**Scarlett Townsend**

**STAT 6620 – Spring 2015**

**CSU East Bay**

**Prof. Seuss**

**Homework #2 – k- Nearest Neighbors Classification**

**Links to sections:**

[**Breast Cancer Wisconsin (Diagnostic) Data Set**](#Cancer)

[**Iris Data Set**](#Iris)

**Breast Cancer Wisconsin (Diagnostic) Data Set**

**Introduction**

*Step 1: Collecting the data*

Data was obtained from the University of California at Irvine Machine Learning Repository which was donated in 1995 for a study on breast cancer. The data contains features of images of the nuclei of cells obtained via fine needle aspiration of breast masses. These features are:

a) radius (mean of distances from center to points on the perimeter)   
b) texture (standard deviation of gray-scale values)   
c) perimeter   
d) area   
e) smoothness (local variation in radius lengths)   
f) compactness (perimeter^2 / area - 1.0)   
g) concavity (severity of concave portions of the contour)   
h) concave points (number of concave portions of the contour)   
i) symmetry   
j) fractal dimension ("coastline approximation" - 1)

All these characteristics were provided as their mean, standard error and worst. Also included are the ID numbers and diagnoses of the patients (B=benign and M=malignant).

**Method**

*Step 2: Exploring and preparing the data*

To prepare the data, the ID column was dropped as it was irrelevant to the analysis. In addition the diagnosis column was reclassified as a factor and several summary statistics were examined to ensure data sufficiency. There were 357 benign instances observed and 212 cases of malignancy present in the data. Examining the proportion we see this is 62.7% benign and 37.3% malignant. Summary statistics for the three main descriptive elements of interest are shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **radius\_mean** | **area\_mean** | **smoothness\_mean** |
| Minimum | 6.981 | 143.5 | 0.05263 |
| 1st Quartile | 11.7 | 420.3 | 0.08637 |
| Median | 13.37 | 551.1 | 0.09587 |
| Mean | 14.127 | 654.9 | 0.09636 |
| 3rd Quartile | 15.78 | 782.7 | 0.1053 |
| Maximum | 28.11 | 501 | 0.1634 |

*Step 3: Training a model on the data*

This data set was divided and 469 examples were used to train a nearest neighbors algorithm which was then applied to the remaining 100 records.

Using R’s class package, the knn function was applied to the data.  This algorithm uses Euclidian distance to determine the k closest points to the test point.  Each “neighbor’s” assigned category, in this case benign or malignant, is given equivalent weight and the category having the most neighbors becomes the category of the test point.

Optimizing and refining the model involved the use of both z-score normalization and trials with a variety of k-values. The results of the normalization on this data for the area are shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **radius\_mean** | **area\_mean** | **smoothness\_mean** |
| Minimum | 0 | 0 | 0 |
| 1st Quartile | 0.2233 | 0.1174 | 0.3046 |
| Median | 0.3024 | 0.1729 | 0.3904 |
| Mean | 0.3382 | 0.2169 | 0.3948 |
| 3rd Quartile | 0.4164 | 0.2711 | 0.4755 |
| Maximum | 1 | 1 | 1 |

**Results**

*Step 4: Evaluating model performance*

The results of the nearest neighbor trials without z-score normalization are shown in the tables below with the k value indicating the number of neighbors used to predict the diagnosis of patient.

|  |  |  |  |
| --- | --- | --- | --- |
| **k value** | **False negatives** | **False positives** | **% Incorrectly classified** |
| 1 | 1 | 3 | 4 |
| 5 | 2 | 0 | 2 |
| 11 | 3 | 0 | 3 |
| 15 | 3 | 0 | 3 |
| 21 | 2 | 0 | 2 |
| 27 | 4 | 0 | 4 |

*Step 5: Improving model performance*

Attempts were made to further refine the model with z-score normalization, however, in this case the results became less accurate than they were with the raw data.  Using k= 21, the number of false negatives calculated was 2.  After the data was normalized, this result increased to 5.

**Conclusion**

While k=21 resulted in the greatest overall accuracy, k=1 was able to reduce the false negatives to one. Although this trial added additional false positives, increasing the overall rate of misclassifications, it can be argued that it would be better to error on the side of diagnosing benign tumors as malignant than the reverse.

It should also be noted, however, that a different set of 100 test patients may have different accuracy results for the number of neighbors used and this should be evaluated before selecting a model.

**Sources**

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>

Lantz, Brett. Machine Learning with R. 2013.

**Appendix 1: Breast Cancer Code**

> # import the CSV file

> wbcd <- read.csv("wisc\_bc\_data.csv", stringsAsFactors = FALSE)

>

> # examine the structure of the wbcd data frame

> str(wbcd)

'data.frame': 569 obs. of 32 variables:

$ id : int 87139402 8910251 905520 868871 9012568 906539 925291 87880 862989 89827 ...

$ diagnosis : chr "B" "B" "B" "B" ...

$ radius\_mean : num 12.3 10.6 11 11.3 15.2 ...

$ texture\_mean : num 12.4 18.9 16.8 13.4 13.2 ...

$ perimeter\_mean : num 78.8 69.3 70.9 73 97.7 ...

$ area\_mean : num 464 346 373 385 712 ...

$ smoothness\_mean : num 0.1028 0.0969 0.1077 0.1164 0.0796 ...

$ compactness\_mean : num 0.0698 0.1147 0.078 0.1136 0.0693 ...

$ concavity\_mean : num 0.0399 0.0639 0.0305 0.0464 0.0339 ...

$ points\_mean : num 0.037 0.0264 0.0248 0.048 0.0266 ...

$ symmetry\_mean : num 0.196 0.192 0.171 0.177 0.172 ...

$ dimension\_mean : num 0.0595 0.0649 0.0634 0.0607 0.0554 ...

$ radius\_se : num 0.236 0.451 0.197 0.338 0.178 ...

$ texture\_se : num 0.666 1.197 1.387 1.343 0.412 ...

$ perimeter\_se : num 1.67 3.43 1.34 1.85 1.34 ...

$ area\_se : num 17.4 27.1 13.5 26.3 17.7 ...

$ smoothness\_se : num 0.00805 0.00747 0.00516 0.01127 0.00501 ...

$ compactness\_se : num 0.0118 0.03581 0.00936 0.03498 0.01485 ...

$ concavity\_se : num 0.0168 0.0335 0.0106 0.0219 0.0155 ...

$ points\_se : num 0.01241 0.01365 0.00748 0.01965 0.00915 ...

$ symmetry\_se : num 0.0192 0.035 0.0172 0.0158 0.0165 ...

$ dimension\_se : num 0.00225 0.00332 0.0022 0.00344 0.00177 ...

$ radius\_worst : num 13.5 11.9 12.4 11.9 16.2 ...

$ texture\_worst : num 15.6 22.9 26.4 15.8 15.7 ...

$ perimeter\_worst : num 87 78.3 79.9 76.5 104.5 ...

$ area\_worst : num 549 425 471 434 819 ...

$ smoothness\_worst : num 0.139 0.121 0.137 0.137 0.113 ...

$ compactness\_worst: num 0.127 0.252 0.148 0.182 0.174 ...

$ concavity\_worst : num 0.1242 0.1916 0.1067 0.0867 0.1362 ...

$ points\_worst : num 0.0939 0.0793 0.0743 0.0861 0.0818 ...

$ symmetry\_worst : num 0.283 0.294 0.3 0.21 0.249 ...

$ dimension\_worst : num 0.0677 0.0759 0.0788 0.0678 0.0677 ...

>

> # drop the id feature

> wbcd <- wbcd[-1]

>

> # table of diagnosis

> table(wbcd$diagnosis)

B M

357 212

>

> # recode diagnosis as a factor

> wbcd$diagnosis <- factor(wbcd$diagnosis, levels = c("B", "M"),

+ labels = c("Benign", "Malignant"))

>

> # table or proportions with more informative labels

> round(prop.table(table(wbcd$diagnosis)) \* 100, digits = 1)

Benign Malignant

62.7 37.3

>

> # summarize three numeric features

> summary(wbcd[c("radius\_mean", "area\_mean", "smoothness\_mean")])

radius\_mean area\_mean smoothness\_mean

Min. : 6.981 Min. : 143.5 Min. :0.05263

1st Qu.:11.700 1st Qu.: 420.3 1st Qu.:0.08637

Median :13.370 Median : 551.1 Median :0.09587

Mean :14.127 Mean : 654.9 Mean :0.09636

3rd Qu.:15.780 3rd Qu.: 782.7 3rd Qu.:0.10530

Max. :28.110 Max. :2501.0 Max. :0.16340

>

> # create normalization function

> normalize <- function(x) {

+ return ((x - min(x)) / (max(x) - min(x)))

+ }

>

> # test normalization function - result should be identical

> normalize(c(1, 2, 3, 4, 5))

[1] 0.00 0.25 0.50 0.75 1.00

> normalize(c(10, 20, 30, 40, 50))

[1] 0.00 0.25 0.50 0.75 1.00

>

> # normalize the wbcd data

> wbcd\_n <- as.data.frame(lapply(wbcd[2:31], normalize))

>

> # confirm that normalization worked

> summary(wbcd\_n$area\_mean)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.1174 0.1729 0.2169 0.2711 1.0000

>

> # create training and test data

> wbcd\_train <- wbcd\_n[1:469, ]

> wbcd\_test <- wbcd\_n[470:569, ]

>

> # create labels for training and test data

>

> wbcd\_train\_labels <- wbcd[1:469, 1]

> wbcd\_test\_labels <- wbcd[470:569, 1]

>

> ## Step 3: Training a model on the data ----

>

> # load the "class" library

> library(class)

>

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test,

+ cl = wbcd\_train\_labels, k=21)

>

> ## Step 4: Evaluating model performance ----

>

> # load the "gmodels" library

> library(gmodels)

>

> # Create the cross tabulation of predicted vs. actual

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred,

+ prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 61 | 0 | 61 |

| 1.000 | 0.000 | 0.610 |

| 0.968 | 0.000 | |

| 0.610 | 0.000 | |

-----------------|-----------|-----------|-----------|

Malignant | 2 | 37 | 39 |

| 0.051 | 0.949 | 0.390 |

| 0.032 | 1.000 | |

| 0.020 | 0.370 | |

-----------------|-----------|-----------|-----------|

Column Total | 63 | 37 | 100 |

| 0.630 | 0.370 | |

-----------------|-----------|-----------|-----------|

>

> ## Step 5: Improving model performance ----

>

> # use the scale() function to z-score standardize a data frame

> wbcd\_z <- as.data.frame(scale(wbcd[-1]))

>

> # confirm that the transformation was applied correctly

> summary(wbcd\_z$area\_mean)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-1.4530 -0.6666 -0.2949 0.0000 0.3632 5.2460

>

> # create training and test datasets

> wbcd\_train <- wbcd\_z[1:469, ]

> wbcd\_test <- wbcd\_z[470:569, ]

>

> # re-classify test cases

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test,

+ cl = wbcd\_train\_labels, k=21)

>

> # Create the cross tabulation of predicted vs. actual

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred,

+ prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 61 | 0 | 61 |

| 1.000 | 0.000 | 0.610 |

| 0.924 | 0.000 | |

| 0.610 | 0.000 | |

-----------------|-----------|-----------|-----------|

Malignant | 5 | 34 | 39 |

| 0.128 | 0.872 | 0.390 |

| 0.076 | 1.000 | |

| 0.050 | 0.340 | |

-----------------|-----------|-----------|-----------|

Column Total | 66 | 34 | 100 |

| 0.660 | 0.340 | |

-----------------|-----------|-----------|-----------|

>

> # try several different values of k

> wbcd\_train <- wbcd\_n[1:469, ]

> wbcd\_test <- wbcd\_n[470:569, ]

>

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test, cl = wbcd\_train\_labels, k=1)

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 58 | 3 | 61 |

| 0.951 | 0.049 | 0.610 |

| 0.983 | 0.073 | |

| 0.580 | 0.030 | |

-----------------|-----------|-----------|-----------|

Malignant | 1 | 38 | 39 |

| 0.026 | 0.974 | 0.390 |

| 0.017 | 0.927 | |

| 0.010 | 0.380 | |

-----------------|-----------|-----------|-----------|

Column Total | 59 | 41 | 100 |

| 0.590 | 0.410 | |

-----------------|-----------|-----------|-----------|

>

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test, cl = wbcd\_train\_labels, k=5)

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 61 | 0 | 61 |

| 1.000 | 0.000 | 0.610 |

| 0.968 | 0.000 | |

| 0.610 | 0.000 | |

-----------------|-----------|-----------|-----------|

Malignant | 2 | 37 | 39 |

| 0.051 | 0.949 | 0.390 |

| 0.032 | 1.000 | |

| 0.020 | 0.370 | |

-----------------|-----------|-----------|-----------|

Column Total | 63 | 37 | 100 |

| 0.630 | 0.370 | |

-----------------|-----------|-----------|-----------|

>

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test, cl = wbcd\_train\_labels, k=11)

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 61 | 0 | 61 |

| 1.000 | 0.000 | 0.610 |

| 0.953 | 0.000 | |

| 0.610 | 0.000 | |

-----------------|-----------|-----------|-----------|

Malignant | 3 | 36 | 39 |

| 0.077 | 0.923 | 0.390 |

| 0.047 | 1.000 | |

| 0.030 | 0.360 | |

-----------------|-----------|-----------|-----------|

Column Total | 64 | 36 | 100 |

| 0.640 | 0.360 | |

-----------------|-----------|-----------|-----------|

>

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test, cl = wbcd\_train\_labels, k=15)

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 61 | 0 | 61 |

| 1.000 | 0.000 | 0.610 |

| 0.953 | 0.000 | |

| 0.610 | 0.000 | |

-----------------|-----------|-----------|-----------|

Malignant | 3 | 36 | 39 |

| 0.077 | 0.923 | 0.390 |

| 0.047 | 1.000 | |

| 0.030 | 0.360 | |

-----------------|-----------|-----------|-----------|

Column Total | 64 | 36 | 100 |

| 0.640 | 0.360 | |

-----------------|-----------|-----------|-----------|

>

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test, cl = wbcd\_train\_labels, k=21)

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 61 | 0 | 61 |

| 1.000 | 0.000 | 0.610 |

| 0.968 | 0.000 | |

| 0.610 | 0.000 | |

-----------------|-----------|-----------|-----------|

Malignant | 2 | 37 | 39 |

| 0.051 | 0.949 | 0.390 |

| 0.032 | 1.000 | |

| 0.020 | 0.370 | |

-----------------|-----------|-----------|-----------|

Column Total | 63 | 37 | 100 |

| 0.630 | 0.370 | |

-----------------|-----------|-----------|-----------|

>

> wbcd\_test\_pred <- knn(train = wbcd\_train, test = wbcd\_test, cl = wbcd\_train\_labels, k=27)

> CrossTable(x = wbcd\_test\_labels, y = wbcd\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_pred

wbcd\_test\_labels | Benign | Malignant | Row Total |

-----------------|-----------|-----------|-----------|

Benign | 61 | 0 | 61 |

| 1.000 | 0.000 | 0.610 |

| 0.938 | 0.000 | |

| 0.610 | 0.000 | |

-----------------|-----------|-----------|-----------|

Malignant | 4 | 35 | 39 |

| 0.103 | 0.897 | 0.390 |

| 0.062 | 1.000 | |

| 0.040 | 0.350 | |

-----------------|-----------|-----------|-----------|

Column Total | 65 | 35 | 100 |

| 0.650 | 0.350 | |

-----------------|-----------|-----------|-----------|

**Iris Data Set**

**Introduction**

*Step 1: Collecting the data*

Data was obtained from the University of California at Irvine Machine Learning Repository which was donated in 1988 on iris classifications. It contains features of the plants which help distinguish between the varieties. These are:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

Also included is the classification of the flower, Iris Setosa, Iris Versicolour  and Iris Virginica occur with 50 samples each.  This data was divided and 120 examples were used to train a nearest neighbors algorithm which was then applied to the remaining 30 records.

**Method**

*Step 2: Exploring and preparing the data*

To prepare the data, the class column was reclassified as a factor and several summary statistics were examined to ensure data sufficiency. Examining the proportion we see 33.3% of each class which is consistent with the data description from the source. Summary statistics for the four main descriptive elements of interest are shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **sepal length** | **sepal width** | **petal length** | **petal width** |
| Minimum | 4.3 | 2 | 1 | 0.1 |
| 1st Quartile | 5.1 | 2.8 | 1.6 | 0.3 |
| Median | 5.8 | 3 | 4.35 | 1.3 |
| Mean | 5.843 | 3.054 | 3.759 | 1.199 |
| 3rd Quartile | 6.4 | 3.3 | 5.1 | 1.8 |
| Maximum | 7.9 | 4.4 | 6.9 | 2.5 |

*Step 3: Training a model on the data*

This data set was divided and 469 examples were used to train a nearest neighbors algorithm which was then applied to the remaining 100 records.

Using R’s class package, the knn function was applied to the data.  This algorithm uses Euclidian distance to determine the k closest points to the test point.  Each “neighbor’s” assigned category, in this case benign or malignant, is given equivalent weight and the category having the most neighbors becomes the category of the test point.

Optimizing and refining the model involved the use of both z-score normalization and trials with a variety of k-values. The results of the normalization on this data for the area are shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **sepal length** | **sepal width** | **petal length** | **petal width** |
| Minimum | 0 | 0 | 0 | 0 |
| 1st Quartile | 0.2222 | 0.3333 | 0.1017 | 0.08333 |
| Median | 0.4167 | 0.4167 | 0.5678 | 0.5 |
| Mean | 0.4287 | 0.4392 | 0.4676 | 0.45778 |
| 3rd Quartile | 0.5833 | 0.5417 | 0.6949 | 0.70833 |
| Maximum | 1 | 1 | 1 | 1 |

**Results**

*Step 4: Evaluating model performance*

The table below shows the results of the trials with a variety of k values. As you can see, for values up to 30, the model is extremely accurate.

|  |  |  |  |
| --- | --- | --- | --- |
| **Raw Data** | | | |
| **k value** | **Correct** | **Incorrect** | **% Incorrectly classified** |
| 1 | 29 | 1 | 3% |
| 5 | 29 | 1 | 3% |
| 10 | 29 | 1 | 3% |
| 15 | 29 | 1 | 3% |
| 30 | 29 | 1 | 3% |
| 50 | 27 | 3 | 10% |

*Step 5: Improving model performance*

Attempts were made to further refine the model with z-score normalization, however, in this case the results became less accurate than they were with the raw data.  The chart below demonstrates this loss in accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Standardized Data** | | | |
| **k value** | **Correct** | **Incorrect** | **% Incorrectly classified** |
| 1 | 28 | 2 | 7% |
| 5 | 29 | 1 | 3% |
| 10 | 29 | 1 | 3% |
| 15 | 28 | 2 | 7% |
| 30 | 28 | 2 | 7% |
| 50 | 26 | 4 | 13% |

**Conclusion**

As a means of classifying irises, the k nearest neighbors algorithm is extremely accurate. While this accuracy was diminished with the introduction of standardization, it should be noted that the process did introduce some variability among the number of neighbors used which can be used to justify the selection of somewhere between 5 and 10 as the most appropriate number of neighbors to select for classification. Also notable and clearly visible in the attached appendix is the observation that the error in classification is most commonly found between versicolor and virginica which indicates that a classification of setosa can be viewed as being more dependable than that of the other two.

**Sources**

<http://archive.ics.uci.edu/ml/datasets/Iris>

Lantz, Brett. Machine Learning with R. 2013.

**Appendix 1: Iris Code**

> # import the CSV file

> iris <- read.csv("iris.csv", stringsAsFactors = FALSE)

>

> # examine the structure of the iris data frame

> str(iris)

'data.frame': 150 obs. of 5 variables:

$ sep\_len: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...

$ sep\_wid: num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...

$ pet\_len: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...

$ pet\_wid: num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...

$ class : chr "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...

>

> # table of classes

> table(iris$class)

Iris-setosa Iris-versicolor Iris-virginica

50 50 50

>

> # recode diagnosis as a factor

> iris$class <- factor(iris$class, levels = c("Iris-setosa", "Iris-versicolor", "Iris-virginica"),

+ labels = c("setosa", "versicolor", "virginica"))

>

> # table or proportions with more informative labels

> round(prop.table(table(iris$class)) \* 100, digits = 1)

setosa versicolor virginica

33.3 33.3 33.3

>

> # summarize three numeric features

> summary(iris[c("sep\_len", "sep\_wid","pet\_len", "pet\_wid")])

sep\_len sep\_wid pet\_len pet\_wid

Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100

1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300

Median :5.800 Median :3.000 Median :4.350 Median :1.300

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

>

> # create normalization function

> normalize <- function(x) {

+ return ((x - min(x)) / (max(x) - min(x)))}

>

> # test normalization function - result should be identical

> normalize(c(1, 2, 3, 4, 5))

[1] 0.00 0.25 0.50 0.75 1.00

> normalize(c(10, 20, 30, 40, 50))

[1] 0.00 0.25 0.50 0.75 1.00

>

> # normalize the iris data

> iris\_n <- as.data.frame(lapply(iris[1:4], normalize))

>

> # confirm that normalization worked

> summary(iris\_n$sep\_len)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.2222 0.4167 0.4287 0.5833 1.0000

>

> # create training and test data

> iris\_train <- iris\_n[-c(2:11,52:61,102:111), ]

> iris\_test <- iris\_n[c(2:11,52:61,102:111), ]

>

> # create labels for training and test data

>

> iris\_train\_labels <- iris[-c(2:11,52:61,102:111), 5]

> iris\_test\_labels <- iris[c(2:11,52:61,102:111), 5]

>

> ## Step 3: Training a model on the data ----

>

> # load the "class" library

> library(class)

>

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test,

+ cl = iris\_train\_labels, k=5)

>

> ## Step 4: Evaluating model performance ----

>

> # load the "gmodels" library

> library(gmodels)

>

> # Create the cross tabulation of predicted vs. actual

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred,

+ prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

> ## Step 5: Improving model performance ----

>

> # use the scale() function to z-score standardize a data frame

> iris\_z <- as.data.frame(scale(iris[-5]))

>

> # confirm that the transformation was applied correctly

> summary(iris\_z$sep\_len)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-1.86400 -0.89770 -0.05233 0.00000 0.67220 2.48400

>

> # create training and test datasets

> iris\_train <- iris\_z[-c(2:11,52:61,102:111), ]

> iris\_test <- iris\_z[c(2:11,52:61,102:111), ]

>

> # re-classify test cases

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test,

+ cl = iris\_train\_labels, k=5)

>

> # Create the cross tabulation of predicted vs. actual

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred,

+ prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

> # try several different values of k with the standardized data

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=1)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 9 | 1 | 10 |

| 0.000 | 0.900 | 0.100 | 0.333 |

| 0.000 | 0.900 | 0.100 | |

| 0.000 | 0.300 | 0.033 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.100 | 0.900 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 10 | 10 | 30 |

| 0.333 | 0.333 | 0.333 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=5)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=10)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=15)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 9 | 1 | 10 |

| 0.000 | 0.900 | 0.100 | 0.333 |

| 0.000 | 0.900 | 0.100 | |

| 0.000 | 0.300 | 0.033 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.100 | 0.900 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 10 | 10 | 30 |

| 0.333 | 0.333 | 0.333 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=30)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 9 | 1 | 10 |

| 0.000 | 0.900 | 0.100 | 0.333 |

| 0.000 | 0.900 | 0.100 | |

| 0.000 | 0.300 | 0.033 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.100 | 0.900 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 10 | 10 | 30 |

| 0.333 | 0.333 | 0.333 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=50)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 8 | 2 | 10 |

| 0.000 | 0.800 | 0.200 | 0.333 |

| 0.000 | 0.800 | 0.200 | |

| 0.000 | 0.267 | 0.067 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 2 | 8 | 10 |

| 0.000 | 0.200 | 0.800 | 0.333 |

| 0.000 | 0.200 | 0.800 | |

| 0.000 | 0.067 | 0.267 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 10 | 10 | 30 |

| 0.333 | 0.333 | 0.333 | |

-----------------|------------|------------|------------|------------|

> # try several different values of k with the non-standardized data

> iris\_train <- iris\_n[-c(2:11,52:61,102:111), ]

> iris\_test <- iris\_n[c(2:11,52:61,102:111), ]

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=1)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=5)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

>

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=10)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=15)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=30)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 10 | 0 | 10 |

| 0.000 | 1.000 | 0.000 | 0.333 |

| 0.000 | 0.909 | 0.000 | |

| 0.000 | 0.333 | 0.000 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 9 | 10 |

| 0.000 | 0.100 | 0.900 | 0.333 |

| 0.000 | 0.091 | 1.000 | |

| 0.000 | 0.033 | 0.300 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|

> iris\_test\_pred <- knn(train = iris\_train, test = iris\_test, cl = iris\_train\_labels, k=50)

> CrossTable(x = iris\_test\_labels, y = iris\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 30

| iris\_test\_pred

iris\_test\_labels | setosa | versicolor | virginica | Row Total |

-----------------|------------|------------|------------|------------|

setosa | 10 | 0 | 0 | 10 |

| 1.000 | 0.000 | 0.000 | 0.333 |

| 1.000 | 0.000 | 0.000 | |

| 0.333 | 0.000 | 0.000 | |

-----------------|------------|------------|------------|------------|

versicolor | 0 | 9 | 1 | 10 |

| 0.000 | 0.900 | 0.100 | 0.333 |

| 0.000 | 0.818 | 0.111 | |

| 0.000 | 0.300 | 0.033 | |

-----------------|------------|------------|------------|------------|

virginica | 0 | 2 | 8 | 10 |

| 0.000 | 0.200 | 0.800 | 0.333 |

| 0.000 | 0.182 | 0.889 | |

| 0.000 | 0.067 | 0.267 | |

-----------------|------------|------------|------------|------------|

Column Total | 10 | 11 | 9 | 30 |

| 0.333 | 0.367 | 0.300 | |

-----------------|------------|------------|------------|------------|