Checkpoint 1

# Why do we do this?

In the beginning of this program, we divided machine learning into two overarching types. We've now thoroughly covered the first class, supervised learning, and so we are ready to move on to the second: unsupervised learning.

This module provides a brief overview of machine learning. This is going to be a programming free module, so pull up a chair and get ready to read. You should come out of this with a solid grasp on the kinds of problems we'll tackle with this class of models and prepared to dive into some actual code in subsequent modules.

## Why do we do this?

To use a supervised technique, you have to have a lot of information about your process and understand some aspects of it pretty well. You have to have some kind of outcome you're interested in, specifically one you can observe and record.

But that isn't always the case, and it isn't always what you're interested in. Enter: unsupervised learning.

Outline

# What is Unsupervised Learning?

So, you have no outcome variable. Maybe you can't actually observe what you're interested in. Maybe it just isn't that kind of problem. But there's no outcome variable. Then you don't really have a training set the way we've thought about them before. Time to pack up and go home, right?

Wrong.

Unsupervised learning looks through the data you do have, a series of independent variables that make up your observed (or observable) data, and allows you to do something with it or say something about it. What can you do or say? Many different things. It depends on what kind of unsupervised model you want to build.

# Clustering

Clustering models are probably the simplest to logically grasp. Let's say you have a bunch of variables that you've observed as part of a process that you're studying. A clustering algorithm will go over that set of variables and say, "I think these are groups of related observations."

That can happen in two ways. Typically you tell it how many groups, or clusters, the algorithm is allowed to generate, and the algorithm goes over the data creating the groups that minimize some cost function. More rarely the algorithm can find the number of groups on its own.

# Variable Importance

You've actually already worked on variable importance in this course. PCA is an unsupervised technique for finding out which variables tend to be most influential in your dataset. There are plenty of other techniques in this space, but we'll leave you with PCA for this course.

# Neural Networks

Lastly come neural networks. Neural networks get their own module later in this course, but they can function as either a supervised or unsupervised technique. Unsupervised neural networks have become an important technique, particularly in things like image recognition where it can be effective in discovering **latent variables** which are impactful but not explicitly defined in the initial dataset.

Outline

This brief overview of techniques gives a sense of what unsupervised techniques can accomplish, but we'll dive in deeper as we introduce specific techniques throughout the unit. There are many topics we won't cover in this course. You can refer to this text for a brief introduction to other techniques and this one for a more in depth and technical survey. Unsupervised learning is a deep and evolving field of which we're only barely able to scratch the surface.

Mark as read

Report a typo or other issue

Next checkpoint

Go to Overview

Go to Overview

Outline