k-Nearest Neighbors

Nik Bear Brown

In this lesson we'll learn the theory behind using k-nearest neighbors (kNN) as a supervised classification technique. We'll then use kNN to classify the UCI wine dataset in R.

# Additional packages needed

To run the code in M06\_Lesson\_01.Rmd you may need additional packages.

* If necessary install the followings packages.

install.packages("ggplot2");  
install.packages("class");

require(ggplot2)

## Loading required package: ggplot2

require(class)

## Loading required package: class

# Data

We will be using the [UCI Machine Learning Repository: Wine Data Set](https://archive.ics.uci.edu/ml/datasets/Wine). These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The attributes are:  
1) Alcohol  
2) Malic acid  
3) Ash  
4) Alcalinity of ash  
5) Magnesium  
6) Total phenols  
7) Flavanoids  
8) Nonflavanoid phenols  
9) Proanthocyanins  
10) Color intensity  
11) Hue  
12) OD280/OD315 of diluted wines  
13) Proline

Feel free to tweet questions to [@NikBearBrown](<https://twitter.com/NikBearBrown>)

# Load our data  
data\_url <- 'http://nikbearbrown.com/YouTube/MachineLearning/M06/wine.csv'  
wn <- read.csv(url(data\_url))  
head(wn)

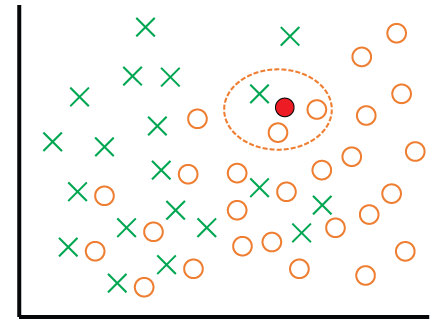
## Cultivar Alcohol Malic.acid Ash Alcalinity.ash Magnesium Total.phenols  
## 1 1 14.23 1.71 2.43 15.6 127 2.80  
## 2 1 13.20 1.78 2.14 11.2 100 2.65  
## 3 1 13.16 2.36 2.67 18.6 101 2.80  
## 4 1 14.37 1.95 2.50 16.8 113 3.85  
## 5 1 13.24 2.59 2.87 21.0 118 2.80  
## 6 1 14.20 1.76 2.45 15.2 112 3.27  
## Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity Hue  
## 1 3.06 0.28 2.29 5.64 1.04  
## 2 2.76 0.26 1.28 4.38 1.05  
## 3 3.24 0.30 2.81 5.68 1.03  
## 4 3.49 0.24 2.18 7.80 0.86  
## 5 2.69 0.39 1.82 4.32 1.04  
## 6 3.39 0.34 1.97 6.75 1.05  
## OD280.OD315 Proline  
## 1 3.92 1065  
## 2 3.40 1050  
## 3 3.17 1185  
## 4 3.45 1480  
## 5 2.93 735  
## 6 2.85 1450

# k-Nearest Neighbors (kNN)

A simple supervised learning algorithm is [k-Nearest Neighbors](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) algorithm (k-NN). KNN is a non-parametric method used for classification and regression.

In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. 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). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.



*k-nearest neighbor voting*

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN has the nice property that a labled subset of a data set could be used to label the whole data set. This is especially important in the analysis of “big-data.” Most big-data sets are only partially labled, as labeling often requires human annotation. While many are looking to unsupervised learning the ‘future’ of big-data, k-Nearest Neighbors is an instance of a supervised learning algorithm that can be used with big-data.

The kNN classification problem is to find the k nearest data points in a data set to a given query data point. The point is then assigned to the group by a majority "vote." For this reason, pick an odd k is prefered as the odd vote can break ties. This operation is also known as a kNN join, and can be defined as: given two data sets and , find the k nearest Neighbor from for every object in . refers to data that has already been classified, the training set. refers to data that is needs to be classified.

The kNN algorithm can be fairly expensive, especially if one chooses a large k, as the k-nearest neighbors in for every point in needs to be calculated.

## Nearest neighbor search

A simple solution to finding nearest neighbors is to compute the distance from the each point in to every point in and keeping track of the "best so far". This algorithm, sometimes referred to as the naive approach, has a running time of O(|R||S|).

One can speed up the search to retrieve a "good guess" of the nearest neighbor. This is often done be limiting the search to a preset radius around a point culling out most of the points in . If k neighbors aren't not found in the radius then the bound can be iteratively expanded until k are found. Altnernatively, the vote could be made using fewer points when k points aren't found within a radius r.

## k-Nearest Neighbors is nonparametric "lazy learning "

K-Nearest Neighbors algorithm (kNN) is a nonparametric method for classifying objects based on the closest training examples in the feature space. kNN is nonparametric becuase it d oes not involve any estimation of parameters. This is sometimes called "lazy learning" or instance-based learning, as the mapping is approximated locally and all computation is deferred until classification.

# kNN Classification and Distance Metrics

Neighbors are defined by a distance or dissimilarity measure. In essence, the only thing that kNN requires is some measure of "closeness" of the points in and . Any distance metric or dissimilarity measure can be used. The most common being the Euclidean distance between the points and is given by the pythagorean formula:

Any measure of "closeness", distance or dissimilarity measure can be used. For example,

* [Chebyshev distance](https://en.wikipedia.org/wiki/Chebyshev_distance) - measures distance assuming only the most significant dimension is relevant.
* [Hamming distance](https://en.wikipedia.org/wiki/Hamming_distance) - identifies the difference bit by bit of two strings
* [Mahalanobis distance](https://en.wikipedia.org/wiki/Mahalanobis_distance) - normalizes based on a covariance matrix to make the distance metric scale-invariant.
* [Manhattan distance](https://en.wikipedia.org/wiki/Taxicab_geometry) - measures distance following only axis-aligned directions.
* [Minkowski distance](https://en.wikipedia.org/wiki/Minkowski_distance) - is a generalization that unifies Euclidean distance, Manhattan distance, and Chebyshev distance

.. and many more.

# kNN Algorithm

## Distance function

The distance function depends on your needs, but in general choosing features and distance metrics in which being "close" makes some sense in your domain are the distance metrics and features to choose. The type of variable, categorical, ordinal or nominal should be considered when choosing a sensible measure of closeness.

## k nearest neighbors

Given an data point p, a training data set , and an integer k, the k nearest neighbors of p from , denoted as kNN(p, S), are a set of k objects from such that:

## kNN join

Given two data sets R and S (where S is a training data set) and an integer k, the kNN join of R and S is defined as:

kNNjoin(R, S) = {(r, s)|∀r ∊ R, ∀s ∊ kNN(r, S)}

Basically, this combines each object r ∊ R with its k nearest neighbors from S.

# Steps in kNN Classification

The kNN algorithm can be summarized in the following simple steps:

* Determine k (the selection of k depends on your data and project requirements; there is no magic formula for k).
* Calculate the distances between the new input and all the training data (as with k, the selection of a distance function also depends on the type of data).
* Sort the distance and determine the k nearest neighbors based on the kth minimum distance.
* Gather the categories of those neighbors.
* Determine the category based on majority vote.

# k-Nearest Neighbors (kNN) in R

k-Nearest Neighbors (kNN) in R

head(wn)

## Cultivar Alcohol Malic.acid Ash Alcalinity.ash Magnesium Total.phenols  
## 1 1 14.23 1.71 2.43 15.6 127 2.80  
## 2 1 13.20 1.78 2.14 11.2 100 2.65  
## 3 1 13.16 2.36 2.67 18.6 101 2.80  
## 4 1 14.37 1.95 2.50 16.8 113 3.85  
## 5 1 13.24 2.59 2.87 21.0 118 2.80  
## 6 1 14.20 1.76 2.45 15.2 112 3.27  
## Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity Hue  
## 1 3.06 0.28 2.29 5.64 1.04  
## 2 2.76 0.26 1.28 4.38 1.05  
## 3 3.24 0.30 2.81 5.68 1.03  
## 4 3.49 0.24 2.18 7.80 0.86  
## 5 2.69 0.39 1.82 4.32 1.04  
## 6 3.39 0.34 1.97 6.75 1.05  
## OD280.OD315 Proline  
## 1 3.92 1065  
## 2 3.40 1050  
## 3 3.17 1185  
## 4 3.45 1480  
## 5 2.93 735  
## 6 2.85 1450

summary(wn)

## Cultivar Alcohol Malic.acid Ash   
## Min. :1.000 Min. :11.03 Min. :0.740 Min. :1.360   
## 1st Qu.:1.000 1st Qu.:12.36 1st Qu.:1.603 1st Qu.:2.210   
## Median :2.000 Median :13.05 Median :1.865 Median :2.360   
## Mean :1.938 Mean :13.00 Mean :2.336 Mean :2.367   
## 3rd Qu.:3.000 3rd Qu.:13.68 3rd Qu.:3.083 3rd Qu.:2.558   
## Max. :3.000 Max. :14.83 Max. :5.800 Max. :3.230   
## Alcalinity.ash Magnesium Total.phenols Flavanoids   
## Min. :10.60 Min. : 70.00 Min. :0.980 Min. :0.340   
## 1st Qu.:17.20 1st Qu.: 88.00 1st Qu.:1.742 1st Qu.:1.205   
## Median :19.50 Median : 98.00 Median :2.355 Median :2.135   
## Mean :19.49 Mean : 99.74 Mean :2.295 Mean :2.029   
## 3rd Qu.:21.50 3rd Qu.:107.00 3rd Qu.:2.800 3rd Qu.:2.875   
## Max. :30.00 Max. :162.00 Max. :3.880 Max. :5.080   
## Nonflavanoid.phenols Proanthocyanins Color.intensity Hue   
## Min. :0.1300 Min. :0.410 Min. : 1.280 Min. :0.4800   
## 1st Qu.:0.2700 1st Qu.:1.250 1st Qu.: 3.220 1st Qu.:0.7825   
## Median :0.3400 Median :1.555 Median : 4.690 Median :0.9650   
## Mean :0.3619 Mean :1.591 Mean : 5.058 Mean :0.9574   
## 3rd Qu.:0.4375 3rd Qu.:1.950 3rd Qu.: 6.200 3rd Qu.:1.1200   
## Max. :0.6600 Max. :3.580 Max. :13.000 Max. :1.7100   
## OD280.OD315 Proline   
## Min. :1.270 Min. : 278.0   
## 1st Qu.:1.938 1st Qu.: 500.5   
## Median :2.780 Median : 673.5   
## Mean :2.612 Mean : 746.9   
## 3rd Qu.:3.170 3rd Qu.: 985.0   
## Max. :4.000 Max. :1680.0

length(wn)

## [1] 14

names(wn)

## [1] "Cultivar" "Alcohol" "Malic.acid"   
## [4] "Ash" "Alcalinity.ash" "Magnesium"   
## [7] "Total.phenols" "Flavanoids" "Nonflavanoid.phenols"  
## [10] "Proanthocyanins" "Color.intensity" "Hue"   
## [13] "OD280.OD315" "Proline"

table(wn$Cultivar)

##   
## 1 2 3   
## 59 71 48

wn$Cultivar

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2  
## [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [106] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3  
## [141] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3  
## [176] 3 3 3

length(wn$Cultivar)

## [1] 178

You can also embed plots, for example:

shuff<-runif(nrow(wn))  
shuff

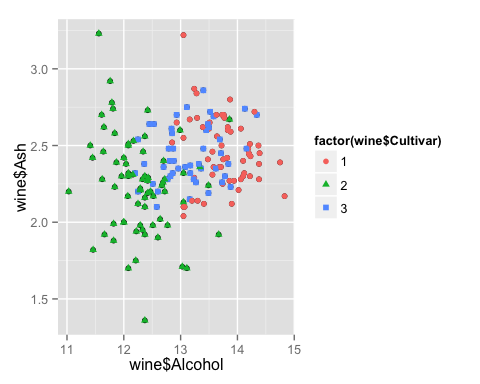
## [1] 0.86040343 0.94726813 0.30854151 0.23959085 0.08480826 0.90766740  
## [7] 0.49355557 0.34927525 0.29890129 0.66643548 0.20373893 0.98909719  
## [13] 0.57859886 0.57515568 0.46491521 0.39184609 0.69297081 0.70753905  
## [19] 0.06801408 0.50979908 0.61556573 0.82425297 0.76172337 0.46517681  
## [25] 0.04502175 0.44817780 0.76922207 0.38229169 0.86672808 0.79873973  
## [31] 0.59141105 0.45173008 0.43330529 0.75857639 0.68586255 0.03160162  
## [37] 0.90430290 0.81327637 0.99834164 0.99017623 0.77754215 0.22803301  
## [43] 0.66059460 0.11873175 0.80512270 0.88449910 0.51162573 0.79342553  
## [49] 0.69487487 0.23144663 0.10531213 0.57450771 0.18929456 0.86227823  
## [55] 0.09790657 0.37628943 0.96045245 0.93053597 0.55738746 0.69452584  
## [61] 0.82301554 0.90429616 0.53973858 0.72891073 0.96758482 0.22583370  
## [67] 0.49464994 0.69703601 0.62995030 0.70906383 0.23094117 0.03398969  
## [73] 0.26262968 0.45263712 0.03732332 0.06022753 0.89635431 0.61679064  
## [79] 0.94272822 0.40605207 0.20112220 0.38634258 0.87754654 0.99627146  
## [85] 0.51006638 0.35148212 0.67942097 0.96637977 0.37329989 0.56125771  
## [91] 0.93725659 0.33371225 0.95651629 0.51433023 0.08773586 0.73872608  
## [97] 0.65852318 0.76311976 0.71970095 0.84212361 0.41814677 0.85304323  
## [103] 0.20914964 0.45809334 0.12371383 0.85257121 0.30673234 0.35068404  
## [109] 0.26980672 0.43960246 0.74519896 0.94761209 0.97125186 0.22780960  
## [115] 0.48392175 0.39218617 0.50588827 0.04040413 0.06148230 0.95152443  
## [121] 0.29015652 0.69412625 0.57035175 0.21436058 0.03898371 0.87940275  
## [127] 0.93771194 0.14677246 0.69265547 0.82830839 0.17884698 0.79174186  
## [133] 0.85996275 0.79282795 0.80388160 0.35236828 0.13648932 0.51246715  
## [139] 0.59179679 0.74992217 0.53123790 0.31944785 0.27521809 0.45490890  
## [145] 0.05698440 0.19627665 0.89848833 0.07411641 0.76568906 0.81402996  
## [151] 0.43160056 0.07096682 0.85320295 0.55168947 0.32361869 0.18922350  
## [157] 0.54068920 0.99569145 0.21614557 0.93672354 0.31411950 0.70021709  
## [163] 0.26024668 0.43457342 0.26147091 0.66416955 0.27287037 0.30255428  
## [169] 0.20918072 0.96833979 0.55112054 0.25786433 0.96566245 0.65779887  
## [175] 0.65977960 0.43371769 0.35579230 0.67297356

wine<-wn[order(shuff),]  
wine$Cultivar

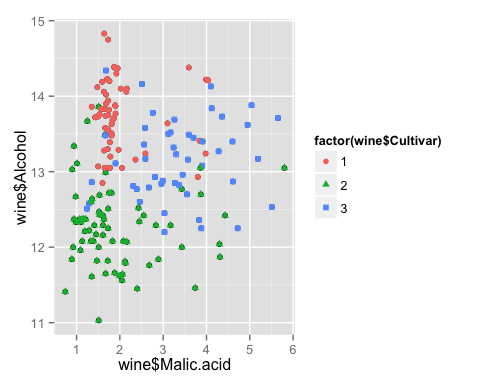
## [1] 1 2 2 2 2 1 3 2 2 1 3 3 1 2 1 1 1 2 3 2 3 3 1 3 2 1 2 3 2 3 2 2 1 2 1  
## [36] 1 3 3 3 2 2 3 3 2 1 3 2 1 3 3 3 2 1 2 2 3 3 2 1 1 2 1 2 2 2 3 1 3 3 2  
## [71] 1 1 2 3 2 1 1 2 1 2 2 1 2 1 3 2 3 2 3 3 3 1 2 2 1 1 1 1 3 1 2 2 3 2 3  
## [106] 1 3 1 3 2 1 2 1 2 2 1 2 3 1 2 2 2 2 2 3 1 1 2 3 1 1 3 3 1 1 3 1 1 3 2  
## [141] 1 2 2 2 2 3 3 1 1 1 2 2 1 2 3 2 1 1 1 3 2 2 2 1 2 2 2 1 3 2 2 3 2 1 1  
## [176] 3 2 1

You can also embed plots, for example:

qplot(wine$Alcohol,wine$Ash,data=wine)+geom\_point(aes(colour = factor(wine$Cultivar),shape = factor(wine$Cultivar)))



qplot(wine$Malic.acid,wine$Alcohol,data=wine)+geom\_point(aes(colour = factor(wine$Cultivar),shape = factor(wine$Cultivar)))



summary(wine)

## Cultivar Alcohol Malic.acid Ash   
## Min. :1.000 Min. :11.03 Min. :0.740 Min. :1.360   
## 1st Qu.:1.000 1st Qu.:12.36 1st Qu.:1.603 1st Qu.:2.210   
## Median :2.000 Median :13.05 Median :1.865 Median :2.360   
## Mean :1.938 Mean :13.00 Mean :2.336 Mean :2.367   
## 3rd Qu.:3.000 3rd Qu.:13.68 3rd Qu.:3.083 3rd Qu.:2.558   
## Max. :3.000 Max. :14.83 Max. :5.800 Max. :3.230   
## Alcalinity.ash Magnesium Total.phenols Flavanoids   
## Min. :10.60 Min. : 70.00 Min. :0.980 Min. :0.340   
## 1st Qu.:17.20 1st Qu.: 88.00 1st Qu.:1.742 1st Qu.:1.205   
## Median :19.50 Median : 98.00 Median :2.355 Median :2.135   
## Mean :19.49 Mean : 99.74 Mean :2.295 Mean :2.029   
## 3rd Qu.:21.50 3rd Qu.:107.00 3rd Qu.:2.800 3rd Qu.:2.875   
## Max. :30.00 Max. :162.00 Max. :3.880 Max. :5.080   
## Nonflavanoid.phenols Proanthocyanins Color.intensity Hue   
## Min. :0.1300 Min. :0.410 Min. : 1.280 Min. :0.4800   
## 1st Qu.:0.2700 1st Qu.:1.250 1st Qu.: 3.220 1st Qu.:0.7825   
## Median :0.3400 Median :1.555 Median : 4.690 Median :0.9650   
## Mean :0.3619 Mean :1.591 Mean : 5.058 Mean :0.9574   
## 3rd Qu.:0.4375 3rd Qu.:1.950 3rd Qu.: 6.200 3rd Qu.:1.1200   
## Max. :0.6600 Max. :3.580 Max. :13.000 Max. :1.7100   
## OD280.OD315 Proline   
## Min. :1.270 Min. : 278.0   
## 1st Qu.:1.938 1st Qu.: 500.5   
## Median :2.780 Median : 673.5   
## Mean :2.612 Mean : 746.9   
## 3rd Qu.:3.170 3rd Qu.: 985.0   
## Max. :4.000 Max. :1680.0

You can also embed plots, for example:

wine.scaled<-as.data.frame(lapply(wine[,c(2:14)], scale))  
head(wine.scaled)

## Alcohol Malic.acid Ash Alcalinity.ash Magnesium  
## 1 0.5904981 -0.4711544 0.15849862 0.3009543 0.01809398  
## 2 1.0585784 -0.7396958 1.10621386 1.6484357 -0.96212770  
## 3 -1.2818231 -1.1156538 -0.24245783 0.4506745 0.08810981  
## 4 -1.3926842 1.7666908 0.08559744 0.4506745 -1.24219104  
## 5 -0.7151996 -0.6501820 -0.64341428 0.8998349 0.57822065  
## 6 0.6151339 -0.4711544 0.88751034 0.1512342 -0.26196936  
## Total.phenols Flavanoids Nonflavanoid.phenols Proanthocyanins  
## 1 0.6469393 0.95181666 -0.8184106 0.47016154  
## 2 1.0463954 0.83167955 -1.2201676 0.48763309  
## 3 1.7334598 0.11085686 -1.8629788 0.10325886  
## 4 0.9025912 1.00187379 -1.2201676 2.30467493  
## 5 -0.4715377 0.06079973 -0.1755994 0.03337264  
## 6 0.3753092 0.58139389 -0.6577078 0.12073042  
## Color.intensity Hue OD280.OD315 Proline  
## 1 0.01807806 0.3611585 1.2089101 0.5497067  
## 2 -0.72384942 1.7611577 0.7722845 -1.0698166  
## 3 -0.79717946 0.1424086 0.7300304 0.4417385  
## 4 -0.97403427 -0.9075907 1.4483500 -1.1650827  
## 5 -1.29323563 0.4486584 0.4905906 -1.2762265  
## 6 -0.66345997 0.7111583 1.7018745 0.3115415

summary(wine.scaled)

## Alcohol Malic.acid Ash   
## Min. :-2.42739 Min. :-1.4290 Min. :-3.66881   
## 1st Qu.:-0.78603 1st Qu.:-0.6569 1st Qu.:-0.57051   
## Median : 0.06083 Median :-0.4219 Median :-0.02375   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.83378 3rd Qu.: 0.6679 3rd Qu.: 0.69615   
## Max. : 2.25341 Max. : 3.1004 Max. : 3.14745   
## Alcalinity.ash Magnesium Total.phenols   
## Min. :-2.663505 Min. :-2.0824 Min. :-2.10132   
## 1st Qu.:-0.687199 1st Qu.:-0.8221 1st Qu.:-0.88298   
## Median : 0.001514 Median :-0.1219 Median : 0.09569   
## Mean : 0.000000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.600395 3rd Qu.: 0.5082 3rd Qu.: 0.80672   
## Max. : 3.145637 Max. : 4.3591 Max. : 2.53237   
## Flavanoids Nonflavanoid.phenols Proanthocyanins   
## Min. :-1.6912 Min. :-1.8630 Min. :-2.06321   
## 1st Qu.:-0.8252 1st Qu.:-0.7381 1st Qu.:-0.59560   
## Median : 0.1059 Median :-0.1756 Median :-0.06272   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.8467 3rd Qu.: 0.6078 3rd Qu.: 0.62741   
## Max. : 3.0542 Max. : 2.3956 Max. : 3.47527   
## Color.intensity Hue OD280.OD315 Proline   
## Min. :-1.6297 Min. :-2.08884 Min. :-1.8897 Min. :-1.4890   
## 1st Qu.:-0.7929 1st Qu.:-0.76540 1st Qu.:-0.9496 1st Qu.:-0.7824   
## Median :-0.1588 Median : 0.03303 Median : 0.2371 Median :-0.2331   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.4926 3rd Qu.: 0.71116 3rd Qu.: 0.7864 3rd Qu.: 0.7561   
## Max. : 3.4258 Max. : 3.29241 Max. : 1.9554 Max. : 2.9631

You can also embed plots, for example:

normalize<- function(x) {  
 return((x-min(x))/(max(x)-min(x)))  
}  
wine.normalized<-as.data.frame(lapply(wine[,c(2:14)],normalize))  
head(wine.normalized)

## Alcohol Malic.acid Ash Alcalinity.ash Magnesium Total.phenols  
## 1 0.6447368 0.21146245 0.5614973 0.5103093 0.3260870 0.5931034  
## 2 0.7447368 0.15217391 0.7005348 0.7422680 0.1739130 0.6793103  
## 3 0.2447368 0.06916996 0.5026738 0.5360825 0.3369565 0.8275862  
## 4 0.2210526 0.70553360 0.5508021 0.5360825 0.1304348 0.6482759  
## 5 0.3657895 0.17193676 0.4438503 0.6134021 0.4130435 0.3517241  
## 6 0.6500000 0.21146245 0.6684492 0.4845361 0.2826087 0.5344828  
## Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity  
## 1 0.5569620 0.2452830 0.4574132 0.3259386  
## 2 0.5316456 0.1509434 0.4605678 0.1791809  
## 3 0.3797468 0.0000000 0.3911672 0.1646758  
## 4 0.5675105 0.1509434 0.7886435 0.1296928  
## 5 0.3691983 0.3962264 0.3785489 0.0665529  
## 6 0.4789030 0.2830189 0.3943218 0.1911263  
## Hue OD280.OD315 Proline  
## 1 0.4552846 0.8058608 0.45791726  
## 2 0.7154472 0.6923077 0.09415121  
## 3 0.4146341 0.6813187 0.43366619  
## 4 0.2195122 0.8681319 0.07275321  
## 5 0.4715447 0.6190476 0.04778887  
## 6 0.5203252 0.9340659 0.40442225

summary(wine.normalized)

## Alcohol Malic.acid Ash Alcalinity.ash   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.3507 1st Qu.:0.1705 1st Qu.:0.4545 1st Qu.:0.3402   
## Median :0.5316 Median :0.2223 Median :0.5348 Median :0.4588   
## Mean :0.5186 Mean :0.3155 Mean :0.5382 Mean :0.4585   
## 3rd Qu.:0.6967 3rd Qu.:0.4629 3rd Qu.:0.6404 3rd Qu.:0.5619   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Magnesium Total.phenols Flavanoids Nonflavanoid.phenols  
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.1957 1st Qu.:0.2629 1st Qu.:0.1825 1st Qu.:0.2642   
## Median :0.3043 Median :0.4741 Median :0.3787 Median :0.3962   
## Mean :0.3233 Mean :0.4535 Mean :0.3564 Mean :0.4375   
## 3rd Qu.:0.4022 3rd Qu.:0.6276 3rd Qu.:0.5348 3rd Qu.:0.5802   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Proanthocyanins Color.intensity Hue OD280.OD315   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.2650 1st Qu.:0.1655 1st Qu.:0.2459 1st Qu.:0.2445   
## Median :0.3612 Median :0.2910 Median :0.3943 Median :0.5531   
## Mean :0.3725 Mean :0.3224 Mean :0.3882 Mean :0.4915   
## 3rd Qu.:0.4858 3rd Qu.:0.4198 3rd Qu.:0.5203 3rd Qu.:0.6960   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Proline   
## Min. :0.0000   
## 1st Qu.:0.1587   
## Median :0.2821   
## Mean :0.3344   
## 3rd Qu.:0.5043   
## Max. :1.0000

nrow(wine)

## [1] 178

You can also embed plots, for example:

wine.normalized.train<-wine.normalized[1:150,]  
wine.normalized.test<-wine.normalized[151:178,]  
wine.normalized.train.target<-wine[1:150,c(1)]  
wine.normalized.test.target<-wine[151:178,c(1)]  
wine.normalized.test.target

## [1] 2 2 1 2 3 2 1 1 1 3 2 2 2 1 2 2 2 1 3 2 2 3 2 1 1 3 2 1

k<-5  
knn.m1<-knn(train = wine.normalized.train, test = wine.normalized.test,wine.normalized.train.target,k)  
knn.m1

## [1] 2 2 1 2 3 3 1 1 1 3 2 2 2 1 2 2 2 1 3 2 2 3 2 1 1 3 3 1  
## Levels: 1 2 3

length(knn.m1)

## [1] 28

cm<-table(wine.normalized.test.target,knn.m1)  
cm

## knn.m1  
## wine.normalized.test.target 1 2 3  
## 1 9 0 0  
## 2 0 12 2  
## 3 0 0 5

# Assingment

* Go to the [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/) and find a dataset for supervised classification. Every student MUST use a different dataset so you MUST get approved for which you can going to use. This can be the same dataset you used for the unsupervised clustering as long as the data has some labeled data.
* Classify your data using k-Nearest Neighbors. Answer the following questions:
  + Does the k for kNN make a difference? Try for a range of values of k.
  + Does scaling, normalization or leaving the data unscaled make a difference for kNN?

# Resources

* [Using R For k-Nearest Neighbors (KNN)](http://blog.datacamp.com/machine-learning-in-r/)
* [Using the k-Nearest Neighbors Algorithm in R](http://blog.webagesolutions.com/archives/1164)
* [kNN PSU](https://onlinecourses.science.psu.edu/stat857/node/129)