

# R Notebook: K Nearest Neighbors for Zoological Classification

Tristen Bristow

April 7, 2017

## Abstract

This study is concerned with categorizing wildlife by traits that could best inform a practical design of zoological display installations. The model investigated is a K-Nearest Neighbor subphylum-classifier, classifying individual animals by a typing that includes the existence or measure of various phenotypical traits. This study concerns the development and testing of a classification model that predicts mammal, bird, reptile, fish, amphibian, insect, or crustacean class of the input species. The KNN model is trained by learning which other animals/insects are of a similar subphylum type based on 16 observable phenotypical features. Each input indicates presence or magnitude of these features for a particular animal or insect.

## Load Data

```
set.seed(1)
library(class)
d = read.table("zoo.DATA", sep=",", header = FALSE)
d = data.frame(d)
```

## Data Conditioning

**Phylogenic traits used for classification:**

```
names(d) <- c("animal", "hair", "feathers", "eggs", "milk", "airborne",
"aquatic", "predator", "toothed", "backbone", "breathes", "venomous",
"fins", "legs", "tail", "domestic", "size", "type")
```

```
types <- table(d$type)
d_target <- d[, 18]
```

```
d_key <- d[, 1]
d$animal <- NULL
```

## Exploratory Investigation

Inspection of the occupancy levels of the classifications (in the merged data set), indicate the necessity for cross validation. Any singularly-induced train test split in the data is unlikely to provide an adequate balance of training examples for each class. From the summary output of the data, it would appear that a very low class occupancy exists for venomous animals (at 7%), which appears as the clearest example this concern. Output classes include,

```
names(types) <- c("mammal", "bird", "reptile", "fish", "amphibian",
"insect", "crustacean")
types
```

```
##      mammal      bird      reptile      fish amphibian      insect
##         41         20          5         13          4          8
## crustacean
##          10
```

```
summary(d)
```

```
##      hair      feathers      eggs      milk
## Min.   :0.0000 Min.   :0.000 Min.   :0.0000 Min.   :0.0000
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000
## Median :0.0000 Median :0.000 Median :1.0000 Median :0.0000
## Mean   :0.4257 Mean   :0.198 Mean   :0.5842 Mean   :0.4059
## 3rd Qu.:1.0000 3rd Qu.:0.000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max.   :1.0000 Max.   :1.000 Max.   :1.0000 Max.   :1.0000
## airborne aquatic predator toothed
## Min.   :0.0000 Min.   :0.0000 Min.   :0.0000 Min.   :0.000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000
## Median :0.0000 Median :0.0000 Median :1.0000 Median :1.000
## Mean   :0.2376 Mean   :0.3564 Mean   :0.5545 Mean   :0.604
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000
## Max.   :1.0000 Max.   :1.0000 Max.   :1.0000 Max.   :1.000
## backbone breathes venomous fins
## Min.   :0.0000 Min.   :0.0000 Min.   :0.00000 Min.   :0.0000
## 1st Qu.:1.0000 1st Qu.:1.0000 1st Qu.:0.00000 1st Qu.:0.0000
```

```
## Median :1.0000 Median :1.0000 Median :0.00000 Median :0.0000
## Mean :0.8218 Mean :0.7921 Mean :0.07921 Mean :0.1683
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:0.0000
## Max. :1.0000 Max. :1.0000 Max. :1.00000 Max. :1.0000
```

```
## legs tail domestic size
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000
## 1st Qu.:2.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
## Median :4.000 Median :1.0000 Median :0.0000 Median :0.0000
## Mean :2.842 Mean :0.7426 Mean :0.1287 Mean :0.4356
## 3rd Qu.:4.000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000
## Max. :8.000 Max. :1.0000 Max. :1.0000 Max. :1.0000
```

```
## type
## Min. :1.000
## 1st Qu.:1.000
## Median :2.000
## Mean :2.832
## 3rd Qu.:4.000
## Max. :7.000
```

```
str(d)
```

```
## 'data.frame': 101 obs. of 17 variables:
## $ hair : int 1 1 0 1 1 1 1 0 0 1 ...
## $ feathers: int 0 0 0 0 0 0 0 0 0 0 ...
## $ eggs : int 0 0 1 0 0 0 0 1 1 0 ...
## $ milk : int 1 1 0 1 1 1 1 0 0 1 ...
## $ airborne: int 0 0 0 0 0 0 0 0 0 0 ...
## $ aquatic : int 0 0 1 0 0 0 0 1 1 0 ...
## $ predator: int 1 0 1 1 1 0 0 0 1 0 ...
## $ toothed : int 1 1 1 1 1 1 1 1 1 1 ...
## $ backbone: int 1 1 1 1 1 1 1 1 1 1 ...
## $ breathes: int 1 1 0 1 1 1 1 0 0 1 ...
## $ venomous: int 0 0 0 0 0 0 0 0 0 0 ...
## $ fins : int 0 0 1 0 0 0 0 1 1 0 ...
## $ legs : int 4 4 0 4 4 4 4 0 0 4 ...
## $ tail : int 0 1 1 0 1 1 1 1 1 0 ...
## $ domestic: int 0 0 0 0 0 0 1 1 0 1 ...
## $ size : int 1 1 0 1 1 1 1 0 0 0 ...
## $ type : int 1 1 4 1 1 1 1 4 4 1 ...
```

## Training

The threshold neighbor-size ( $k$ ), for membership is set to the square root of the number of predictors, plus a constant that assigns it to the nearest odd number. A KNN Model is formed from the data using Leave One Out Cross Validation.

```
k = sqrt(17) + 1
m1 <- knn.cv(d, d_target, k, prob = TRUE)
prediction <- m1

cmat <- table(d_target, prediction)
acc <- (sum(diag(cmat)) / length(d_target)) * 100
print(acc)

## [1] 90.09901
```

## Confusion Matrix

`data.frame(types)`

```
##      Var1 Freq
## 1  mammal   41
## 2   bird   20
## 3 reptile    5
## 4   fish   13
## 5 amphibian  4
## 6   insect   8
## 7 crustacean 10
```

`cmat`

```
##      prediction
## d_target  1  2  3  4  5  6  7
##      1 41  0  0  0  0  0  0
##      2  0 20  0  0  0  0  0
##      3  0  1  0  3  1  0  0
##      4  0  0  0 13  0  0  0
##      5  0  0  0  0  4  0  0
##      6  0  0  0  0  0  8  0
##      7  0  0  0  2  1  2  5
```

## Accuracy (%)

`acc`

```
## [1] 90.09901
```

## **Discussion**

Classification accuracy of the LOOCV-trained knn model indicates that 90% accuracy is to be generally expected by applying this model to out-of-sample data. However, from an inspection of the confusion matrix output, it appears that there is generally a 0% accuracy for the identification of reptiles and crustaceans. It is unclear if the low-performing class outputs is due to there being a limited amount of data on hand, or if poor class separation of the two types is the cause. Overall, this model appears to be a robust predictor for all classes excluding reptiles and crustaceans. This classifier can be of pragmatic use for classifying various animal and insect into subphylum, providing a record of observable phylogenetic traits is available for each instance.

## **Conclusion**

It would appear that subphylum is a practical level of species categorization, and is an attainable training pattern for the K Nearest Neighbors algorithm.