Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Clustering

Why do clustering?

In chapter 4, we discussed **classification**: sorting data points into different groups, or **classes**.

Example: We might sort animals by **genetics**, or different sub-diseases that need different **treatments**.

Simplifying our data into **categories** can allow us to do better work, more easily.

This had lots of benefits:

- It could be used to make **decisions**. For example, a binary classifier could be used to decide "yes" or "no".
- We could use this to understand the item It could be used to make yes or no decisions. and **distribution** of our data.
- We could sort different types of data to be processed **separately**.

The problem is, this relied on us **knowing** what classes we plan to sort into.

This may seem obvious, but what if we're looking at something **new**? A disease we don't fully **understand**, or animals we've never **seen** before? How do we classify them?

In the past, doing this ourselves has given rise to many of the **classes** we use to sort things today. But, computers allow us to do this in situations we **never** could before:

- **High-dimensional** datasets, with too much **complex** information for a human to make sense of.
- Discovering new classes **faster** than ever using computers.

- Finding **patterns** in creative ways humans would never think to, especially for really **abstract** problems.

Concept 1

Clustering is like **classification**, where we want to assign things to **classes**: we call them **clusters**.

But, we use it when we **don't know** what groupings we want, so we have to **find** them.

We have some challenges ahead of us, though. Not only do we need to create **new classes**, we *still* need to classify our points based on them!

Clustering Formalisms

Unsupervised Learning

The first thing we should note:

This problem is similar to classification, a **supervised** problem.

It was **supervised** because we knew the **correct** labels for our data in our advance. We just wanted to **teach** it to our computer.

The problem here: we **don't** know the correct labels! In fact, we're making them up as we go. Because we aren't being "supervised" by a correct answer, we call this **unsupervised learning**.

Concept 2

Clustering is a type of **unsupervised learning**: meaning, we don't have a **correct** answer in advance.

The labels we create are not based on a **known** truth.

The **label** for data point $x^{(i)}$ is written as $y^{(i)}$.

What is clustering?

So, if we don't know **what** our classes are, how do we figure out **which** classes to create?

Intuitively, we think of classes as a **collection** of things that are **similar** to each other. Before, we've considered things "**close**" if they have a **low distance** in input space.

Remember that input space is where we represent each data point using input variables.

Example: We can tell two animals are **both** cats because they both have fur and sharp claws, among other things: they're **similar**.

Meanwhile, two points in different classes are more **different**: they're further apart in input space.

Example: We can tell a dolphin is **distinct** from a cat because one lives in water and one doesn't: their clusters are more **different**.

We call each of these groupings **clusters**.

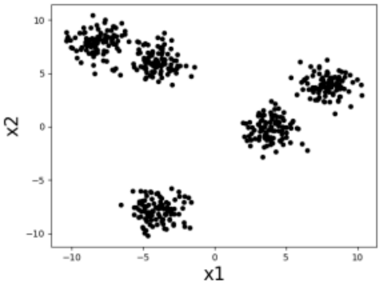
Definition 3

Informally, a **cluster** is a collection of **data points** that are all

- **Near** each other
- **Far** from the other clusters.

We use clusters as our way to **discover** new classifications.

Example: Below, we can visually mark out what looks like 5 distinct **clusters** in input space $(x_1, x_2) \in \mathbb{R}^2$:



This is an informal way to understand clusters, though. If we want to be more precise, we need to ask ourselves questions like:

- What does it mean for points to be "close" or "far"? How are we measuring distance?
- How many clusters do we want?
- How do we evaluate our clustering?