure', 'age', 'marital', 'address', 'income', 'ed', 'employ', 're  
X[0:5]
```

```
Out[8]: array([[ 2., 13., 44.,  1.,  9., 64.,  4.,  5.,  0.,  0.,  2.],  
               [ 3., 11., 33.,  1.,  7., 136.,  5.,  5.,  0.,  0.,  6.],  
               [ 3., 68., 52.,  1., 24., 116.,  1., 29.,  0.,  1.,  2.],  
               [ 2., 33., 33.,  0., 12.,  33.,  2.,  0.,  0.,  1.,  1.],  
               [ 2., 23., 30.,  1.,  9.,  30.,  1.,  2.,  0.,  0.,  4.]])
```

What are our labels?

```
In [9]: y = df['custcat'].values  
y[0:5]
```

```
Out[9]: array([1, 4, 3, 1, 3])
```

## Normalize Data

Data Standardization gives the data zero mean and unit variance, it is good practice, especially for algorithms such as KNN which is based on the distance of data points:

```
In [10]: X = preprocessing.StandardScaler().fit(X).transform(X.astype(float))
X[0:5]
```

```
Out[10]: array([[ -0.02696767, -1.055125 ,  0.18450456,  1.0100505 , -0.25303431,
        -0.12650641,  1.0877526 , -0.5941226 , -0.22207644, -1.03459817,
        -0.23065004],
       [ 1.19883553, -1.14880563, -0.69181243,  1.0100505 , -0.4514148 ,
        0.54644972,  1.9062271 , -0.5941226 , -0.22207644, -1.03459817,
        2.55666158],
       [ 1.19883553,  1.52109247,  0.82182601,  1.0100505 ,  1.23481934,
        0.35951747, -1.36767088,  1.78752803, -0.22207644,  0.96655883,
        -0.23065004],
       [-0.02696767, -0.11831864, -0.69181243, -0.9900495 ,  0.04453642,
        -0.41625141, -0.54919639, -1.09029981, -0.22207644,  0.96655883,
        -0.92747794],
       [-0.02696767, -0.58672182, -0.93080797,  1.0100505 , -0.25303431,
        -0.44429125, -1.36767088, -0.89182893, -0.22207644, -1.03459817,
        1.16300577]])
```

## Train Test Split

Out of Sample Accuracy is the percentage of correct predictions that the model makes on data that the model has NOT been trained on. Doing a train and test on the same dataset will most likely have low out-of-sample accuracy, due to the likelihood of our model overfitting.

It is important that our models have a high, out-of-sample accuracy, because the purpose of any model, of course, is to make correct predictions on unknown data. So how can we improve out-of-sample accuracy? One way is to use an evaluation approach called Train/Test Split. Train/Test Split involves splitting the dataset into training and testing sets respectively, which are mutually exclusive. After which, you train with the training set and test with the testing set.

This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is not part of the dataset that has been used to train the model. It is more realistic for the real world problems.

```
In [11]: from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=42)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

```
Train set: (800, 11) (800,)
Test set: (200, 11) (200,)
```

## Classification

### K nearest neighbor (KNN)

#### Import library

Classifier implementing the k-nearest neighbors vote.

```
In [12]: from sklearn.neighbors import KNeighborsClassifier
```

## Training

Let's start the algorithm with k=4 for now:

```
In [13]: k = 4
#Train Model and Predict
neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)
neigh
```

```
Out[13]: KNeighborsClassifier(n_neighbors=4)
```

## Predicting

We can use the model to make predictions on the test set:

```
In [14]: yhat = neigh.predict(X_test)
yhat[0:5]
```

```
Out[14]: array([1, 1, 3, 2, 4])
```

## Accuracy evaluation

In multilabel classification, **accuracy classification score** is a function that computes subset accuracy. This function is equal to the jaccard\_score function. Essentially, it calculates how closely the actual labels and predicted labels are matched in the test set.

```
In [15]: from sklearn import metrics
print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh.predict(X_train))
print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))
```

```
Train set Accuracy:  0.5475
```

```
Test set Accuracy:  0.32
```

## Practice

Can you build the model again, but this time with k=6?

```
In [16]: k = 6
neigh6 = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)
yhat6 = neigh6.predict(X_test)
print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh6.predict(X_train))
print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat6))
```

```
Train set Accuracy:  0.51625
```

```
Test set Accuracy:  0.31
```

Click here for the solution `python k = 6 neigh6 = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train) yhat6 = neigh6.predict(X_test) print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh6.predict(X_train))) print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat6))`

## What about other K?

K in KNN, is the number of nearest neighbors to examine. It is supposed to be specified by the user. So, how can we choose right value for K? The general solution is to reserve a part of your data for testing the accuracy of the model. Then choose  $k=1$ , use the training part for modeling, and calculate the accuracy of prediction using all samples in your test set. Repeat this process, increasing the  $k$ , and see which  $k$  is the best for your model.

We can calculate the accuracy of KNN for different values of  $k$ .

In [17]:

```
Ks = 10
mean_acc = np.zeros((Ks-1))
std_acc = np.zeros((Ks-1))

for n in range(1,Ks):

    #Train Model and Predict
    neigh = KNeighborsClassifier(n_neighbors = n).fit(X_train,y_train)
    yhat=neigh.predict(X_test)
    mean_acc[n-1] = metrics.accuracy_score(y_test, yhat)

    std_acc[n-1]=np.std(yhat==y_test)/np.sqrt(yhat.shape[0])

mean_acc
```

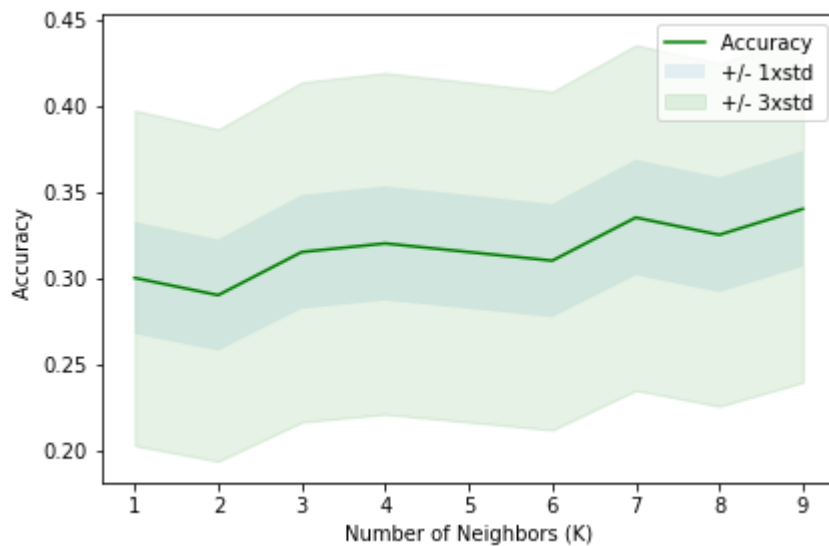
Out[17]: array([0.3 , 0.29 , 0.315, 0.32 , 0.315, 0.31 , 0.335, 0.325, 0.34 ])

## Plot the model accuracy for a different number of neighbors.

In [18]:

```
plt.plot(range(1,Ks),mean_acc,'g')
plt.fill_between(range(1,Ks),mean_acc - 1 * std_acc,mean_acc + 1 * std_acc, alpha=0.
plt.fill_between(range(1,Ks),mean_acc - 3 * std_acc,mean_acc + 3 * std_acc, alpha=0.
plt.legend(('Accuracy ', '+/- 1xstd', '+/- 3xstd'))
plt.ylabel('Accuracy ')
plt.xlabel('Number of Neighbors (K)')
plt.tight_layout()
plt.show()
```





```
In [19]: print( "The best accuracy was with", mean_acc.max(), "with k=", mean_acc.argmax()+1)
```

The best accuracy was with 0.34 with k= 9

## Want to learn more?

IBM SPSS Modeler is a comprehensive analytics platform that has many machine learning algorithms. It has been designed to bring predictive intelligence to decisions made by individuals, by groups, by systems – by your enterprise as a whole. A free trial is available through this course, available here: [SPSS Modeler](#)

Also, you can use Watson Studio to run these notebooks faster with bigger datasets. Watson Studio is IBM's leading cloud solution for data scientists, built by data scientists. With Jupyter notebooks, RStudio, Apache Spark and popular libraries pre-packaged in the cloud, Watson Studio enables data scientists to collaborate on their projects without having to install anything. Join the fast-growing community of Watson Studio users today with a free account at [Watson Studio](#)

**Thank you for completing this lab!**

## Author

Saeed Aghabozorgi

## Other Contributors

[Joseph Santarcangelo](#)

## Change Log

<b>Date (YYYY-MM-DD)</b>	<b>Version</b>	<b>Changed By</b>	<b>Change Description</b>
2021-01-21	2.4	Lakshmi	Updated sklearn library
2020-11-20	2.3	Lakshmi	Removed unused imports
2020-11-17	2.2	Lakshmi	Changed plot function of KNN
2020-11-03	2.1	Lakshmi	Changed URL of csv
2020-08-27	2.0	Lavanya	Moved lab to course repo in GitLab

© IBM Corporation 2020. All rights reserved.