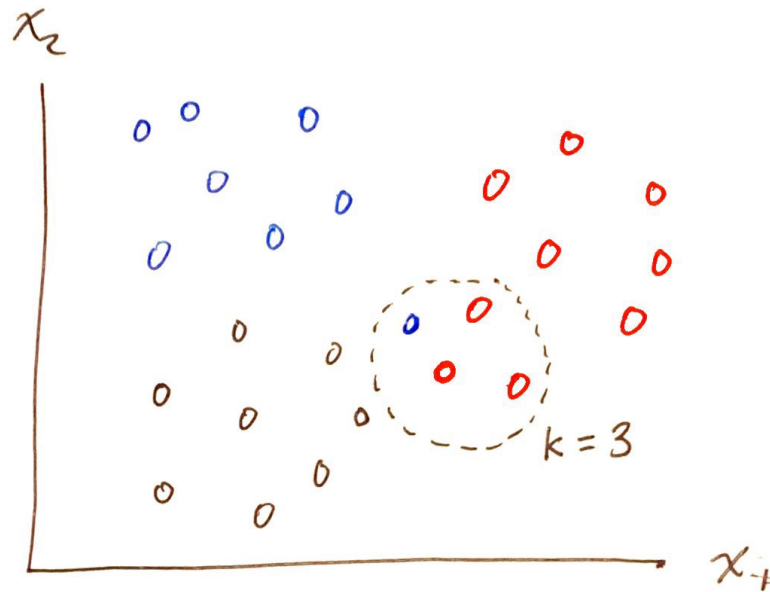# K - Nearest Neighbor


```
>>  group Data = read table ("groupData.csv")

>>  group Data . group = categorical ( ...

    group Data . group )


>>  cvpt = cv partition ( group Data . group ...
                "Hold Out", 0.35)

>>  data Train = group Data ( training (cvpt), :)

>>  data Test = group Data ( test (cvpt), :)
```

Construct a k-nearest neighbor classifier in MATLAB using fitcknn with k = 3 on the response variable "group"

```
>>    mdl = fitcknn ( dataTrain,
            "group", "NumNeigbors", 3)
```

Predict the groups for the test data

```
>>    label = predict (mdl, dataTest)
```
label

Plot the results over the original data

```
>>  gscatter ( groupData.x, groupData.y, ...
                    groupData.group)
```

```
>> hold on
```

```
>> gscatter ( dataTest.x, dataTest.y, ...
            label, [], "o", 15)
```
predicted results

Notice the point near $(7,1)$ has been misclassified. The true class blue is different from the predicted class green

What proportion of points have been misclassified?

```
>>  err3 = loss (mdl, dataTest)
```
            0.1880

Use more neighbors for the predictions $(k=10)$

```
>> mdl.NumNeighbors = 10
>>  err10 = loss (mdl, data Test)
```
            0.1171

Heart Disease Analysis

Bring in numerical heart data:

```
>> heartData = readtable ("heartDataNum.csv")
```

Response Variable ;

```
>> heartData. HeartDisease = categorical ( ...
              heartData. HeartDisease)
```

Set up Testing and Training Data

```
>> pt = cvpartition (heartData. HeartDisease, ...
            "HoldOut", 0.3)
```

```
>> hdTrain = heartData (training (pt), :)
>> hdTest = heartData (test (pt), :)
```

Create a k-nearest neighbor classification model:

```
>> mdl = fitcknn (hdTrain, "HeartDisease")
```

Change the number of neighbors to 5

>> mdl.NumNeighbors = 5

Set the distance weighting function
to "squared inverse"

>> mdl.DistanceWeight = "squared inverse"

Calculate the error for training (in-sample error)

>> errTrain = resubLoss(mdl)

0

Estimate the out-of-sample error using testing data:

>> errTest = loss(mdl, hdTest)

0.34388