CMPT419 Assignment 2

Yu Ke
(301414915)

1. Softmax for Multi-Class Classification

1) Because the green point is the common point, so

$$p(C_1|x) = p(C_2|x) = p(C_3|x) = \frac{1}{3}$$

2) The propabilities alone each of red lines are the same for two neighbor regions.

When moving along a red line away from the green point, the propabilities of two neighbor region are close to $\frac{1}{2}$, and the third's propability becomes close to 0.

3) As we move far away from the intersection point, staying in middle of one region, the value becomes bigger and bigger and close to 1, the other two will be close to 0.

2. Error Backpropagation

① In finial output node : $h(a) = a$

$$\Rightarrow \quad a_1^{(3)} = y_1^{(3)}$$

$$\delta_1^{(3)} = \frac{\partial E_n(w)}{\partial a_1^{(3)}} = \frac{\partial}{\partial a_1^{(3)}} \cdot \frac{1}{2}(a_1^{(3)} - t_n)^2 = a_1^{(3)} - t_n$$

$$\frac{\partial E_n(w)}{\partial w_{12}^{(3)}} = \frac{\partial E_n(w)}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial w_{12}^{(3)}} = \delta_1^{(3)} \cdot \frac{\partial a_1^{(3)}}{\partial w_{12}^{(3)}}$$

$$\because \quad a_1^{(3)} = w_{11}^{(3)} z_1^{(2)} + w_{12}^{(3)} \cdot z_2^{(2)} + w_{13}^{(3)} \cdot z_3^{(2)}$$

$$\therefore \quad \frac{\partial E_n(w)}{\partial w_{12}^{(3)}} = \delta_1^{(3)} \cdot z_2^{(2)} = z_2^{(2)} \cdot (a_1^{(3)} - t_n)$$

② $$\frac{\partial E_n(w)}{\partial a_1^{(2)}} = \frac{\partial E_n(w)}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial a_1^{(2)}}$$

$$= \delta_1^{(3)} \cdot \frac{\partial}{\partial a_1^{(2)}} \left(w_{11}^{(3)} z_1^{(2)} + w_{12}^{(3)} \cdot z_2^{(2)} + w_{13}^{(3)} \cdot z_3^{(2)} \right)$$

$$= \delta_1^{(3)} \cdot w_{11}^{(3)} \cdot h'(a_1^{(2)})$$

$$\bullet \quad \frac{\partial E_n(w)}{\partial w_{11}^{(2)}} = \frac{\partial E_n(w)}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial w_{11}^{(2)}}$$

$$\because\ a_1^{(2)} = W_{11}^{(2)} z_1^{(1)} + W_{12}^{(2)} \cdot z_2^{(1)} + W_{13}^{(2)} \cdot z_3^{(1)}$$

$$\therefore\ \frac{\partial a_1^{(2)}}{\partial W_{11}^{(2)}} = z_1^{(1)}$$

$$\Rightarrow \frac{\partial E_n(W)}{\partial W_{11}^{(2)}} = \delta_1^{(3)} \cdot W_{11}^{(3)} \cdot h'(a_1^{(2)}) \cdot z_1^{(1)}$$

③ $$\frac{\partial E_n(W)}{\partial a_1^{(1)}} = \delta_1^{(1)} = \sum_{k=1}^{3} \frac{\partial E_n(W)}{\partial a_k^{(2)}} \cdot \frac{\partial a_k^{(2)}}{\partial a_1^{(1)}}$$

$$= \sum_{k=1}^{3} \delta_k^{(2)} \cdot \frac{\partial a_k^{(2)}}{\partial a_1^{(1)}} = h'(a_1^{(1)}) \cdot \sum_{k=1}^{3} \left(W_{k1}^{(1)} \cdot \delta_k^{(2)} \right)$$

$$\frac{\partial E_n(W)}{\partial W_{11}^{(1)}} = \frac{\partial E_n(W)}{\partial a_1^{(1)}} \cdot \frac{\partial a_1^{(1)}}{\partial W_{11}^{(1)}}$$

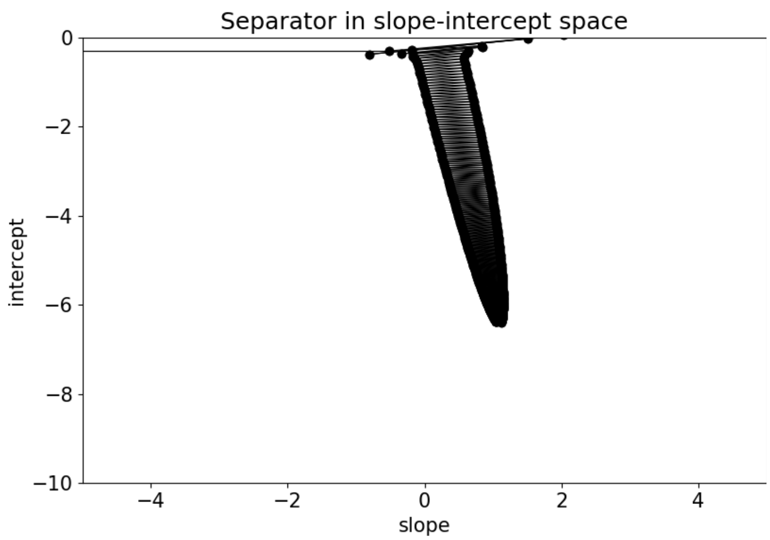
$$= \delta_1^{(1)} \cdot \frac{\partial}{\partial W_{11}^{(1)}} \left(W_{11}^{(1)} x_1 + W_{12}^{(1)} x_2 + W_{13}^{(1)} x_3 \right)$$

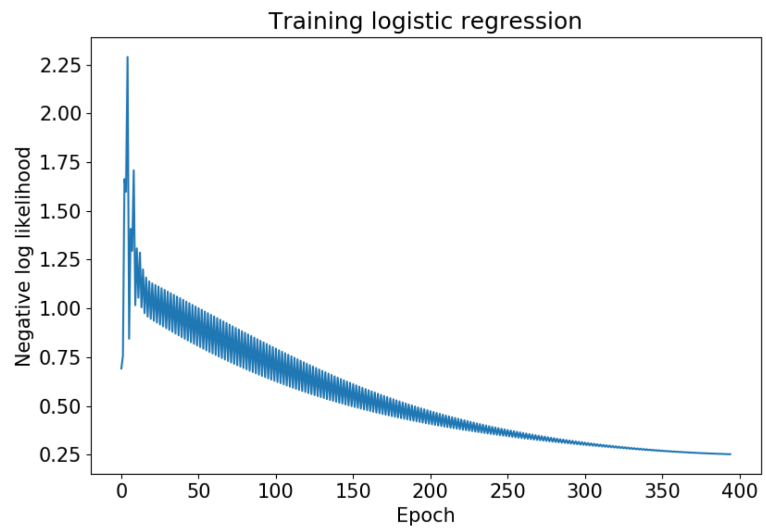
$$= \delta_1^{(1)} \cdot x_1$$

$$= x_1 \cdot h'(a_1^{(1)}) \cdot \sum_{k=1}^{3} \left(W_{k1}^{(1)} \cdot \delta_k^{(2)} \right)$$

3 Logistic Regression

1）

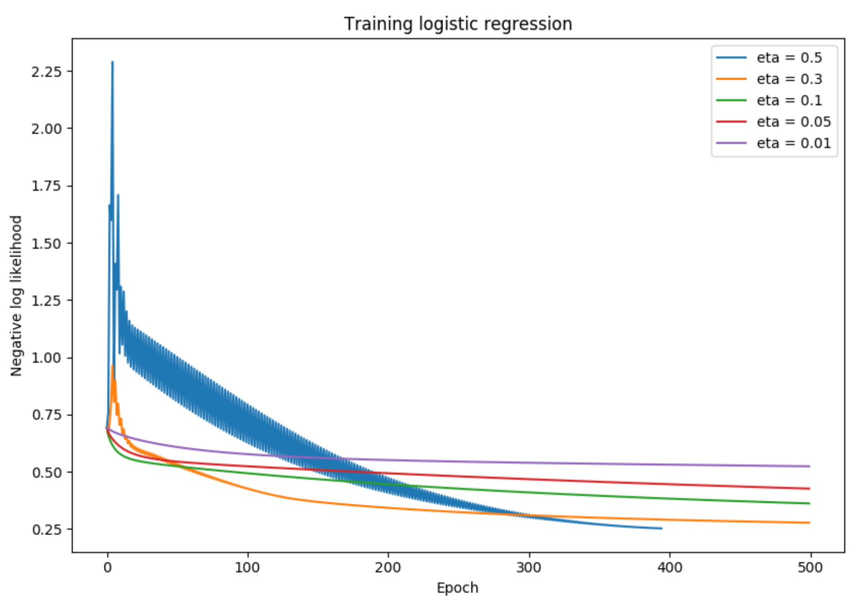




These plots are oscillating because the eta=0.5, which is relatively large

The curve as a whole shows a downward trend, but a large value of eta will cause the error in some areas to become larger, which will cause oscillation

2)



Compared to eta=0.5, eta=[0.3, 0.1, 0.05, 0.01] causes less oscillation because for each step, the gradient decent is smaller. On the other hand, eta=0.5 has its advantage. The negative log likelihood can become small faster than eta<0.5.

Above all, there is no good or bad value of eta, and it depends on your goal.

3)

According to the plot, the stochastic gradient descent faster than gradient, but it has more oscillation

4 Fine-Tuning a Pre-Trained Network

See details in file: ***P4/README.md***