Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Independent and Identically Distributed

Let's look at our underlying assumptions: the rest of this class relies on these assumptions.

1. An assumption about data

Let's return back to our original goal: we want to use **data** to teach our machine to give us **results** we want. Just like how a person might learn from their **experience** and use it to make **judgments**.

However, there's an **assumption** built in to this statement, one we need to look at more closely: we are assuming that **past** data allows us to predict **future** data.

This may seem obvious, but it isn't always: past data may not be **representative** of the future, for example.

- **Example:** We can't use the weather over the month of July to predict the weather in the month of December.

This often called the problem of **induction**: using the past to predict the future.

2. Is our data representative?

First, let's solve the problem presented above:

- **Example:** We got our weather from a **different** month than we're trying to predict.

So, it seems our problem is that our **data** and what we're trying to **predict** are from **two different sources**.

We want them to come from the **same source**, then. In this case, we could say we want them to be from the **same** month. Great. But how do we say this in general?

3. How do we compare data?

We got down to the real problem: we want our new data to be from a similar source to the old data. One month couldn't **represent** another, because they **behave** differently.

- **Example:** For different months, we get different rainy days, different temperature ranges, so on: they can't be compared.

In general, we need a way to describe what we mean by "different": what describes one of these months?

- **Example:** To us, all that matters is the weather: how **likely** are we have a rainy day, for example? In fact, we'd like to know how **likely** every outcome is.

We represent this with something called a **distribution**. A distribution gives us exactly what we just described: **how likely** different events are to occur.

This is how our system "behaves", in a way.

Definition 1

A **distribution** is a **function** that gives us the **probability** of different **outcomes**.

Why is it called a distribution? Well, we're taking the **odds**, and spreading them out (or **distributing** them) over multiple different outcomes!

Example: The **distribution** of outcomes on a coin is 50% chance of heads, 50% chance of tails.

Notice that distributions are **probabilistic**: outcomes have a certain **chance** of occurring. Otherwise, these problems would be simple.

4. Identically Distributed Data

We can think of this distribution as a **simplified** view of the **source** of our data. Each "outcome" is a data point; one we can use to **learn**.

We want our **past** data we **learn** from, and our **future** data with **test** with, to have the **same** distribution.

We also want different points in the **same** dataset (past *or* future) to be from the same **distribution**: if they aren't, then why are we lumping them together?

We want them to be the "same", or **identical**: they have **exactly** the same chances for each outcome.

We want to focus on one problem at a time - one distribution.

In other words: we want our sets of data to be **identically distributed**.

Definition 2

If two **data points**(or datasets) are **identically distributed**, then they have the **same** underlying **distributions**.

In other words, they have the **same probabilities** for each possible **outcome**.

Example: Two fair coins will behave the same as each other: they both have 50-50 odds. Thus, they're **identically distributed**.

5. Independence (Review)

There's a second assumption that is just as important: when we draw two different data points, we are also **assuming** that the results of one do not **affect** the other.

If one point **depended** on another, then there's no **new** information: you could have used the last point to guess this one.

This means you're **not learning**, which is a problem: you need many experiences to come to a good **conclusion**, that will apply well in the future.

Because we don't want the result of one data point to **depend** on another, we call this assumption **independence**.

Definition 3

Two **data points** are **independent** if **knowledge** of the outcome for one data point does not affect the **probabilities** for the other.

This definition is a bit informal: the proper definition is to say that, for two events A and B, $P(A)P(B) = P(A \text{ and } B)$

Example: If you flip two coins, knowing that one coins comes up heads does not tell you anything about the other coin: the two coin tosses are **independent**.

6. Independent and Identically Distributed

We combine both of these assumptions into our final result: we want our data points and data sets to be both **independent and identically distributed.**

Definition 4

IID, or **Independent and Identically Distributed**, means that if you draw two data points, they

- Come from the **same distribution**: they have the same **probabilities** for each outcome,
- They **aren't related** in any other way: they are **independent**, meaning the **outcome** of one **does not** affect the other.

Example: Based on the two examples above, flipping two coins (or rolling a die twice) is IID.

We shorten this to one acronym, which tells you how important it is: it is the base assumption in many different statistics, inference, and machine learning settings.

We will assume this to be true, and use that assumption throughout the class. We expect our data to be IID in most cases.