# Building a Zoo



## Introduction

The Minnesota Zoo has just gotten a new animal called the *aurum tridecemlineatus*. Because they are math nerds, the zoo’s directors have decided to house the *aurum tridecemlineatus* near other animals that share similar characteristics. Today, your task is to help determine where *aurum tridecemlineatus*’ new home will be located by measuring its similarity to other Minnesota Zoo animals based on a given set of characteristics.

## Animal **Similarity**

Below is a table of characteristics for *aurum tridecemlineatus* and four other animals.

| **Animal** | **Hair** | **Toothed** | **Eggs** | **Legs** |
| --- | --- | --- | --- | --- |
| aurum tridecemlineatus | Yes | No | Yes | 4 |
| cavy | Yes | Yes | No | 4 |
| seahorse | No | Yes | No | 0 |
| tortoise | No | No | Yes | 4 |
| fruit bat | Yes | Yes | No | 2 |

1. Which animal seems most similar to *aurum tridecemlineatus*? Explain.
2. Which animal seems most different from the *aurum tridecemlineatus*? Explain
3. How did you determine which animals were similar and which was different?.

## Mixed Attributes and Gower Distance

One thing that complicates measuring similarity in this example is that some of the attributes are binary categorical attributes and some are quantitative. (We would call this a mixed set of attributes.) This makes it difficult to employ any of the similarity measures we have learned about so far.

One method of quantifying similarity when you have mixed attributes is *Gower’s distance.* The essential idea is that we compute the distance between measures on each attribute separately and then combine them by averaging those distances measures. The distance is computed differently depending on whether the attribute is categorical or quantitative. The rules are:

**Categorical Attribute:**

* If the values on the attribute are different, the distance is 1.
* If the values on the attribute are the same, the distance is 0.

**Quantitative Attribute:**

* The distance is the absolute value of the difference between values divided by the range of the attribute. Mathematically,

Where Xi is the value on the quantitative attribute *X* for case *i*, Xj is the value on the quantitative attribute *X* for case *j*, XMax is the maximum value on the quantitative attribute *X* in the data (across all cases), and XMin is the minimum value on the quantitative attribute *X* in the data.

Once we have the distance for each attribute, we average those values to compute Gower’s distance.

Where *di* is the distance measure for the *i*th attribute, and *k* is the total number of attributes.

### Math Extension: Normalizing

In the formula for computing the distance between quantitative measures, the numerator gives us the distance between the two cases’ measures. So why do we then divide that by the range? Dividing by the range is called *normalizing* the distance. By normalizing the distance, we are re-scaling that distance to lie between 0 and 1. This puts the distance between quantitative measures on the same scale as the distance values for the categorical attributes. This is important so that when we find the average, of all the distance measures the quantitative distances don’t unduly influence the average.

### Example

Say we want to compute Gower’s distance between *aurum tridecemlineatus* and the cavy.

| **Animal** | **Hair** | **Toothed** | **Eggs** | **Legs** |
| --- | --- | --- | --- | --- |
| aurum tridecemlineatus | Yes | No | Yes | 4 |
| cavy | Yes | Yes | No | 4 |
| seahorse | No | Yes | No | 0 |
| tortoise | No | No | Yes | 4 |
| fruit bat | Yes | Yes | No | 2 |

We first determine the distance between *aurum tridecemlineatus* and the cavy for each attribute:

* **Hair:** d = 0 (they are both “Yes”)
* **Toothed:** d = 1 (they are not the same; One is “Yes”, the other is “No”)
* **Eggs:** d = 1 (they are not the same; One is “Yes”, the other is “No”)
* **Legs:** d = = 0

Next, we compute the average of those four distances:

1. Compute the Gower distance between aurum tridecemlineatus and each of the other three animals given in the table. Show your work.
2. Based on the Gower distance, which animal is *aurum tridecemlineatus* most similar to? Explain?
3. Based on the Gower distance, which animal is *aurum tridecemlineatus* least similar to? Explain?

## Binary Attributes: Dummy Coding

When we have categorical binary attributes, we often re-code the values into the numbers 0 and 1. This is called dummy coding. In our example, we could re-code so that:

* “Yes” = 1, and
* “No” = 0

Our new data would then look like the following:

| **Animal** | **Hair** | **Toothed** | **Eggs** | **Legs** |
| --- | --- | --- | --- | --- |
| aurum tridecemlineatus | 1 | 0 | 1 | 4 |
| cavy | 1 | 1 | 0 | 4 |
| seahorse | 0 | 1 | 0 | 0 |
| tortoise | 0 | 0 | 1 | 4 |
| fruit bat | 1 | 1 | 0 | 2 |

Note that we only dummy code categorical binary attributes; we do not change any quantitative attributes.

The advantage of doing this is that now we can compute the distance between animals for each categorical attribute in the same way we do for the quantitative attributes:

Computing the distances for each attribute between *aurum tridecemlineatus* and the cavy we find:

* **Hair:**
* **Toothed:**
* **Eggs:**
* **Legs:** d = = 0

Next, we compute the average of those four distances:

Notice that the range (in the denominator) for our categorical binary attributes will always be 1–0=1. So, we could ignore that and just use:

1. What if we had coded “Yes” = 0 and “No” = 1? Use that coding to recode both *aurum tridecemlineatus* and the cavy. Then compute Gower’s distance based on this new coding scheme.

| **Animal** | **Hair** | **Toothed** | **Eggs** | **Legs** |
| --- | --- | --- | --- | --- |
| aurum tridecemlineatus |  |  |  |  |
| cavy |  |  |  |  |

## 

## 

1. Does it matter which value you code as 0 and which you code as 1? Explain.

## Gower Distance Using Technology

The data in the [zoo-data Google Sheet](https://docs.google.com/spreadsheets/d/1PIvpz0nB_BUE-82hcdPOlTR5uSRCt2cGqvmFpjDoyCk/edit#gid=0) includes 16 characteristics for the 98 different animals that reside at the Zoo. The characteristics for *aurum tridecemlineatus* are given below.

| Hair | Yes |  | Backbone | Yes |
| --- | --- | --- | --- | --- |
| Feathers | No |  | Breathes | Yes |
| Eggs | Yes |  | Venomous | No |
| Milk | Yes |  | Fins | No |
| Airborn | No |  | Tail | Yes |
| Aquatic | Yes |  | Domestic | No |
| Predator | Yes |  | Catsize | Yes |
| Toothed | No |  | Legs | 4 |

## 

1. Dummy code all of the categorical binary attributes in the zoo-data sheet. Also dummy code the characteristics for *aurum tridecemlineatus* using the same coding scheme.
2. Create a new column in the Google sheet called “Gower’s Distance”. Fill in this column by using a formula to compute Gower’s distance between each animal and *aurum tridecemlineatus*.

## Classification Using kNN

Now that you have the distance measures, we can use the K-NN algorithm to determine which area of the zoo *aurum tridecemlineatus* should be housed in.

1. Compute and report the optimal value for *k* to determine how many nearest neighbors we should base our classification on. Show your work.
2. Use the optimal number of nearest neighbors (that you reported in the previous question) to determine the zoo area that *aurum tridecemlineatus* should be housed in. Also report the “vote” tally for each class (zoo area).

## Using kNN when the Classes are Imbalanced

There is one issue with using the kNN algorithm as we have been for these data…namely that the class sizes are not the same (or even similar). For example, there are 39 animals that live in the Mammal Mansion, while only three animals live in the Amphibian Abode. When the classes are greatly imbalanced it becomes almost impossible for the kNN algorithm to choose a small class.

To offset this, we weight the votes in our *k* nearest neighbors by multiplying the number of votes by the inverse of the class frequency. For example, suppose we had two potential classes with 40 cases that are from Class A and 10 cases from Class B. Further suppose that we were basing our classification on the 7 nearest neighbors and that those neighbors were:

{A, B, A, A, B, A, A}

Typically our classification would be Class A since the vote tally is 5 (Class A) to 2 (Class B). However, because the classes are imbalance we need to update the votes as:

**Class A:**

**Class B:**

Based on these updated tallies, we would classify the unknown case as Class B.

1. Use the vote tallies you reported in Question 12, and update them to account for the imbalance in the zoo data. Show your work.
2. Which zoo area should *aurum tridecemlineatus* be housed in?