Introduction to Data Science: Principles

Héctor Corrada Bravo

University of Maryland, College Park, USA

2020-01-28

Measurements and Data Types

A data analysis to get us going

Analysis of Baltimore crime data.

Downloaded from Baltimore City's awesome open data site (this was downloaded a couple of years ago so if you download now, you will get different results).

The repository for this particular data is here.
https://data.baltimorecity.gov/Crime/BPD-Arrests/3i3v-ibrt

Getting data

We've prepared the data previously into a comma-separated value file (`.csv` file):

- each column defines attributes that describe arrests
- each line contains attribute values (separated by commas) describing specific arrests.

Getting data

Note: To download this dataset to follow along you can use the following code:

```
if (!dir.exists("data")) dir.create("data")

download.file("https://www.hcbravo.org/IntroDataSci/misc/BPD_Arrests.csv", destfile="data
```

Getting data

To make use of this dataset we want to assign the result of calling `read_csv` (i.e., the dataset) to a variable:

```
library(tidyverse)

arrest_tab <- read_csv("data/BPD_Arrests.csv")

arrest_tab
```

```
## # A tibble: 104,528 x 15

##    arrest   age race  sex   arrestDate arrestTime arrestLocation

##     <dbl> <dbl> <chr> <chr> <chr>      <time>     <chr>

##  1 1.11e7    23 B     M     01/01/2011 00'00"     <NA>

##  2 1.11e7    37 B     M     01/01/2011 01'00"     2000 Wilkens …
```

Getting data

Now we can ask what *type* of value is stored in the `arrest_tab` variable:

```
class(arrest_tab)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

Getting data

The `data.frame` is a workhorse data structure in R. It encapsulates the idea of *entities* (in rows) and *attribute values* (in columns). We call these *rectangular datasets*. The other types `tbl_df` and `tbl` are added by `tidyverse` for improved functionality.

Getting data

The `data.frame` is a workhorse data structure in R. It encapsulates the idea of *entities* (in rows) and *attribute values* (in columns). We call these *rectangular datasets*. The other types `tbl_df` and `tbl` are added by `tidyverse` for improved functionality.

Later, we will see how the `pandas` Python package provides the same semantics.

Getting data

We can ask other features of this dataset:

```
# This is a comment in R, by the way

# How many rows (entities) does this dataset contain?
nrow(arrest_tab)
```

```
## [1] 104528
```

```
# How many columns (attributes)?
ncol(arrest_tab)
```

```
## [1] 15
```

Getting data

Now, in Rstudio you can view the data frame using `View(arrest_tab)`.

Entities and attributes

We use the term *entities* to refer to the objects represented in a dataset refers to.

In our example dataset each arrest is an *entity*.

Entities and attributes

We use the term *entities* to refer to the objects represented in a dataset refers to.

In our example dataset each arrest is an *entity*.

In a rectangular dataset (a data frame) this corresponds to rows in a table.

Entities and attributes

A dataset contains *attributes* for each entity.

Attributes of each arrest would be:

the person's *age*, the type of offense, the location, etc.

Entities and attributes

A dataset contains *attributes* for each entity.

Attributes of each arrest would be:

the person's *age*, the type of offense, the location, etc.

In a rectangular dataset, this corresponds to the columns in a table.

Entities and attributes

This language of *entities* and *attributes* is commonly used in the database literature.

In statistics you may see *experimental units* or *samples* for *entities* and *covariates* for *attributes*.

In other instances *observations* for *entities* and *variables* for *attributes*.

In Machine Learning you may see *example* for *entities* and *features* for *attributes*.

For the most part, all of these are exchangable.

Entities and attributes

This table summarizes the terminology:

Field	Entities	Attributes
Databases	Entities	Attributes
Machine Learning	Examples	Features
Statistics	Observations/Samples	Variables/Covariates

Categorical attributes

A categorical attribute for a given entity can take only one of a finite set of examples.

For example, the `sex` variable can only have value `M`, `F`, or (we'll talk about missing data later in the semester).

Categorical attributes

The result of a coin flip is categorical

The outcome of rolling an 8-sided die, is also categorical

Can you think of other examples?

Categorical attributes

Categorical data may be *ordered* or *unordered*

In our example, all categorical data is unordered.

Categorical attributes

Categorical data may be *ordered* or *unordered*

In our example, all categorical data is unordered.

Examples of *ordered categorical data* are grades in a class, Likert scale categories, e.g., `strongly agree`, `agree`, `neutral`, `disagree`, `strongly disagree`, etc

Discrete numeric attributes

These are attributes that can take specific values from elements of ordered, discrete (possibly infinite) sets. The most common set in this case would be the non-negative positive integers.

Discrete numeric attributes

These are attributes that can take specific values from elements of ordered, discrete (possibly infinite) sets. The most common set in this case would be the non-negative positive integers.

This data is commonly the result of counting processes. In our example dataset, age, measured in years, is a discrete attribute.

Discrete numeric attributes

Frequently, we obtain datasets as the result of summarizing, or aggregating other underlying data.

In our case, we could construct a new dataset containing the number of arrests per neighborhood (we will see how to do this later)

Discrete numeric attributes

```
## # A tibble: 6 x 2
##   neighborhood      number_of_arrests
##   <chr>                         <int>
## 1 Abell                            62
## 2 Allendale                       297
## 3 Arcadia                          78
## 4 Arlington                       694
## 5 Armistead Gardens               153
## 6 Ashburton                        78
```

Discrete Numeric Attributes

In this new dataset, the *entities* are each neighborhood, the `number_of_arrests` attribute is a *discrete numeric* attribute.

Discrete Numeric Attributes

Other examples:

- the number of students in a class is discrete,
- the number of friends for a specific Facebook user.

Can you think of other examples?

Discrete Numeric Attributes

Distinctions between ordered categorical and discrete numerical data:

ordered categorical data do not have magnitude

Discrete Numeric Attributes

For instance, is an 'A' in a class twice as good as a 'C'?

Is a 'C' twice as good as a 'D'?

Discrete Numeric Attributes

For instance, is an 'A' in a class twice as good as a 'C'?

Is a 'C' twice as good as a 'D'?

Not necessarily.

Grades don't have an inherent magnitude.

Discrete Numeric Attributes

However, if we *encode* grades as 'F=0,D=1,C=2,B=3,A=4', etc. they do have magnitude.

In that case, an 'A' *is* twice as good as a 'C', and a 'C' *is* twice as good as a 'D'.

Discrete Numeric Attributes

In summary,

if ordered data has magnitude, then *discrete numeric*

if not, *ordered categorical*.

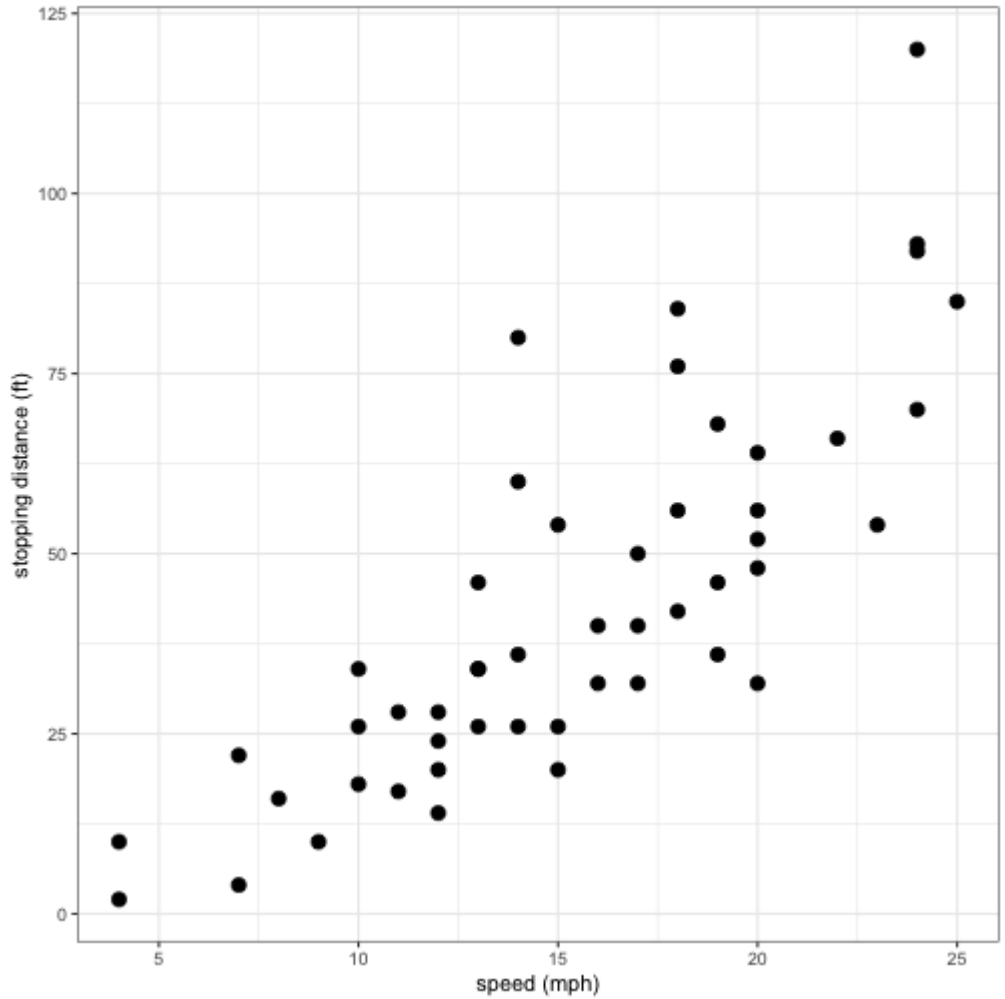
Continuous numeric data

Attributes that can take any value in a continuous set.

For example, a person's height, in say inches, can take any number (within the range of human heights).

Continuous numeric data

Different dataset: entities are cars and we look at continuous numeric attributes `speed` and `stopping distance`



Continuous Numeric Attributes

The distinction between *continuous* and *discrete* can be tricky:

measurements that have finite precision are, in a sense, discrete.

Continuous Numeric Attributes

The distinction between *continuous* and *discrete* can be tricky:

measurements that have finite precision are, in a sense, discrete.

Remember, continuity is *not* a property of the specific dataset you have in hand,

It *is* a property of the process you are measuring.

Continuous Numeric Attributes

The number of arrests in a neighborhood cannot be fractional, regardless of the precision at which we measure this.

Continuous Numeric Attributes

The number of arrests in a neighborhood cannot be fractional, regardless of the precision at which we measure this.

On the other hand, if we had the appropriate tool, we could measure a person's height with infinite precision.

Continuous Numeric Attributes

This distinction is very important when we build statistical models of datasets for analysis.

For now, think of discrete data as the result of counting, and continuous data the result of some physical measurement.

Continuous Numeric Attributes

This distinction is very important when we build statistical models of datasets for analysis.

For now, think of discrete data as the result of counting, and continuous data the result of some physical measurement.

Here's a question: is `age` in our dataset a continuous or discrete numeric value?

Other examples

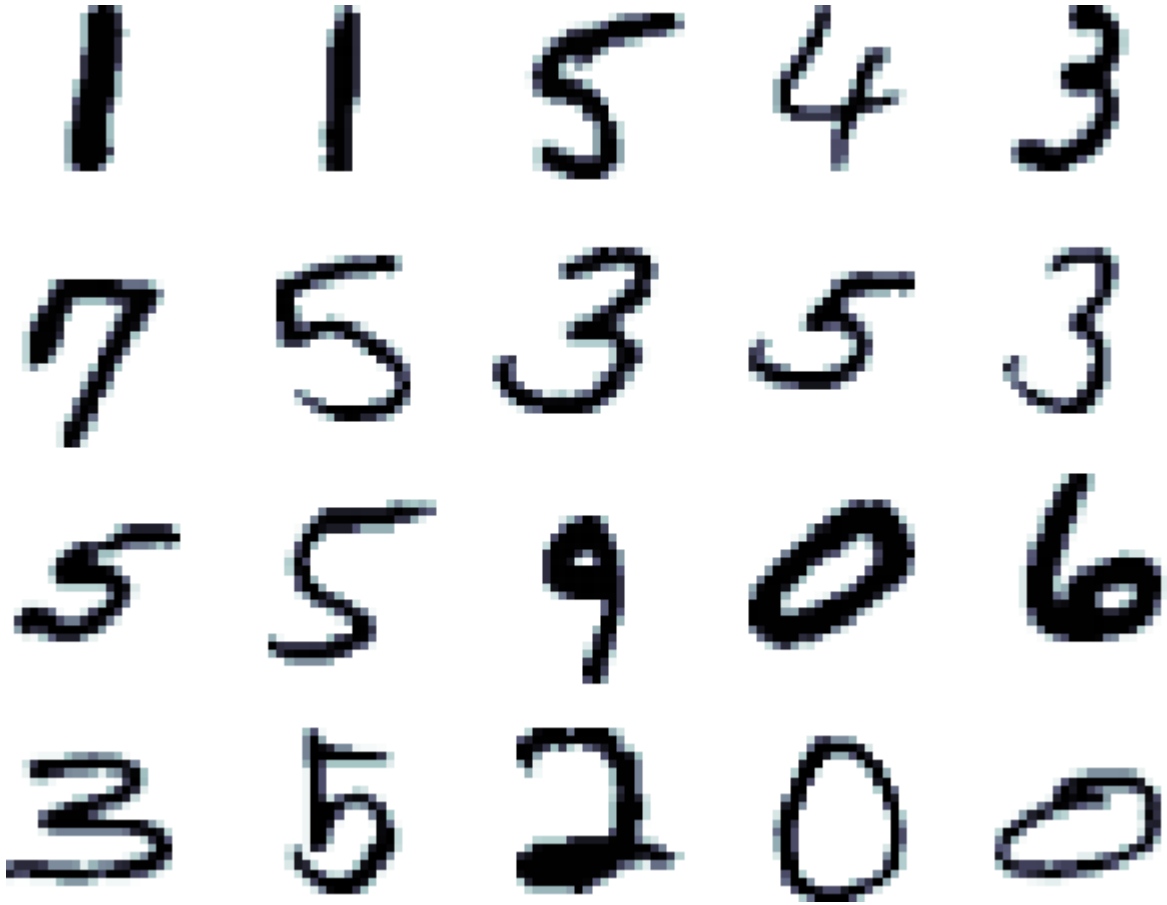
MNIST dataset of handwritten digits.

Each image is an *entity*.

Each image has a *label* attribute which states which of the digits 0,1,...9 is represented by the image.

What type of data is this (categorical, continuous numeric, or discrete numeric)?

Other examples



Other examples

Each image is represented by grayscale values in a 28x28 grid.

That's 784 attributes, one for each square in the grid, containing a grayscale value.

What type of data are these other 784 attributes?

Other important datatypes

- Text: Arbitrary strings that do not encode a categorical attribute.
- Datetime: Date and time of some event or observation (e.g., `arrestDate`, `arrestTime`)
- Geolocation: Latitude and Longitude of some event or observation (e.g., `Location.`)
- Relationships: links between entities, with links having their own attributes (e.g., social network, how long have two people followed each other)

Units

Something that we tend to forget but is **extremely** important for the modeling and interpretation of data.

Attributes are *measurements* and that they have *units*.

For example, age of a person can be measured in different units: *years*, *months*, etc.

Units

These can be converted to one another, but nonetheless in a given dataset, that *attribute* or measurement will be recorded in some specific units.

Similar arguments go for distances and times, for example.

Units

In other cases, we may have unitless measurements (we will see later an example of this when we do *dimensionality reduction*).

In these cases, it is worth thinking about *why* your measurements are unit-less.

Units

When performing analyses that try to summarize the effect of some measurement or attribute on another, units matter a lot!

We will see the importance of this in our *regression* section.

For now, make sure you make a mental note of units for each measurement you come across.

Important when modeling and interpreting the results of these models.