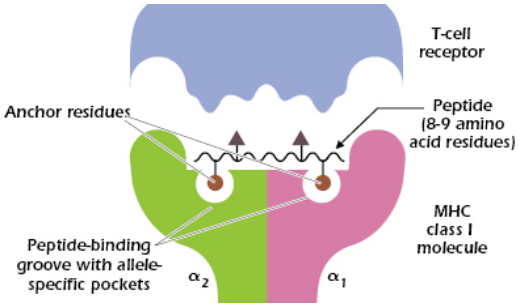
Introduction to Deep Learning - Exercise #1

submission date: 30/4/2024

Programing Task: Antigen Discovery for SARS-CoV-2 ("Corona") Virus Vaccine

Let's find potential specific antigens to the SARS-CoV-2 virus (the virus behind the COVID-19 pandemic). The antigens are sub-sequences of the virus protein sequence recognizable by our immune system. Our adaptive immune system consists of 6 HLA (class I) alleles that allow it to selectively identify small fragments of proteins, known as peptides (as illustrated by the figure on the right). The system evolved to recognize only peptides of a foreign body and by that invoke an immune proliferation and response of B- and T-cells that destroy the intruder. However, unfortunately not all foreign peptides are recognized. For those of you who are interested in learning more about this biological mechanism, I suggest this Wiki page, as well as this page.



In this exercise you will train a deep neural network to identify the peptides detected by six popular HLA alleles in the western population, as well as negative examples. A peptide is expected to be detected by a single allele if any. The training data consists of six ~12,000-25,000 positive and ~245,000 negative peptides. Each peptide consists of 9 amino acids (of 20 types). At a second stage, you will use your trained predictor to identify sequences of 9-amino-acids peptides from the Spike protein of the SARS-CoV-2 virus.

Formally,

1. You will find the training data at the course's moodle page. Split it to train and test sets randomly at a 1:9 ratio.
2. We'd like to set up a small multi-layered perceptron (MLP) network to accept this data(*) and output the proper prediction (detect / not detect).
    a. How would you represent these 9-mers of amino acids (a vocabulary of 20 amino acids)? How would you represent the associate alleles? Explain your reasoning.
    b. What will the network's input dimension be under this representation choice? Implement an MLP that keeps this dimension for 2 inner layers, and plot the resulting train and test losses. Does this dimension pose a training problem?
    c. Describe a network architecture that avoids the problems seen, explain all your design choices. Plot the train / tests losses. Include the plots obtained in your report.
    d. Remove the non-linear operators from your network, and report the results obtained. How do the results compare?

(*) Notice that the training data contains more negative examples than positive. Use some machine learning strategy to deal with this form of unbalance, as well as when measuring your performance.

e. Use your model to predict the detection of 9-amino-acids peptides from the **Spike** protein of the SARS-CoV-2, that is: all its consecutive 9-mer segments out of its 1273-amino-acid sequence. You can download this sequence from this page: https://www.uniprot.org/uniprotkb/P0DTC2/entry. Include in your report (as well as notify the CDC) the top 3 most detectable peptides in this protein (and by which allele).

Theoretical Questions:

1. Show that the composition of linear functions is a linear function. Show that the composition of affine transformations remains an affine function.
2. The Gradient Descent (GD) method and the learning rate:
   a. Write down the GD update equations for x and y when minimizing:

$$\min 10(x-1)^2 + (y+1)^2/10$$

   b. What's the maximal learning rate possible - what happens when it is exceeded? (hint: reach a form of $x^{k+1} = Ax^k + B$ for both x and y)
   c. What value in the minimization term determines this maximal learning rate?
   d. Which variable takes the longest to converge? What value in the minimization term determines that?
3. Assume a network is required to predict an angle (0-360 degrees). How will you define a prediction loss that accounts for the circularity of this quantity, i.e., the loss between 2 and 360 is not 358, but 2 (since 0 is 360..). Write your answer in a pseudo-code.

4. Chain Rule. Differentiate the following terms (specify the points of evaluation of each function):
   a.

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

   b.

$$f_1\Big(f_2\big(\ldots f_n(x)\big)\Big)$$

   c.

$$f_1\Big(x, f_2\big(x, f_3(\ldots f_{n-1}(x, f_n(x)))\big)\Big)$$

   d.

$$f\Big(x + g\big(x + h(x)\big)\Big)$$

Submission Guidelines:

The submission is in **<u>pairs</u>.** Please submit a single zip file named “ex1_ID1_ID2.zip”. This file should contain your code, along with an ”ex1.pdf” file which should contain your answers to the theoretical part as well as the figures/text for the practical part. Furthermore, include in this compressed file a README with your names and cse usernames.
Please write readable code, with documentation where needed, as the code will also be checked manually.

Late submission - 10 points reduction for each day. Submissions will not be accepted after 4 days.