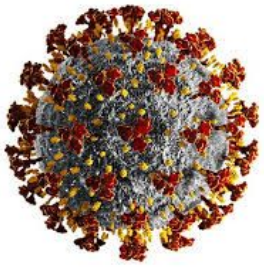
Introduction to Deep Learning - Exercise #1

submission date: 30/11/2020

Programing Task: Antigen Discovery for SARS-CoV-2 (“Corona”) Vaccine

Time to end the COVID-19 pandemic by finding potential antigens. The latter are sub-sequences of the virus proteins that can be recognized 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. The system is evolved to recognize only peptides of a foreign body and by that invoke an immune proliferation and response of T-cells that destroy the intruder (unfortunately not all foreign peptides are recognized). For those of you who are interested in learning more about this mechanism, I suggest this Wiki page.



In this exercise you will train a deep neural network to identify the peptides that a specific HLA allele detects, namely the HLA_A0201 which is a very common allele shared by 24% of the western population. The training data consists of ~3,000 positive and ~24,500 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, and welcome to use the TensorFlow 2 quickstart for experts as your starting point.
2. Set up a multi-layered perceptron network to accept this data and output the proper prediction (detect / not detect). Test different architectures (number of levels, neurons at each level, etc.), and non-linearities (RelU, sigmoid) and pick the one achieving the highest accuracy on the test set. List the tests you conducted in the submission. Detail the chosen architecture in your document.
3. Load the data from the files, and map it to the proper mathematical representation. Split it into 90%/10% train/test sets (picked at random at each run to avoid over-fitting).
4. Train the network till convergence of the (train/test) loss plots. Make sure your learning rates are not too small and certainly not too large. Detail the chosen parameters in your document and add the train/test loss plots to the submission pdf.
5. Once you find your best configuration, plot the train and test accuracies, and add it to your submission.
6. Use your model to predict the detection of 9-amino-acids peptides from the Spike protein of the SARS-CoV-2 (all consecutive 9-mer segments from the entire 1273-amino-acid protein sequence). You can download this protein sequence from: https://viralzone.expasy.org/8996

7. Notify the CDC of your predicted antigens, as well as submit the top 5 scoring peptides as part of your exercise.

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 calculus behind the Gradient Descent method
    a. What is the stopping condition of this iterative scheme,

$$f(x^{n+1}) = f(x^n) - \alpha \nabla f(x^n)$$

    b. Use the second-order multivariate Taylor theorem,

$$f(x + dx) = f(x) + \nabla f(x) \cdot dx + dx^T \cdot H(x) \cdot dx + O(\|dx\|^3),$$
$$H_{ij}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

    to derive the conditions for classifying a stationary point as local maximum or minimum.
3. Assume the network is required to predict an angle (0-360 degrees). How will you define a prediction loss which 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 tensorflow-codable form.
4. Explain why Cybenko and Hornik theorems also imply that linear combinations of translated and dilated RelU functions form a dense set in C[0,1].
5. Generalize the construction of a deep network that expresses a shallow network in O(N) neurons, that we saw in class, to signed functions.

Submission Guidelines:

The submission is in **pairs.** 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 compress 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.