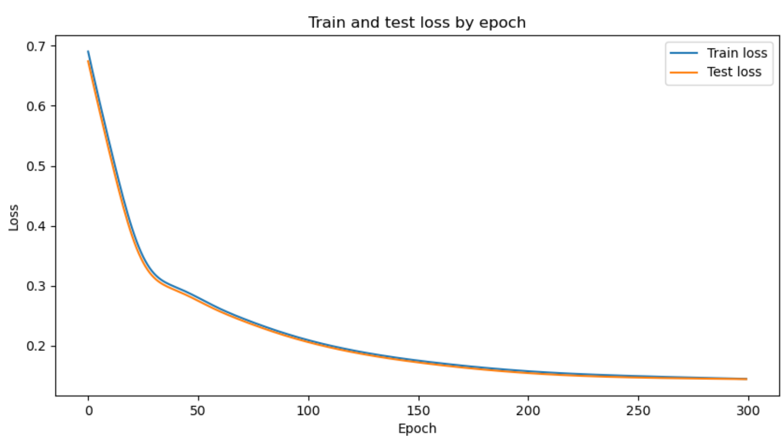
Exercise 1 Intro to Deep Learning: Nadav Eisen & Yonatan Miroshnik

May 30, 2024

Practical Part:

2.a. For every character in the 9 character peptide, I gave it a unique index, and from that I created a one-hot embedding for each character based on this unique index. Each 9 character length peptide is thus represented by a 180 bit long binary number of the consecutive one-hot embeddings of each character in the peptide.

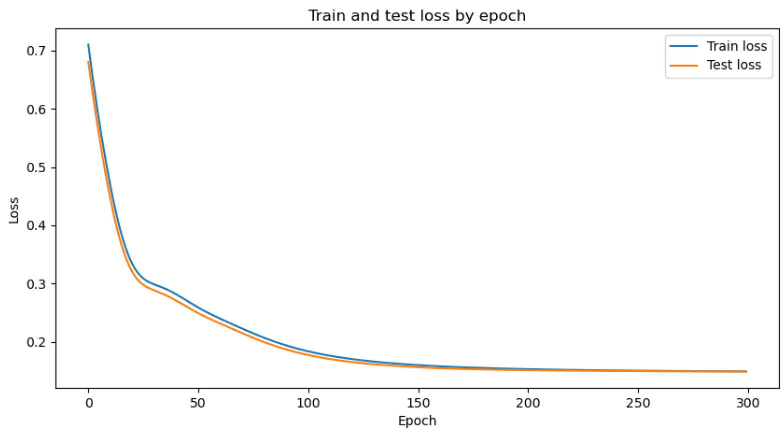
2.b. The input dimension, as explained above, is a 180 bit long input. The offered neural network set up has a problem in our need to decide the size of the hidden layers and thus the number of weights, if we decide on a hidden linear layer size similar to the input, we start seeing overfitting at some point.



2.c. We propose reducing the hidden layer size to 128, this makes the model overfit less and thus produces better training results.

2.d. The final test loss result is: 0.1439

The totally linear model(similar to the previous one, without the middle ReLU layer) learns the problem well, but stops bettering itself through training at a former phase.



2.e. We first go over the spike protein as given and divide it into all possible consecutive 9-peptides, we then pass it through the trained network, and return the most confidently positive peptides of the prediction, our resulting peptides are: NLREFVFKN, LLIVNNATN, FDNPVLPFN

Question 1:

From basic linear algebra, we know that any linear functions $f \in Hom\left(\mathbb{R}^n, \mathbb{R}^k\right)$, and $g \in Hom\left(\mathbb{R}^k, \mathbb{R}^m\right)$ can be represented by matrices $A \in M_{n\times k}(\mathbb{R}), B \in M_{k\times m}(\mathbb{R})$ appropriately, such that for all $x \in \mathbb{R}^n$:

$$(f \circ g)(x) = (A \circ B)(x) = (AB)(x)$$

And the multiplication of matrices AB is also a matrix, and therefore also a linear function $\in Hom(\mathbb{R}^n, \mathbb{R}^m)$. For any affine $f : \mathbb{R}^n \to \mathbb{R}^k$, $g : \mathbb{R}^k \to \mathbb{R}^m$ can be represented as $f(x) = Ax + c$, $g(x) = Bx + d$ where $A \in M_{n\times k}(\mathbb{R}), B \in M_{k\times m}(\mathbb{R})$ and $c \in \mathbb{R}^n$, $d \in \mathbb{R}^k$, such that for any $x \in \mathbb{R}^n$:

$$(f \circ g)(x) = f(Bx + d) = A(Bx + d) + c = (AB)x + (Ad + c)$$

From this form, since AB is a matrix and $Ad + c$ is a vector in the image dimension, we can see that $f \circ g$ is an affine transformation

Question 2:

a)

To find the gradient descent step, we of course need to find the gradient. We know that:

$$\frac{df}{dx} = 20\left(x-1\right), \frac{df}{dy} = \frac{1}{5}\left(y+1\right)$$

Therefore:

$$\nabla f\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 100\left(x-1\right) \\ y+1 \end{pmatrix} = \begin{bmatrix} 20 & \\ & \frac{1}{5} \end{bmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -20 \\ \frac{1}{5} \end{pmatrix}$$

Therefore the gradient descent step from x^k to x^{k+1} is:

$$x^{k+1} = x^k - t\nabla f\left(x^k\right) = x^k - t\left(\begin{bmatrix} 20 & \\ & \frac{1}{5} \end{bmatrix} x^k + \begin{pmatrix} -20 \\ \frac{1}{5} \end{pmatrix}\right)$$

$$= \begin{bmatrix} 1-20t & \\ & 1-\frac{t}{5} \end{bmatrix} x^k + \begin{pmatrix} 20t \\ -\frac{t}{5} \end{pmatrix}$$

Where t is the step size, or the learning reate.

b)

We can see of course that the function reaches a minimum at $x = 1, y = -1$, because it reaches $f\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 0$ and otherwise it is bounded from below by 0.

From the GD step we saw in the last section, we can see that every coordinate of x^k is independent in every step, such that:

$$x_1^{k+1} = 20t + x_1^k\left(1-20t\right)$$

$$x_2^{k+1} = -\frac{t}{5} + x_2^k\left(1-\frac{t}{5}\right)$$

For any series of the form $a_0 = x$ and $a_{k+1} = a_k x + b$, we can see that:

$$a_0 = x$$

$$a_1 = ax + b$$

$$a_2 = a^2x + ab + b$$

$$a_3 = a^3x + a^2b + ab + b$$

We can therefore see recursively that $a_k = a^k x + \frac{a^k - 1}{a-1}b$, as $a_0 = x + \frac{1-1}{a-1}b = x$, and recursively:

$$a_{k+1} = a\left(a^k x + \frac{a^k - 1}{a-1}b\right) = a^{k+1}x + \sum_{i=1}^{k-1} a^i b + b$$

$$= a^{k+1}x + \sum_{i=0}^{k-1} a^i b = a^{k+1}x + \frac{a^{k+1} - 1}{a-1}b$$

Therefore, we can see that:

$$x_1^k = (1 - 20t)^k x_1^0 + \frac{1 - (1-20t)^k}{20t} 20t$$

$$= (1 - 20t)^k \left(x_1^0 - 1\right) + 1$$

And:

$$x_2^k = \left(1 - \frac{t}{5}\right)^k x_2^0 + \frac{\left(1 - \frac{t}{5}\right)^k - 1}{-\frac{t}{5}} \left(-\frac{t}{5}\right)$$

$$= \left(1 - \frac{t}{5}\right)^k \left(x_2^0 + 1\right) - 1$$

Therefore, we can see that these series converge for all x^0 iff $|1 - 20t|, \left|1 - \frac{t}{5}\right| < 1$, and therefore $t < 10, \frac{1}{10}$, or just $t < \frac{1}{10}$.

Therefore, the maximal learning rate is $\frac{1}{10}$, such that when $t = \frac{1}{10}$, x_1^k bounces back and forth between x_1^0 and $1 + \left(1 - x_1^0\right)$, and when $t > \frac{1}{10}$ the series x_1^k is unbounded.

c)

The value 10, as this creates a very large gradient on the x coordinate which requires a small step size in order to converge.

d)

As we saw with the series:

$$x_1^k = (1-20t)^k \left(x_1^0 - 1\right) + 1$$

$$x_2^k = \left(1 - \frac{t}{5}\right)^k \left(x_2^0 + 1\right) - 1$$

In general, the y coordinate takes the longest to converge, as for almost all $t < \frac{1}{10}$ we have $|1-20t| < \left|1 - \frac{t}{5}\right|$, and therefore $(1-20t)^k$ converges faster than $\left(1 - \frac{t}{5}\right)^k$.

Question 3:

If α_p is the predicted angle, and α_r is the real angle, we can use the following loss function to predict the distance between them:

$$L_{\alpha_r}(\alpha_p) = (\sin(\alpha_p) - \sin(\alpha_r))^2 + (\cos(\alpha_p) - \cos(\alpha_r))^2$$

This is a smooth function with a very intuitive geometric meaning: we know that the unit vector with the angle α_p is $\hat{\alpha_p} = (\cos(\alpha_p), \sin(\alpha_p))$, and the unit vector with the angle is α_r is $\hat{\alpha_r} = (\cos(\alpha_r), \sin(\alpha_r))$. The distance between these vectors is roughly equivalent to the distance between the angles, whose square is:

$$\|\hat{\alpha_p} - \hat{\alpha_r}\|_2^2 = (\sin(\alpha_p) - \sin(\alpha_r))^2 + (\cos(\alpha_p) - \cos(\alpha_r))^2 = L_{\alpha_r}(\alpha_p)$$

Question 4:

a)

$$\frac{d}{dx}\left(f\left(x+y,2x,z\right)\right)=\frac{d}{dx}\left(f\circ\left(x+y,2x,z\right)\right)=\left(Df\right)_{(x+y,2x,z)}\begin{pmatrix}1\\2\\0\end{pmatrix}$$

b)

$$\frac{d}{dx}\left(f_1\left(f_2\left(...f_n\left(x\right)\right)\right)\right)=f`_1\left(f_2\left(...f_n\left(x\right)\right)\right)\left(\frac{d}{dx}\left(f_2\left(...f_n\left(x\right)\right)\right)\right)=$$

$$=f`_1\left(f_2\left(...f_n\left(x\right)\right)\right)\left(f_2`\left(...f_n\left(x\right)\right)\right)\left(\frac{d}{dx}\left(f_3\left(...f_n\left(x\right)\right)\right)\right)=...$$

$$...=\left(f_1`\left(f_2\left(...f_n\left(x\right)\right)\right)\right)\left(f_2`\left(...f_n\left(x\right)\right)\right)...\left(f_{n-1}`\left(f_n\left(x\right)\right)\right)\left(f_n\left(x\right)\right)$$

c)

$$\frac{d}{dx}\left(f_1\left(x,f_2\left(x,...f_{n-1}\left(x,f_n\left(x\right)\right)\right)\right)\right)=$$

$$=\frac{d\left(f_1\left(x,f_2\left(x,...f_{n-1}\left(x,f_n\left(x\right)\right)\right)\right)\right)}{d\left(x,f_2\left(x,...f_{n-1}\left(x,f_n\left(x\right)\right)\right)\right)}\frac{d\left(x,f_2\left(x,...f_{n-1}\left(x,f_n\left(x\right)\right)\right)\right)}{dx}=$$

$$=\left(Df_1\right)_{(x,f_2(x,...f_{n-1}(x,f_n(x))))}\begin{pmatrix}1\\\frac{d}{dx}\left(f_2\left(x,...f_{n-1}\left(x,f_n\left(x\right)\right)\right)\right)\end{pmatrix}$$

$$=\left(\left(\nabla f_1\right)_{(x,f_2(x,...f_{n-1}(x,f_n(x))))}\right)_1+$$

$$+\left(\left(\nabla f_1\right)_{(x,f_2(x,...f_{n-1}(x,f_n(x))))}\right)_2\left(\frac{d}{dx}\left(f_2\left(x,...f_{n-1}\left(x,f_n\left(x\right)\right)\right)\right)\right)=...$$

$$...=\left(\left(\nabla f_1\right)_{(x,f_2(x,...f_{n-1}(x,f_n(x))))}\right)_1+\left(\left(\nabla f_1\right)_{(x,f_2(x,...f_{n-1}(x,f_n(x))))}\right)_2\cdot$$

$$(\left(\left(\nabla f_2\right)_{(x,f_3(x,...f_{n-1}(x,f_n(x))))}\right)_1+$$

$$+\left(\left(\nabla f_2\right)_{(x,f_3(x,...f_{n-1}(x,f_n(x))))}\right)_2\left(...+\left(\left(\nabla f_{n-1}\right)_{(x,f_n(x))}\right)_1+\left(\left(\nabla f_{n-1}\right)_{(x,f_n(x))}\right)_2f_n`\left(x\right)\right))$$

d)

$$\frac{d}{dx}\left(f\left(x+g\left(x+h\left(x\right)\right)\right)\right)=\frac{d\left(f\left(x+g\left(x+h\left(x\right)\right)\right)\right)}{d\left(x+g\left(x+h\left(x\right)\right)\right)}\frac{d\left(x+g\left(x+h\left(x\right)\right)\right)}{dx}$$

$$=f‘\left(x+g\left(x+h\left(x\right)\right)\right)\left(1+\frac{d}{dx}g\left(x+h\left(x\right)\right)\right)=$$

$$=f‘\left(x+g\left(x+h\left(x\right)\right)\right)\left(1+\frac{dg\left(x+h\left(x\right)\right)}{d\left(x+h\left(x\right)\right)}\frac{d\left(x+h\left(x\right)\right)}{dx}\right)=$$

$$=f‘\left(x+g\left(x+h\left(x\right)\right)\right)\left(1+g‘\left(x+h\left(z\right)\right)\left(1+h‘\left(x\right)\right)\right)$$