

**Goals.** The main purpose of this assignment is to help review prerequisite concepts in this course, including linear algebra and probability. The latter sections of this assignment will also serve as a warmup to get you familiar with common concepts in NLP.

## Assignment

### Part 1: Review of Linear Algebra

How is linear algebra related to NLP? Thanks to neural networks, they’re very closely intertwined these days. We usually represent words as vectors of real numbers (**embeddings**, or embedding vectors). These vectors are refined over many layers of a neural network into increasingly abstract representation vectors.<sup>1</sup> We also learn many **weight matrices** that will be multiplied with a representation vector as part of the process of computing the hidden representation vector for the next layer. Thus, a neural network is composed largely of matrix–vector and matrix–matrix multiplications, followed by some calculus to update the values of the weight matrices.

We also often use linear algebraic concepts like  $L_2$  norms to compute the magnitude of a representation, and dot products or cosine similarities to compare the similarities of two representation vectors.

**Q1.** Perform the following matrix multiplications. Write “undefined” if the matrix multiplication is not possible.

$$(a) \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}$$

$$\begin{bmatrix} 23 \\ 53 \end{bmatrix}$$

$$(b) \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 2 & 3 & 5 \end{bmatrix}$$

undefined

$$(c) \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^\top \begin{bmatrix} 1 & 1 & 2 \\ 3 & 5 & 8 \end{bmatrix}$$

$$\begin{bmatrix} 13 & 21 & 34 \\ 17 & 27 & 44 \\ 21 & 33 & 54 \end{bmatrix}$$

**Q2.** Compute the Frobenius ( $L_2$ ) norm of this matrix:  $\begin{bmatrix} 1 & 2 & 3 \\ 5 & 8 & 13 \\ 5 & 7 & 9 \end{bmatrix}$   $\sqrt{427} \approx 20.66$

**Q3.** Write the inner product (dot product) of the following two vectors:  $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 5 & 10 & 15 \end{bmatrix}$   $70$

**Q4.** Write the outer product of the following two vectors:  $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 5 & 10 & 15 \end{bmatrix}$   $\begin{bmatrix} 5 & 10 & 15 \\ 10 & 20 & 30 \\ 15 & 30 & 45 \end{bmatrix}$

**Q5.** What is the cosine similarity of these vectors?  $\begin{bmatrix} 2 & 5 & 9 \end{bmatrix} \begin{bmatrix} -1 & 3 & 3 \end{bmatrix}$   $\frac{40}{\sqrt{110} \times \sqrt{19}} \approx 0.875$

**Q6.** What is the rank of this matrix?  $\begin{bmatrix} 1 & 5 & 9 \\ 3 & 15 & 27 \\ 12 & 60 & 108 \end{bmatrix}$   $1$

<sup>1</sup>Some use “embedding” to refer to representation vectors only before the first layer of a neural network, while others use it to refer to representation vectors in *any* layer. In this course, I will use “embedding” in the latter sense, but will usually formally clarify exactly what is meant using mathematical notation.

## Part 2: Review of Probability

More obviously related to NLP is probability. Language models are just machines that take prior context as input and produce probability distributions over continuations.

**Q7.** Assume we have a probability distribution over 6 outcomes  $y_i \in Y$ , where  $Y$  is a random variable. (Think of rolling a six-sided die). The probability distribution over  $y_i \in Y$  is uniform.

- What is the probability of rolling a 6 in one roll?  $\frac{1}{6} = 0.1\bar{6}$
- In three rolls, what is the probability of rolling a 6, a 3, and a 1, in that order?  $\frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216} \approx 0.005$
- What is the probability of rolling a 6, a 3, or a 1 in one roll?  $\frac{1}{2} = 0.5$
- In three rolls, what is the probability of rolling a 6, a 3, and a 1, in *any* order? For example, a 1-3-6 would fit this criterion, as would a 3-1-6. There are  $3! = 6$  permutations of a 1-3-6 sequence. We divide this by the total number of possible sequences of 3 rolls, which is  $(\frac{1}{6})^3$ . This yields  $\frac{6}{216} = 0.02\bar{7}$

**Q8.** Assume we have a random variable  $X$ .  $X$  can take one of four integer values. The probability distribution over these values is as follows:  $p(1) = 0.5$ ,  $p(2) = 0.2$ ,  $p(5) = 0.25$ ,  $p(10) = 0.05$ .

- What is the expectation over this distribution?  $(0.5 \times 1) + (0.2 \times 2) + (0.25 \times 5) + (0.05 \times 10) = 2.65$
- What is the entropy of this distribution? Use base-2 logarithms.  $\approx 1.68$
- If you could rearrange the probabilities (but not the number of values nor the values themselves that  $X$  can take), what is the maximal entropy that you could obtain over this distribution? Use base-2 logarithms. 2

**Q9.** Take the following joint probability distribution over random variables  $X$  and  $Y$ :

$p(X, Y)$	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0.05	0.30	0.05
$X = 2$	0.30	0.05	0.05
$X = 3$	0.05	0.05	0.30

- Are  $X$  and  $Y$  independent? Why or why not? No; if they were independent, their joint probabilities would equal the product of individual probabilities. For example,  $p(X = 1) = 0.4$  and  $p(Y = 1) = 0.4$ , so if  $X$  and  $Y$  were independent, then  $p(X = 1, Y = 1)$  would be  $0.4 \times 0.4 = 0.16$ . However, the table says that  $p(X = 1, Y = 1) = 0.05$ .  $0.05 \neq 0.16$ , so  $X$  and  $Y$  are not independent.
- What is  $p(X = 2 | Y = 3)$ ?  $p(X = 2|Y = 3) = \frac{p(X=2,Y=3)}{p(Y=3)} = \frac{0.05}{0.4} = 0.125$

**Q10.** Assume the probability that a student does HW-1 for NLP is 0.6; we'll represent this as  $p(H = 1) = 0.6$ . The probability that the student passes the course assuming they did HW-1  $p(P = 1|H = 1)$  is 0.9; the probability that the student passes assuming they did *not* do HW-1  $p(P = 1|H = 0)$  is 0.7. If we randomly select a student from the course, what is the probability that they will pass? In other words, what is  $p(P = 1)$ ? Hint: recall the chain rule of probabilities. We need  $p(P = 1)$ . This is equal to  $p(P = 1, H = 0) + p(P = 1, H = 1)$ . We can use the chain rule to break each term down:  $p(P = 1, H = 0) = p(P = 1|H = 0)p(H = 0)$ . Thus,

$p(P = 1) = p(P = 1|H = 0)p(H = 0) + p(P = 1|H = 1)p(H = 1) = 0.7 \times 0.4 + 0.9 \times 0.6 = 0.82.$

The actual probability is hopefully much higher! :)

## Part 3: Review of Differential Calculus

To update the weight matrices of a neural network, we use derivatives. We'll go more in-depth on how this works in class and in the book, but for now, it will be helpful to refresh your knowledge of the basics.

**Q11.** What is  $\frac{\partial}{\partial x}x^3y^3$ ?  $3x^2y^3$

**Q12.** The sigmoid function, often denoted  $\sigma(x)$ , is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

What is  $\frac{d\sigma(x)}{dx}$ ?  $\frac{1}{1+e^{-x}} \cdot (1 - \frac{1}{1+e^{-x}})$ , or equivalently,  $\sigma(x) \cdot (1 - \sigma(x))$

**Q13.** We will often denote exponential functions like  $e^x$  as  $\exp(x)$ . What is  $\frac{\partial}{\partial x}\exp(x^2y^2)$ ?  $2xy^2\exp(x^2y^2)$