Data 100, Fall 2020

Discussion #1

September 2, 2020

Welcome to Data 100!

Linear Algebra Fundamentals

1. Linear algebra is what powers linear regression, logistic regression, and PCA (concepts we will be studying in this course). Moving forward, you will need to understand how matrix-vector operations work. That is the aim of this problem.

   Fernando, Anthony, and Kobe are shopping for fruit at Berkeley Bowl. Berkeley Bowl, true to its name, only sells fruit bowls. A fruit bowl contains some fruit and the price of a fruit bowl is the total price of all of its individual fruit.

   Berkeley Bowl has apples for \$2, bananas for \$1, and cantaloupes for \$4. (expensive!). The price of each of these can be written in a vector:

$$\vec{v} = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$$

   Berkeley Bowl sells the following fruit bowls:

   1. 2 of each fruit
   2. 5 apples and 8 bananas
   3. 2 bananas and 3 cantaloupes
   4. 10 cantaloupes

   (a) Define a matrix B such that

$$B\vec{v}$$

   evaluates to a length 4 column vector containing the price of each fruit bowl. The first entry of the result should be the cost of fruit bowl 1, the second entry the cost of fruit bowl 2, etc.

(b) Fernando, Anthony, and Kobe make the following purchases:

- Fernando buys 2 fruit bowl 1s and 1 fruit bowl 2.
- Anthony buys 1 of each fruit bowl.
- Kobe buys 10 fruit bowl 4s (he really like cantaloupes).

Define a matrix A such that the matrix expression

$$AB\vec{v}$$

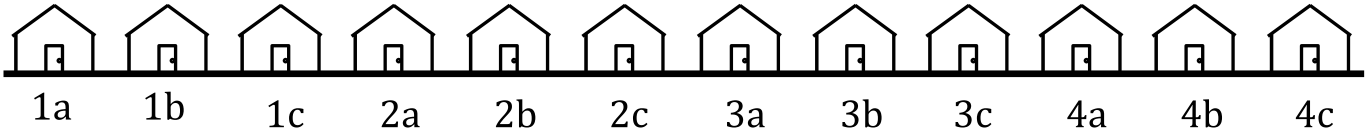
evaluates to a length 3 column vector containing how much each of them spent. The first entry of the result should be the total amount spent by Fernando, the second entry the amount sent by Anthony, etc.

(c) Let's suppose Berkeley Bowl changes their fruit prices, but you don't know what they changed their prices to. Fernando, Anthony, and Kobe buy the same quantity of fruit baskets and the number of fruit in each basket is the same, but now they each spent these amounts:

$$\vec{x} = \begin{bmatrix} 80 \\ 80 \\ 100 \end{bmatrix}$$

In terms of A, B, and $\vec{x}$, determine $\vec{v_2}$ (the new prices of each fruit).

Probability & Sampling



2. Kalie wants to measure interest for a party on her street. She assigns numbers and letters to each house on her street as illustrated above. She picks a letter “a”, “b”, or “c” at random and then surveys every household on the street ending in that letter.

   (a) What is the chance that two houses next door to each other are both in the sample?

   (b) Now, suppose that Kalie decides to collect an SRS of one house instead. What is the probability that house 1a is **not** selected in Kalie’s SRS of one house?

   (c) Kalie decides to collect a SRS of four houses instead of a SRS of one house. What is the probability that house 1a is **not** in Kalie’s simple random sample of four houses?

   (d) Instead of surveying every member of each house from the SRS of four houses, Kalie decides to only survey two members in each house. Four people live in house 1a, one of whom is Bob. What is the probability that Bob is **not** chosen in Kalie’s new sample?

Proportions

In Data 100 you will typically work with multiple variables and large data sets. But before we get carried away by complexity, let's make sure we have our feet on the ground when it comes to interpreting simple quantities like proportions.

3. Investigators at the scene of a crime find a footprint that shows a distinctive pattern on the sole. They identify the type of shoe, and then they find a person owns that kind of shoe and could have committed the crime. They put this person on trial for the crime.

   After looking at sales patterns and so on, the investigators find that of the 10,000 other people who could have committed the crime, 1 in 1,000 own that kind of shoe.

   The prosecution says that given these findings, the chance that the defendant is not the guilty person is 1 in 1,000.

   The prosecution has made an error called the "prosecutor's fallacy." Unfortunately it's rather common. Let's see what the error is and what conclusions we can draw from the evidence.

   (a) There are 10,001 people who could have committed the crime. Define a person to be "Matching the Footprint" if the person owns the kind of shoe identified by the investigators. Fill in the table below with the counts of people in the four categories. The four counts should add up to 10,001, and you should assume, as the prosecution did, that only one person is guilty.

	Guilty	Not Guilty
Matching the Footprint		
Not Matching the Footprint		

   (b) The prosecution has reported a proportion as a chance. Whether they know it or not, this implies they are assuming that the defendant is like a person drawn at random from the group who could have committed the crime. So let's assume that too. That is, we assume the defendant is drawn at random from 10,001 people of whom 1 is guilty.

   Use the table in Part **a** to fill in the blanks with choices from among "Guilty", "Not Guilty", "Matching the Footprint", and "Not Matching the Footprint". The vertical bar is the usual notation for "given".

   Under this assumption, $\frac{1}{1000} = P(\_\_\_\_\_\_\_ \mid \_\_\_\_\_\_\_)$.

(c) What the investigators know is that the defendant has the fateful type of shoe. Fill in the blanks:

Given the findings of the investigators, the chance that the defendant is not guilty is $P(\_\_\_\_\_\_\_ \mid \_\_\_\_\_\_\_) = \_\_\_\_\_\_\_$.

The last blank should be filled with a fraction, and the first two should be filled choices from among "Guilty", "Not Guilty", "Matching the Footprint", and "Not Matching the Footprint".

Note: The prosecution's error is to confuse the probabilities in Parts **b** and **c**.