Classification and Clustering

Machine Learning, or Statistical Learning, is the ability of a computer to improve its own performance through the use of software that employs artificial intelligence techniques to mimic the ways by which humans seem to learn. It focuses on the development of computer programs that can teach themselves to change and behave in some improved way when exposed to new data.

The ‘learning’ is actually achieved by looking for and detecting patterns in the data and adjusting the program’s actions accordingly, similar to the way in which humans try to improve their understanding through observation, study, and repetition. While artificial intelligence and machine learning are thought of as sciences, each is also considered an art, in that creativity is required to achieve the goals.

We have talked about association rules and linear regression, which are both techniques of machine learning. **Classification** and **Clustering** are two others, and we will introduce them in concept:

**Classification**

Classification attempts to use an algorithm to group objects into classes, or groups, when you have some ‘practice data’ with groups that are known. For example, animals may be classified into species; symptoms may be classified into health conditions. Here are the steps:

* The features of members of each class are first analyzed or learned,
* They are then generalized to build an understanding of what it means to be a member of each class.
* Finally, the generalizations are compared with potential new members and used to ‘classify’ each into the best fitting class.

Classification is considered **supervised learning**, because we have some information on the class members at the beginning. The goal is to create an algorithm that can predict what group future examples fall into, e.g., if a person has certain symptoms, what disease do we predict they have?

**Clustering**

Clustering is similar to Classification, except that the groups are **not** known beforehand. Groupings are invented by looking at individuals, and grouping the ones that seem to have shared characteristics.

* Given a set of potential class members, consider many of their characteristics,
* Look for two or more clusters of similar members, and evaluate how ‘good’ the grouping is,
* Try another grouping and see if it seems better,
* Repeat (‘iterate’) until the grouping seems good enough.

Clustering is an example of **unsupervised learning**, because the classes are unknown at the start. The goal is to find clusters of like individuals. For example, a marketer might want to segment the population into general types of consumers, and market to each group in a specific way. A tricky aspect of clustering is that, once the groups are created, it is up to the person using the algorithm to characterize or describe the groups.

**Questions:**

1. **Classification:** Attached, find a picture of many different dogs and cats. Classify each picture as a cat or a dog. How many cats did you come up with? How many dogs? Was this easy to do?

13 dogs 13 cats

1. **Classification:** What if you did not know what a cat or dog was? (For example, if you had lived in isolation and had not seen a cat or dog ‘in person’ or on television, video, etc.) Could you have done this task? Suppose you were going to write instructions for a person who does not know what a cat or a dog is. Try writing down a list of criteria that will help them. Are your criteria perfect – are there any exceptions to your ‘rules’?

Cats have pointy ears

Cats have smaller, triangle noses

Cats have longer whiskers

Cats Pupils more visible

Cats flatter faces

Cats have curved backs

Dogs are bigger

Dogs more often have tongue hanging out

Dogs have more visible musculature.

Dogs more likely to have a collar

Dogs eyes are wider set

Usually cats have pointier ears.

1. **Clustering:** Using the same images, forget about ‘cats’ and ‘dogs’ for a moment.. Create categories that put the images into **two** groups based on some criteria – (but **not** the groups of ‘cat’ and ‘dog.’) Try doing this at least 2 or 3 different ways, writing down the criteria for each grouping.

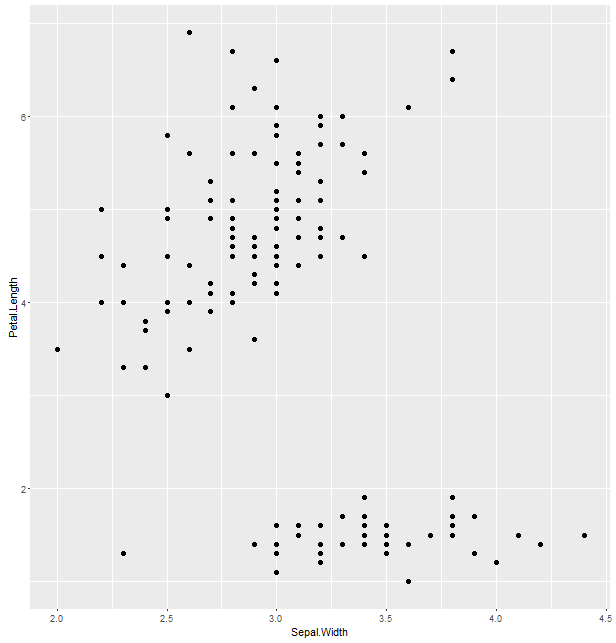
**Background vs no background**

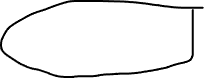
**Sitting vs standing**

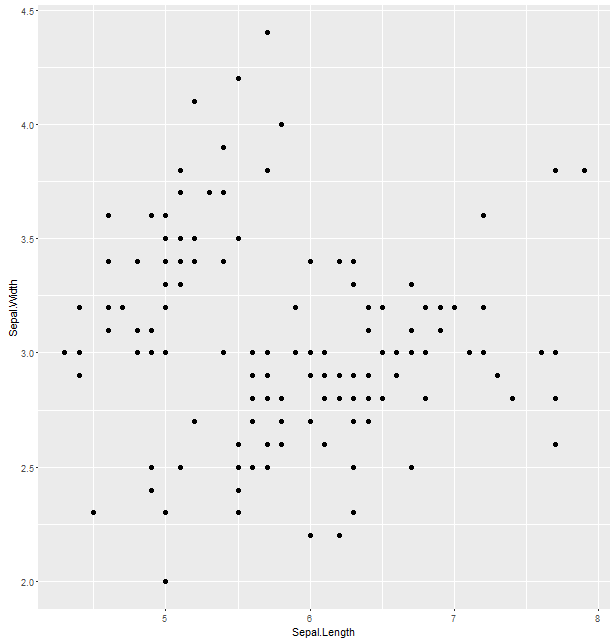
1. **Clustering:** Repeat what you just did, but creating categories that put the images into **three** groups this time.

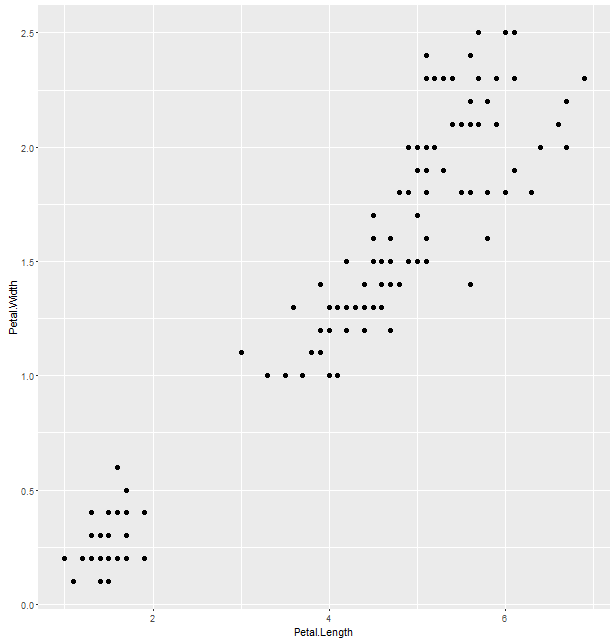
**Amused, Unamused, Can’t tell**

1. Which did you find to be easier, Clustering or Classification? Why? Explain.
2. Now, consider another example. Measurements were taken for three different species of iris. The measurements are: Petal Length, Petal Width, Sepal Length and Sepal Width. Suppose you want to group the irises into three species but you do not have any previous data on species – all you know is that you want to form three groups. Is this a problem of Classification or Clustering? Explain.
3. Here are three plots that could be used to form three groups of irises. Each plot includes measurements of irises on its axes, and each plot has dots from all three groups. Think about where the groups might lie on each of the following plots:











1. Which of the three plots gives you the most information to divide the irises into three (equal-sized ) groups? Which gives the least? Explain why.
2. Using one or two of the plots, suggest where you believe the groups lie. That is, circle or highlight the dots belonging to each group. (If you can’t draw on the graph, describe where you’d put your divisions.)
3. Now, to check your work, get the actual data in R. It is a stored data set, so just type (or copy/paste) in R:

**library(tidyverse)**

**names(iris) # to see what is in the data set**

**ggplot(data = iris,**

**mapping = aes(x = Sepal.Width, y = Petal.Length,**

**color=Species)) +**

**geom\_point()**

Repeat for the other two plots (Petal.Width by Petal.Length, then Sepal.Width by Sepal. Length). How did you do with your guesses above?

There are many methods for deciding quantitatively how the groups should be chosen. Take a look at Wikipedia to browse some of the methods. (Google cluster analysis, Wikipedia). Scroll down to some of the graphical illustrations. Below, we’ll use the Ward Hierarchical Clustering method on some other data. More details here on the method: <http://www.statmethods.net/advstats/cluster.html>

1. **Clustering:** Try it, using R. View the data set **mtcars**, a data frame that is stored in R.

We will try out an algorithm called Hierarchical Clustering, using R. We’ll run the code in the Rmd script.

1. Based on what you know about cars, do these groups make sense? Try changing the code to create 3 groups, or some other number. (you only need to rerun the plot, groups, and rect.hclust lines of code)
2. Google these two articles. Determine if they are using techniques of Classification or Clustering. Do you think these represent good use of the machine learning techniques?
3. *“People with depression are more likely to say certain words,”* by Mohammed Al-Mosaiwi, February 6, 2018
4. *The Hidden Tribes of America*, on hiddentribes.us, a project begun in late 2018 by More in Common.

14. **Classification:** We’ll try a classification algorithm called Decision Rules. We’ll use a different set of car data, so you’ll download the data from the DATA link on Blackboard.

This data is in two parts: Training data and Test data. It is very common in machine learning to divide the data set this way, with most of the cases in the Training data. A model is built using the Training data, then tested on the new, Test data. We’ll run the code in the Rmd script.