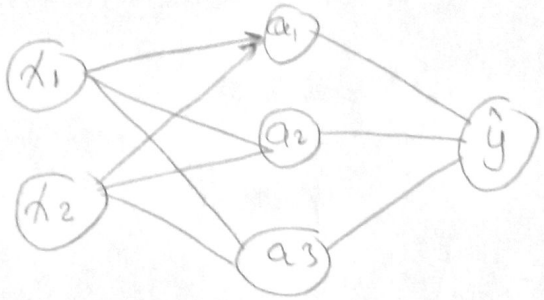
Assignment 1

1



1a) First Layer

$$\underset{(3\times1)}{V^{(1)}} = \left(\underset{(3\times2)}{W^{(1)T}}\ \underset{(2\times1)}{X} + \underset{(3\times1)}{B^{(1)}}\right) = \begin{bmatrix} 2 & 2 \\ 1 & -1 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 8+1 \\ -2 \\ 6-1 \end{bmatrix}$$

$$V^{(1)} = \begin{bmatrix} 9 \\ -2 \\ 5 \end{bmatrix}$$

$$\underset{(1\times1)}{\hat{y}} = \underset{(1\times3)}{W^{(2)}}\ \underset{(3\times1)}{V^{(1)}} + \underset{(1\times1)}{b^{(2)}} = [3 \quad 1 \quad 2]\begin{bmatrix} 9 \\ -2 \\ 5 \end{bmatrix} + 1 = 27-2+10+1$$

$$\hat{y} = 36$$

1b) $Relu(x) = max(0, x)$

$$Relu(V^{(1)}) = \begin{bmatrix} 9 \\ 0 \\ 5 \end{bmatrix}$$

$$\hat{y} = [3 \quad 1 \quad 2]\begin{bmatrix} 9 \\ 0 \\ 5 \end{bmatrix} + 1 = 27 + 10 + 1 = 38$$

$$\hat{y} = 38$$

1c)

$$J = (\hat{y} - y)^2$$

$$\frac{\partial J}{\partial b_1^{(2)}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial b_1^{(2)}} = 2(\hat{y} - y)(1)$$

$$\hat{y} = V_1^{(1)} W_1^{(2)} + V_2^{(1)} W_2^{(2)} + V_3^{(1)} W_3^{(2)} + b^{(2)}$$

$$\boxed{\frac{\partial J}{\partial b_1^{(2)}} = 2(36 - 32) = 8}$$

$$\frac{\partial J}{\partial W_{21}^{(2)}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial W_{21}^{(2)}} = 2(\hat{y} - y) V_2^{(1)}$$

$$\boxed{\frac{\partial J}{\partial W_{21}^{(2)}} = 2(36-32)(-2) = -16}$$

$$\frac{\partial J}{\partial b_2^{[1]}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{V_2^{(1)}} \frac{\partial V_2^{(1)}}{\partial b_2^{(1)}}$$

$$\frac{\partial J}{\partial b_2^{(1)}} = 2(\hat{y} - y)(W_{21}^{(2)})(1) = 2(36-32)(1)(1)$$

$$\boxed{\frac{\partial J}{\partial b_2^{(1)}} = 8}$$

$$\frac{\partial J}{\partial W_{13}^{(1)}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial V_3^{(1)}} \frac{\partial V_3^{(1)}}{\partial W_{13}^{(1)}}$$

$$= 2(\hat{y} - y)(W_{31}^{(2)})(x_1)$$

$$= 2 * 4 * 2 * 3$$

$$\boxed{\frac{\partial J}{\partial W_{13}^{(1)}} = 48}$$

(1d)

$$b_2^{(1)}{}_{n+1} = b_2^{(1)}{}_{n} - \lambda \frac{\partial J}{\partial b_2^{(1)}{}_{n}}$$

$$b_2^{(1)}{}_{n+1} = 0 - 2(8)$$

$$\boxed{b_2^{(1)}{}_{n+1} = -16}$$

$$W_{13}^{(1)}{}_{n+1} = W_{13}^{(1)}{}_{n} - \lambda \frac{\partial J}{\partial W_{13}^{(1)}}$$

$$W_{13}^{(1)}\big|_{n+1} = 3 - 2(48)$$

$$\boxed{W_{13}\big|_{n+1} = -93}$$

yes use the test set once to calculate

Question 1 (e)

Some how it will be good to use test set as an indicator to measure out of sample error, but data set could be correlated some how, sow we need to test on more data sets to validate our model

(2) $f: \mathbb{R}^2 \to \mathbb{R}$

$f = \sin g_1 + g_2^2$

$g: \mathbb{R}^2 \to \mathbb{R}^2$

$g(x) = \begin{bmatrix} x_1 e^{x_2} \\ x_1 + x_2^2 \end{bmatrix}$

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \frac{\partial g}{\partial x}$$

(1×2)

$$\begin{bmatrix} \frac{\partial f}{\partial g_1} & \frac{\partial f}{\partial g_2} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial f}{\partial g_1} \frac{\partial g_1}{\partial x_1} + \frac{\partial f}{\partial g_2} \frac{\partial g_2}{\partial x_1} & \frac{\partial f}{\partial g_1} \frac{\partial g_1}{\partial x_2} + \frac{\partial f}{\partial g_2} \frac{\partial g_2}{\partial x_2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial f}{\partial x_1} = (\cos g_1)(e^{x_2}) + (2g_2)(1)$$

$$\frac{\partial f}{\partial x_2} = (\cos g_1)(x_1 e^{x_2}) + (2g_2)(2x_2)$$

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{m} \frac{\partial f}{\partial g_j} \frac{\partial g_j}{\partial x_i}$$

(3)

(3.1)

$$f = \frac{1}{1+e^{-z}}$$

$$\frac{df}{dz} = \frac{-1}{(1+e^{-z})^2} * (-e^{-z})$$

$$= \frac{e^{-z}}{(1+e^{-z})^2} = \left(\frac{e^{-z}}{1+e^{-z}}\right)\left(\frac{1}{1+e^{-z}}\right) = \left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) f(z)$$

$$\boxed{\frac{df}{dz} = (1-f(z))(f(z))}$$

(3.2)

$$f(w) = \frac{1}{1-e^{w^T x}}$$

(1×1) $(1\times D)$ $(D\times1)$

$$w^T x = w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$$

$$\boxed{\frac{\partial f}{\partial w_i} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial w_i} = \left[(f(w))(1-f(w))\right][x_i]}$$

$$\frac{df}{dw} = \left[\frac{\partial f}{\partial w_1} \quad \frac{\partial f}{\partial w_2} \quad \cdots \quad \frac{\partial f}{\partial w_D}\right]$$

$$\boxed{\frac{df}{dw} = x^T \left[(f(w))(1-f(w))\right]}$$

$\frac{df}{dw}$: $(1\times D)$ $\quad x^T$: $(1\times D)$

(3.3) $J(w) = \frac{1}{2}\sum_{i=1}^{m} \left| w^T x^{(i)} - y^{(i)} \right|$

(1×1) $(1\times D)$ $(D\times1)$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \\ w_D \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ x \\ x_D \end{bmatrix}$$

$$w^T x = w_1 x_1 + w_2 x_2 + \dots + w_D x_D$$

$$\frac{dJ}{dw} = \begin{bmatrix} \frac{\partial J}{\partial w_1} & \frac{\partial J}{\partial w_2} & \dots & \frac{\partial J}{\partial w_D} \end{bmatrix}$$

$(1\times D)$

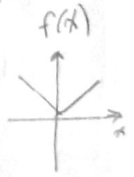
$$\frac{\partial J}{\partial w_1} = \frac{1}{2}\sum_{i=1}^{m} \text{sign}\left(w^T x^{(i)} - y^{(i)}\right)\left(x_1^{(i)}\right)$$

$$\frac{dJ}{dw} = \frac{1}{2}\sum_{i=1}^{m} \underbrace{\text{sign}\left(w^T x^{(i)} - y^{(i)}\right)}_{(1\times1)} \underbrace{\left(x^{(i)T}\right)}_{(1\times D)}$$

$(1\times D)$

Side note

$f(x) = |x| = \sqrt{x^2}$



$$\frac{df}{dx} = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$$

$f(x) = \sqrt{x^2}$

$\frac{df}{dx} = \frac{1}{2}(x^2)^{-1/2} * 2x$

$$\frac{df}{dx} = \frac{x}{\sqrt{x^2}} = \frac{x}{|x|}$$

$\frac{df}{dx} = \text{sign}(x)$

$x \neq 0$

$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \end{cases}$$

(3.4)

$$J(w) = \frac{1}{2}\left[\sum_{i=1}^{m}\left(w^T x^{(i)} - y^{(i)}\right)^2\right] + \lambda \|w\|_2^2$$

$J(w)$: (1×1); $w^T x^{(i)}$: $(1\times D)\,(D\times 1)$; $y^{(i)}$: (1×1); $\|w\|_2^2$: (1×1)

$$\frac{dJ}{dw} = \left[\frac{dJ}{dw_1} \quad \frac{dJ}{dw_2} \quad + \quad \cdots\cdots \quad + \quad \frac{dJ}{dw_D}\right]$$

$(1\times D)$

$$\frac{dJ}{dw_k} = \left[\sum_{i=1}^{m}\left(w^T x^{(i)} - y^{(i)}\right) x_k^{(i)}\right] + 2\lambda w_k$$

$$\frac{dJ}{dw} = \sum_{i=1}^{m}\left(\left(w^T x^{(i)} - y^{(i)}\right) x^{(i)T}\right) + 2\lambda w^T$$

$\frac{dJ}{dw}$: $(1\times D)$; $\left(w^T x^{(i)} - y^{(i)}\right)$: $(1\times D)$; $x^{(i)T}$: $(1\times D)$; $2\lambda w^T$: $(1\times D)$

3.5

$$J(w) = \sum_{i=1}^{m} y^{(i)} \log\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(1-\frac{1}{1+e^{-w^T x^{(i)}}}\right)$$

$J(w)$: (1×1)

$$\frac{dJ}{dw} = \begin{bmatrix} \frac{dJ}{aw_1} & \frac{aJ}{aw_2} + \cdots & \frac{aJ}{aw_D} \end{bmatrix}$$

$(1\times D)$

Side note

$\log_e \equiv \log$

$\frac{d}{dx}(\log x) = \frac{1}{x}$

let $\sigma(x) = \frac{1}{1+e^{-x}} \rightarrow \frac{d\sigma(x)}{dx} = (1-\sigma(x))(\sigma(x))$

$$J(w) = \sum_{i=1}^{m} y^{(i)} \log\left(\sigma(w^T x^{(i)})\right) + (1-y^{(i)}) \log\left(1-\sigma(w^T x^{(i)})\right)$$

$$\frac{aJ}{aw_1} = \sum_{i=1}^{m} (y^{(i)}) \left(\frac{1}{\sigma(w^T x^{(i)})}\right) (1-\sigma(w^T x^{(i)})) (\sigma(w^T x^{(i)})) (x_1^{(i)})$$

$$+ (1-y^{(i)}) \left(\frac{1}{1-\sigma(w^T x^{(i)})}\right) (-1) (1-\sigma(w^T x^{(i)})) (\sigma(w^T x^{(i)})) (x_1^{(i)})$$

$$\frac{aJ}{aw_1} = \sum_{i=1}^{m} \left[y^{(i)} (1-\sigma(w^T x^{(i)}) - (1-y^{(i)}) (\sigma(w^T x^{(i)})) \right] x_1^{(i)}$$

$$\frac{dJ}{dw} = \sum_{i=1}^{m} \left(\left[y^{(i)} (1-\sigma(w^T x^{(i)}) - (1-y^{(i)}) (\sigma(w^T x^{(i)})) \right] x^{(i)T} \right)$$

$\frac{dJ}{dw}$: $(1\times D)$; $x^{(i)T}$: $(1\times D)$; $(1\times D)$

3.6

$$f(w) = \tanh[w^T x]$$
$$(1\times 1) \quad (1\times D) \quad (D\times 1)$$

$$\nabla_w f = \left[\frac{\partial f}{\partial w_1} \quad \frac{\partial f}{\partial w_2} \quad \cdots \quad \frac{\partial f}{\partial w_D}\right]$$

$$\frac{\partial f}{\partial w_1} = 4(1-6(2w^T x))(6(2w^T x))(x_1)$$

$$\nabla_w f = \underset{(1\times D)}{x^T} \underbrace{\left[4(1-6(2w^T x))(6(2w^T x))\right]}_{(1\times 1)}$$
$$(1\times D)$$

Side note

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

$$= \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

$$= -\frac{-2 + 1 + e^{-2x}}{1 + e^{-2x}}$$

$$\frac{2}{1+e^{-2x}} - 1$$

$$\tanh(x) = 2\sigma(2x) - 1$$

$$\frac{d}{dx}\tanh(x) = 2(1-\sigma(2x))(\sigma(2x))(2)$$

$$= 4(1-\sigma(2x))(\sigma(2x))$$

Assignment 1CSE 616

Abdullah Aml, 2101398

github llink for the project

I have trained two data sets: Iris,and DryBean.

The code is:

- customized to run generic number of hidden layers
- creates a new directory to save: plots, scores, lossses, weights in “train-dryBean” directory

Iris

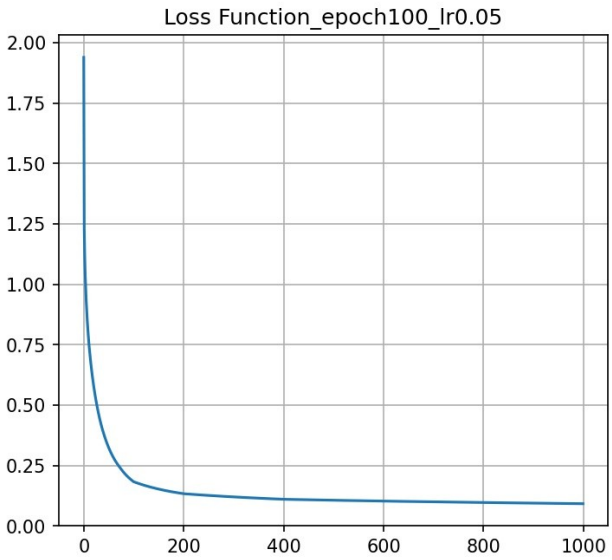
Very simple dataset I used to learn and build working model step by step: it has 4 attributes and 150 sample in total with 3 classes:

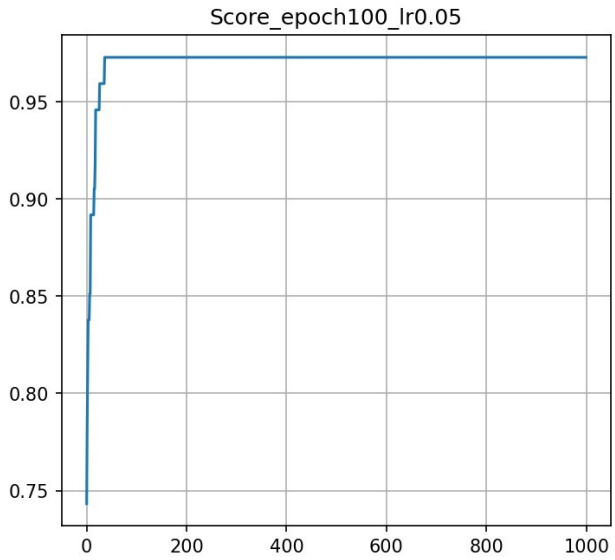
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

I used:

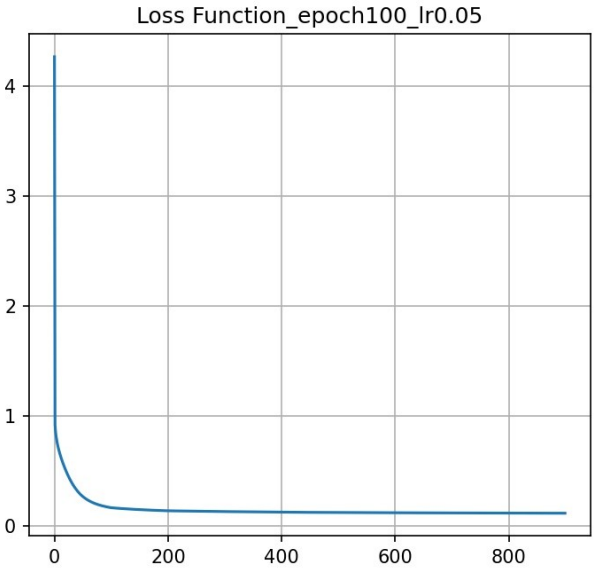
- a single hidden layer with 10 neurons
- LeakRelu in the middle and sigmoid at the last layer
- cross entropy loss

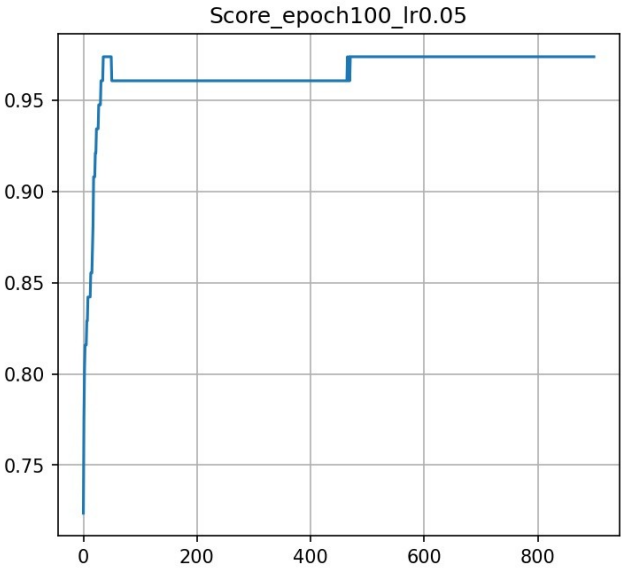
Learning Rate Sequence	Loss	Epocs	Average Score	Top Score
.4, .2, .1, 6*[.05]	0.091728	1000	0.970635	0.972972
.4, .2, 3*[.1], 4*[.05]	0.112709	900	0.964006	0.973684
.4, .2, .1	.0095086	900	0.949541	0.96
.4, .2, .1, 3*[.05]	0.060299	600	0.94444	.9692308



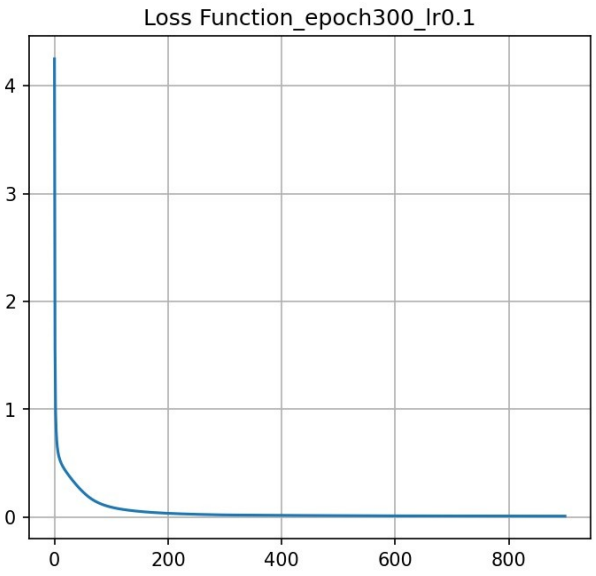


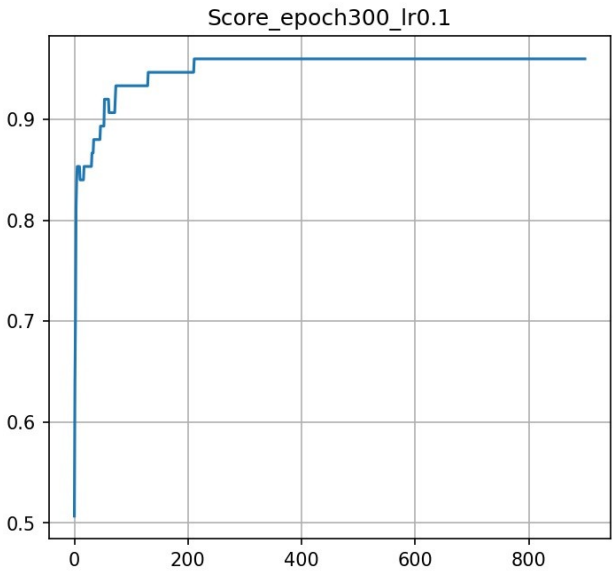
Learning rates=.4, .2, .1, 6*[.05]



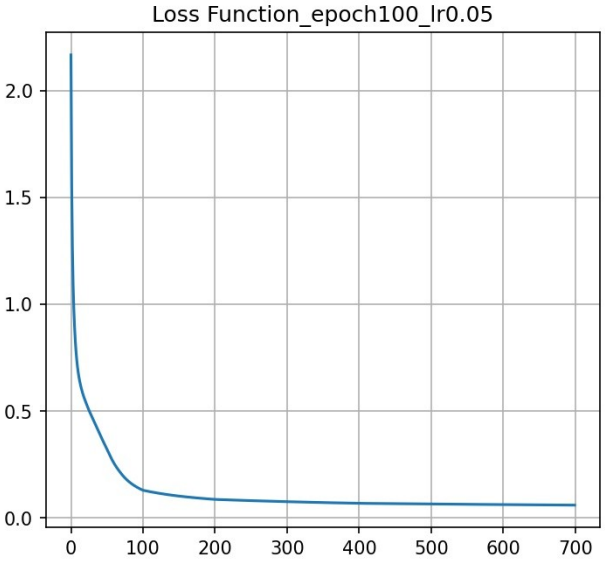


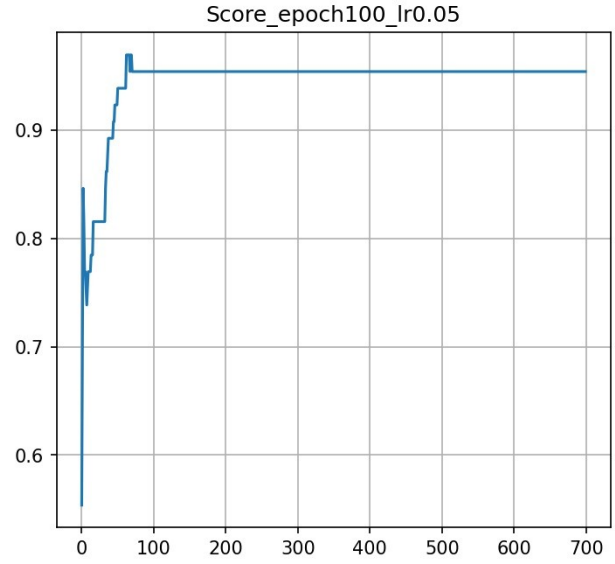
Learning rates=.4, .2, 3*[.1], 4*[.05]





Learning rates=.4, .2, .1





Learning rates=.4, .2, .1, 3*[.05]

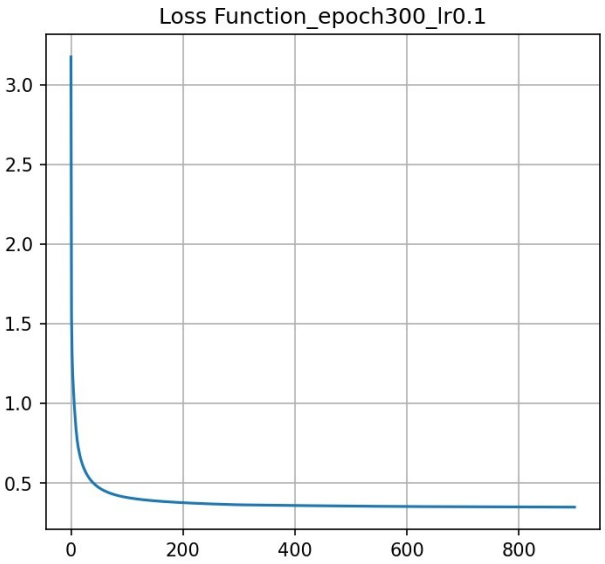
DryBean

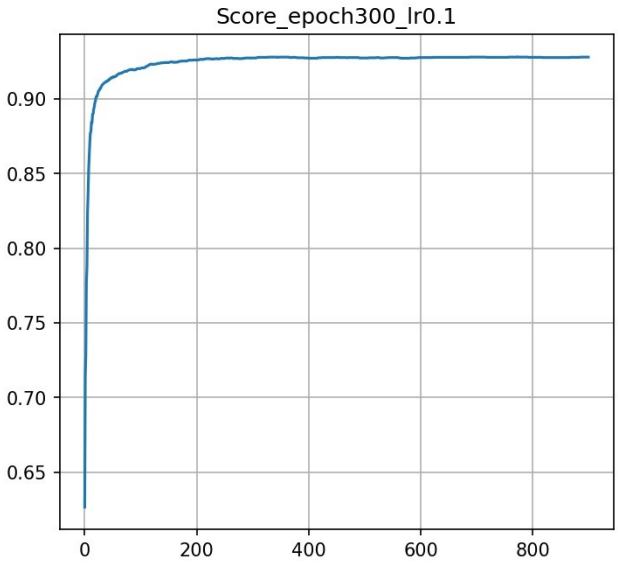
The data-set has 16 attributes and 13611 sample in total with 7 classes: (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira)

I used:

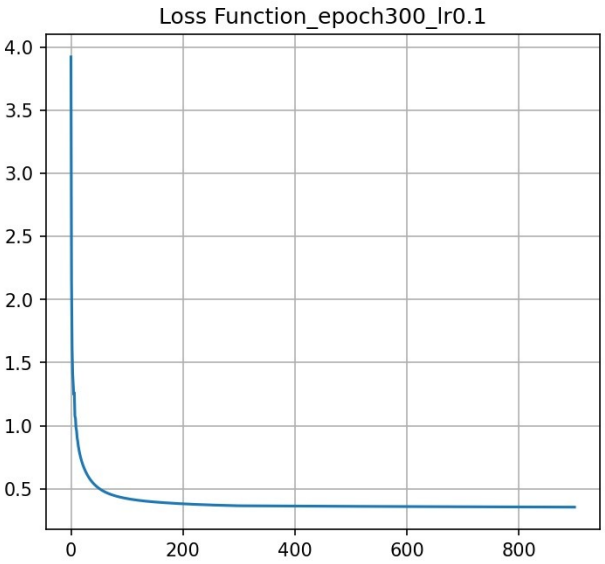
- a single hidden layer with 50 neurons
- LeakRelu in the middle and sigmoid at the last layer
- cross entropy loss

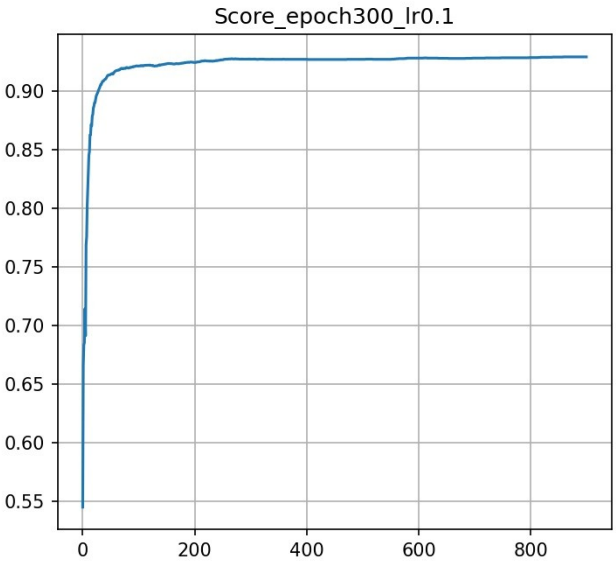
Learning Rate Sequence	Loss	Epocs	Average Score	Top Score
.4, .2, .1	0.348955	900	0.92415	0.92817
.4, .1, .1	0.3558576	900	0.92283	0.92923
.4, .3, .1, .01	0.347872	1200	0.92005	0.927168
.4, .2, .1, .05	0.334702	1200	0.92313	0.926969



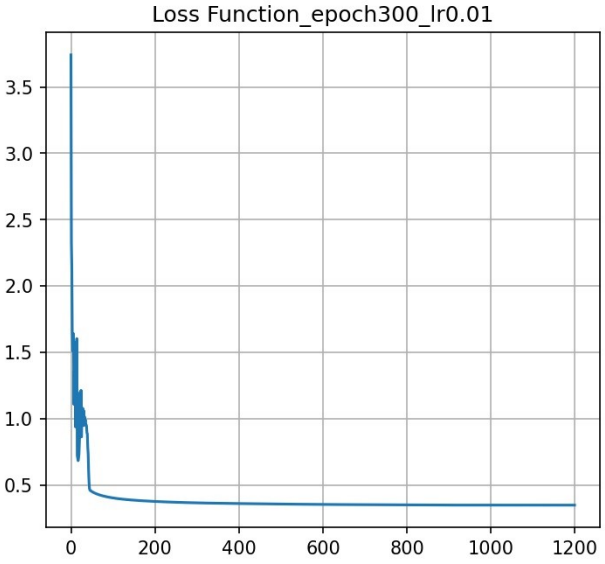


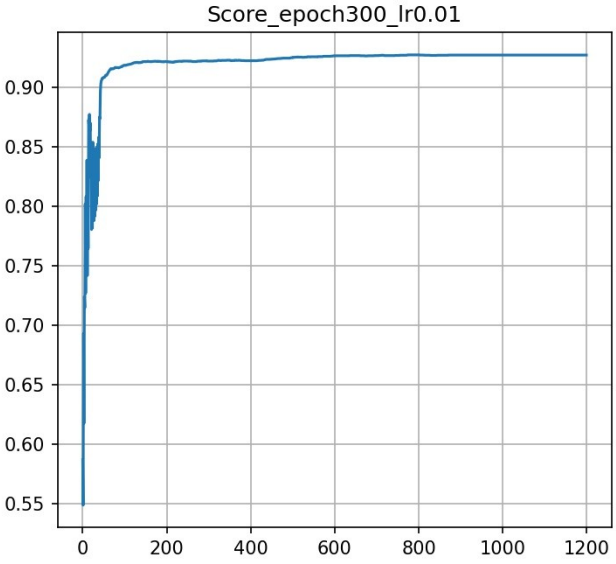
Learning rate = .4, .2, .1



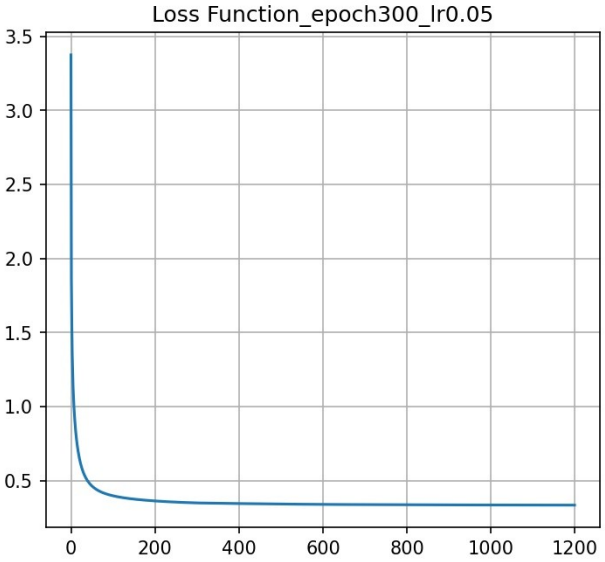


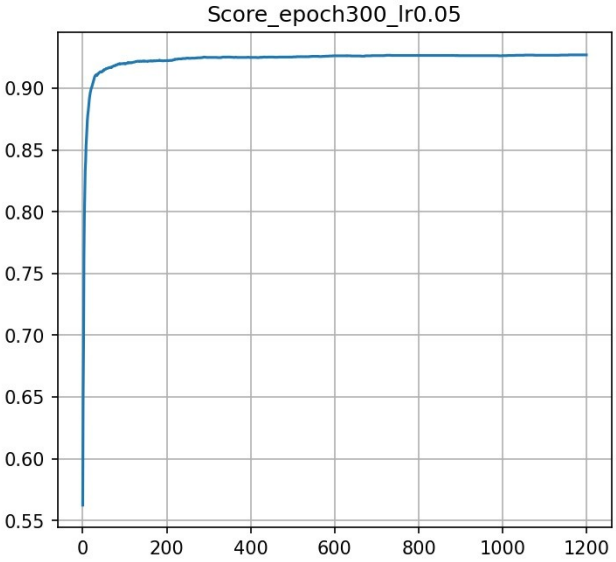
learning rate = .4, .1, .1





learning rate = .4, .3, .1, .01





learning rate = .4, .2, .1, .05