Data 100, Spring 2022

Discussion #1 Solutions

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.

   Josh, Lisa, 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.

   Solution:

   $$B = \begin{bmatrix} 2 & 2 & 2 \\ 5 & 8 & 0 \\ 0 & 2 & 3 \\ 0 & 0 & 10 \end{bmatrix}$$

(b) Josh, Lisa, and Kobe make the following purchases:

- Josh buys 2 fruit bowl 1s and 1 fruit bowl 2.
- Lisa 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 Josh, the second entry the amount sent by Lisa, etc.

Solution:

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 10 \end{bmatrix}$$

(c) Let's suppose Berkeley Bowl changes their fruit prices, but you don't know what they changed their prices to. Josh, Lisa, 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).

Solution:
We know that $\vec{x} = AB\vec{v_2}$ from the previous part. To solve for $\vec{v_2}$ we need to left-multiply both sides of the above equation by $(AB)^{-1}$. Doing so yields $\vec{v_2} = (AB)^{-1}\vec{x}$.
This assumes that the product AB is invertible. If you work out the product, you will see that it is full rank and thus invertible.

2. As a warm up for the homework, we will introduce matrix inverses and matrix rank.

- The inverse of a square invertible matrix M, M^{-1} is defined as a matrix such that $MM^{-1} = I$ and $M^{-1}M = I$. The matrix I is a special matrix denoted as the identity matrix where the diagonal elements are 1 and the non-diagonal elements are 0.
- Linear dependence among a set of vectors $\{v_1, v_2, v_3, ..., v_n\}$ is defined as follows. If any (non-trivial) linear combination of the vectors can produce the zero vector, then the set of vectors is linearly dependent.

In other words, if we can multiply the vectors v_i with some scalar α_i and sum the quantity to obtain the zero vector (given at least one $\alpha_j \neq 0$, then the set is linearly dependent.

$$\sum_{i=1}^{n} \alpha_i v_i = 0 \text{ such that some } \alpha_j \neq 0 \implies \text{linear dependence}$$

Any set of vectors such that we cannot obtain the zero vector as described above is linearly independent.

- The (column) rank of a matrix M is the maximal number of linearly independent column vectors in M. A full rank matrix has a column rank equal to the number of column vectors.

We will go over all of these definitions applied to relevant practical examples in the following subparts.

(a) Consider the matrix $M = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$ containing two column vectors $v_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $v_2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$. Is it possible to construct the zero vector using a linear combination of the column vectors? What can be concluded about the rank of the matrix M?

Solution: No, it is impossible to construct the zero vector if we use at least one $\alpha_i \neq 0$ since the first vector can't affect the second dimension and vice versa. Hence, neither of the vectors can "undo" each other. As a more formal proof:

$$\alpha_1 v_1 + \alpha_2 v_2 = \begin{bmatrix} 2\alpha_1 \\ 3\alpha_2 \end{bmatrix}$$

If at least one of the α values are not 0, then the above quantity can never be 0. Hence, this matrix is full rank (i.e. the set of vectors is linearly independent).

(b) Consider the inverse matrix $M^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ of M. Carry out the matrix multiplication MM^{-1}, and determine what M^{-1} must be.

Solution: We carry out the matrix multiply:

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 2a & 2b \\ 3c & 3d \end{bmatrix}$$

Hence, since we know that $MM^{-1} = I$, $c = d = 0$ and $2a = 1$ and $3d = 1$. Thus, the inverse matrix is:

$$M^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

An interesting fact about diagonal matrices (i.e. matrices that are non-zero on the diagonal entries but zero everywhere else) is that their inverse is simply the elementwise multiplicative inverse of the diagonal entries!

(c) Consider a different matrix $Q = \begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & 5 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$. What is the column rank of the matrix? Is the matrix invertible?

Solution: We can construct the zero vector with all of v_1, v_2 and v_3, so this initial set is linearly dependent since $5v_1 + 5v_2 - v_3 = 0$. Hence, the maximal number of linearly independent vectors we can have is 2, where we remove any one of the vectors among v_1, v_2 and v_3. A similar argument as from the first subpart can be applied to explain why there is no linear dependence in any 2 of the vectors (i.e. $\{v_1, v_2\}$, $\{v_2, v_3\}$, and so on). The column rank is 2.

The matrix is not invertible since it is not square.

(d) Consider a matrix R, which is equal to the transpose of the matrix Q: $R = Q^T$. What is the column rank of the matrix R? Is the matrix R invertible?

Solution: We take the transpose:

$$R = Q^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 5 & 5 \end{bmatrix}$$

The column rank is 2 because we can never "undo" the first vector dimension with the second and vice versa, similarly to previous subparts. Thus, this matrix is full column rank.

The matrix is not invertible since it is not square, but take a look at the next bonus question to find out the conditions under which the matrix $R^T R$ can be inverted (this figures into how we study linear regression later on)!

3. *(Bonus)* We will explore a few properties of a special symmetric matrix that we will use quite a bit when we cover linear regression and regularization. Consider a matrix X, of shape m by n for some $m \geq n$. We will work with a matrix that is created by matrix multiplying X^T with X: X^TX. We will prove an interesting property:

   If X is of full column rank, then X^TX is invertible.

   (a) Explain why the matrix X^TX is symmetric. Recall that the transpose operation turns the first column into the first row, second column into the second row, and so on.

   Solution: Consider that X is a collection of column vectors x_i for $i < n$. We write the relevant computation as follows:

   $$\begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \\ \dots \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 & \dots \end{bmatrix} = \begin{bmatrix} x_1^Tx_1 & x_1^Tx_2 & x_1^Tx_3 & \dots \\ x_2^Tx_1 & x_2^Tx_2 & x_2^Tx_3 & \dots \\ \dots & & & \end{bmatrix}$$

   Note that all the diagonals are simply the squared norms of the column vectors x_i. The off-diagonals are symmetric since $x_i^Tx_j = x_j^Tx_i$.

   (b) Recall that invertibility for square matrices depends on the matrix's rank. Simply put, redundant (or linearly dependent) vectors with respect to a matrix's span reduce its rank. We will determine whether X^TX is invertible by using an equivalent condition without proof (for those curious, this leverages the fact that a square matrix is positive definite implies invertibility).

   Lemma 0.1. *If for all non-zero vectors v, $v^TMv > 0$, then M is invertible.*
   What condition is required to invert the matrix X^TX?

   Solution:
   Applying the statement to this case, we get that for all non-zero vectors v, $v^TX^TXv > 0$ for the matrix to be invertible. We will simplify this quantity later on to get a more interpretable result!

   (c) The nullspace of a matrix M is defined as the set of all vectors that when multiplied by the matrix M yield 0. In other words, it is the set $\{v : Mv = 0\}$. Prove that if X has a trivial nullspace (i.e. a nullspace with only the zero vector), then X^TX is invertible.

   In case you're curious, this is the exact condition that we will need for least squares regression to work later on!

Hint: Using the fact that $(Av)^T = v^T A^T$, simplify the condition in the previous part by creating a new vector w and leverage the fact that $w^T w = \|w\|^2$.

Solution:
Let $w = Xv$. Then, we simplify the condition from above:

$$v^T X^T X v = w^T w = \|w\|^2 = \|Xv\|^2 > 0$$

Simplify further, and given that a norm of zero is only produced with the zero vector:

$$\|Xv\| > 0 \implies Xv \neq 0$$

Hence, we have simplified the condition to say that if for all non-zero vectors v, $Xv \neq 0$, then $X^T X$ is invertible. Note that any vector v such that $Xv = 0$ must be in the nullspace of X by its definition. Hence, we can conclude that if no non-zero vector exists in the matrix X's nullspace (otherwise known as a trivial nullspace), then $X^T X$ is invertible.

As an aside, this statement can be equivalently written that if X is full column rank, then $X^T X$ is invertible since a trivial nullspace implies full column rank.

(d) Do the same analysis with $X^T X + \lambda I$ for $\lambda > 0$. You should notice that the conditions required for invertibility are much more lax; this is the setup for ridge regression, which we'll study later too!

Solution: We use the same lemma as above:

$$v^T (X^T X + \lambda I) v = v^T X^T X v + \lambda v^T I v = \|Xv\|^2 + \lambda \|v\|^2$$

Note that this quantity is always positive since v is non-zero by definition per the lemma, and hence $\lambda \|v\|^2 > 0$. The squared quantity $\|Xv\|^2$ is at least zero, hence, the overall sum is positive.

This implies that this matrix is always invertible given that $\lambda > 0$ - and this is one of the advantages of using ridge regression that we will study!

Calculus

In this class, we will have to determine which inputs to a functions minimize the output (for instance, when we choose a model and need to fit it to our data). This process involves taking derivatives.

In cases where we have multiple inputs, the derivative of our function with respect to one of our inputs is called a *partial derivative*. For example, given a function $f(x, y)$, the partial derivative with respect to x (denoted by $\frac{\partial f}{\partial x}$) is the derivative of f with respect to x, taken while treating all other variables as if they're constants.

4. Suppose we have the following scalar-valued function on x and y:

$$f(x, y) = x^2 + 4xy + 2y^3 + e^{-3y} + \ln(2y)$$

(a) Compute the partial derivative of $f(x, y)$ with respect to x.

Solution:

$$\frac{\partial}{\partial x} f(x, y) = 2x + 4y$$

(b) Compute the partial derivative of $f(x, y)$ with respect to y.

Solution:

$$\frac{\partial}{\partial y} f(x, y) = 4x + 6y^2 - 3e^{-3y} + \frac{2}{2y}$$

(c) The gradient of a function $f(x, y)$ is a vector of its partial derivatives. That is,

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T$$

$\nabla f(x, y)$ tells us the magnitude and direction in which f is moving, at point (x, y). This is analogous to the single variable case, where $f'(x)$ is the rate of change of f, at the point x.

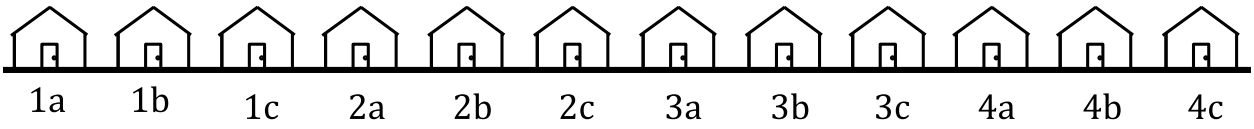
Using your answers to the above two parts, compute $\nabla f(x, y)$ and evaluate the gradient at the point $(x = 2, y = -1)$.

Solution:

$$\nabla f(x, y) = [2x + 4y, 4x + 6y^2 - 3e^{-3y} + \frac{2}{2y}]^T$$

$$\nabla f(2, -1) = [0, 13 - 3e^3]^T$$

Probability & Sampling



5. 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?

   Solution: None of the adjacent houses end in the same letter, so the chance is zero.

   (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?

   Solution: The probability that house 1a is selected is $\frac{1}{12}$. Therefore, the probability that house 1a is not selected is $1 - \frac{1}{12} = \frac{11}{12}$.

   (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?

   Solution: This time, we are taking a sample of 4 houses. But, we can apply a similar approach from part b to determine the probability of missing house 1a in each of the four selected houses. Then we multiply the four probabilities together to get our answer. The probability that house 1a is not in Kalie's sample is $\frac{11}{12} \cdot \frac{10}{11} \cdot \frac{9}{10} \cdot \frac{8}{9} = \frac{8}{12} = \frac{2}{3}$.

   (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?

   Solution: The probability that house 1a is included in Kalie's initial SRS is $\frac{1}{3}$. Given that house 1a is selected, the probability that Bob is one of the two people

surveyed is $\frac{1}{2}$. Therefore, the probability that Bob *is* surveyed is $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$. Thus the probability that Bob is **not** selected is $1 - \frac{1}{6} = \frac{5}{6}$.

Proportions (Bonus)

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.

6. 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		

Solution:

	Guilty	Not Guilty
Matching the Footprint	1	10
Not Matching the Footprint	0	9990

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

> **Solution:** P(Matching the Footprint | Not Guilty) because it's 10/(10+9990)

(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".

> **Solution:** P(Not Guilty | Matching the Footprint) = 10/(1+10) = 10/11

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

Election Forecasts (Bonus)

7. People have a hard time understanding polls. In September 2016, the New York Times tried to explain aspects of polls that tend to get overlooked. To illustrate the issues, they gave all the data in one of their own polls to four well-known forecasters and asked them to make predictions.

   The data were from a poll of 867 Florida voters and the exercise was to predict Trump/-Clinton result in Florida. In the election, Trump beat Clinton in Florida, 49% to 47.8%.

   Here are the forecasts. The Times' own results, derived jointly with researchers at Siena College, are in the last line of the table.

Pollster	Clinton	Trump	Margin
Charles Franklin Marquette Law	42%	39%	Clinton +3%
Patrick Ruffini Echelon Insights	39%	38%	Clinton +1%
Omero, Green, Rosenblatt Penn Schoen Berland Research	42%	38%	Clinton +4%
Corbett-Davies, Gelman, Rothschild Stanford University/Columbia University/Microsoft Research	40%	41%	Trump +1%
NYT Upshot/Siena College	41%	40%	Clinton +1%

(a) Pick one of the options (i) and (ii); if you pick (ii), provide the reason.

The predictions were different from each other because

(i) samples come out differently due to randomness so the forecasters all had different data.

(ii) _____

Solution: The forecasters used the data differently so they made different predictions even though the data were the same. Specifically, they used the same data but different weights and models.

(b) Point out one other interesting aspect of the data in the table. This question doesn't have just one right answer; just describe something you noticed.

Solution:

Possible answers:

- Gelman's team picked Trump to have a 1% margin over Clinton, which was almost correct. But they were quite far off both candidates' percents.
- All the forecasters' Clinton+Trump percents were much less than 100. Undecided or "other" voters in September made a difference in November.
- Several of the forecasters ended up with pretty small margins, which is essentially saying that the election result could go either way. After the election, Nate Silver listed "a failure to appreciate uncertainty" as one of the shortcomings of political coverage by the media.

(c) If you were going to forecast an election result, which of the following groups would you most want to focus on, and why? Pick at most two groups.

(i) adults in the Census

(ii) eligible voters

(iii) registered voters

(iv) likely voters

(v) undecided voters

Solution: Elections are decided by who actually votes, so registered voters and likely voters are worth focusing on.

(d) Of the two main methods for identifying likely voters, described below, one does a better job at predicting whether the person will show up and vote. Which do you think it is, and why? Could it systematically exclude some likely voters?

- Self-reported voting intention
- Voting history (in which past elections did the person vote; data available in the voter registration database)

Solution: Voting history. What people do is a better predictor than what they say. But using history alone (as Gelman et al and Ruffini did in the table) excludes those who registered after the last election, often young voters.