Problem Set 3 for Machine Learning 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 Neural Networks

For convenience, we can denote the nodes in the first layer as n_1, n_2, the second layer n_3.

1.1 Neural network for regression

1.1.1. Simulating linear regression

In this scenario, every node should use S.

$$n_1 = c(x_1w_1 + x_2w_3), \quad n_2 = c(x_1w_2 + x_2w_4) \tag{1}$$

$$y = cn_3 = c(n_1w_5 + n_2w_6) \tag{2}$$

$$= c^2((x_1w_1 + x_2w_3)w_5 + (x_1w_2 + x_2w_4)w_6) \tag{3}$$

$$= c^2(w_1w_5 + w_2w_6)x_1 + c^2(w_3w_5 + w_4w_6)x_2 \tag{4}$$

$$\beta_1 = c^2(w_1w_5 + w_2w_6), \quad \beta_2 = c^2(w_3w_5 + w_4w_6) \tag{5}$$

1.1.2. Derive β_1 and β_2

In this scenario, n_1 and n_2 should use L, and n_3 should use S.

$$n_1 = c(x_1w_1 + x_2w_3), \quad n_2 = c(x_1w_2 + x_2w_4) \tag{6}$$

$$p(n_3 = 1 \mid X) = \frac{1}{1 + exp(-(n_1w_5 + n_2w_6))} \tag{7}$$

$$= \frac{exp(n_1w_5 + n_2w_6)}{1 + exp(n_1w_5 + n_2w_6)} \tag{8}$$

According to the defination of S, we know that the Y = 1 when n_3 = 1, and Y = -1 when n_3 = 0. Therefore, we can derive that:

$$P(Y = 1 \mid X) = p(n_3 = 1 \mid X) = \frac{exp(n_1w_5 + n_2w_6)}{1 + exp(n_1w_5 + n_2w_6)} \tag{9}$$

$$\beta_1 = c(w_1w_5 + w_2w_6), \quad \beta_2 = c(w_3w_5 + w_4w_6) \tag{10}$$

1.1.3. Derive α_1 and α_2

In this scenario, n_1 and n_2 should use S, and n_3 should use L. For f1 and f2, they employ the same distribution as 1.1.2, so:

$$p(Y_1 = 1 \mid X, f1) = p(n_1 = 1 \mid x) = \frac{exp(w_1 x_1 + w_3 x_2)}{1 + exp(w_1 x_1 + w_3 x_2)} \tag{11}$$

$$p(Y_2 = 1 \mid X, f2) = p(n_2 = 1 \mid x) = \frac{exp(w_2 x_1 + w_4 x_2)}{1 + exp(w_2 x_1 + w_4 x_2)} \tag{12}$$

For n_3 and Y, we have:

$$n_3 = n_1 w_5 + n_2 w_6, \quad y = c n_3 = sign(\alpha_1 Y_1 + \alpha_2 Y_2) \tag{13}$$

$$\alpha_1 = c w_5, \quad \alpha_2 = c w_6 \tag{14}$$

1.2 Convolutional Neural Networks

1.3 Gradient vanishing/explosion

1.3.1 Derive b1, the first layer bias

We can derive the derivative of L w.r.t. b1 using the chain rule:

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_m} \frac{\partial z_m}{\partial z_{m-1}} \cdots \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial b_1} \tag{15}$$

$$= \frac{\partial L}{\partial z_m} \cdot \sigma'(w_1 x + b_1) \cdot \prod_{k=1}^{m} \frac{\partial z_{k+1}}{\partial z_k} \tag{16}$$

1.3.2. (a) explain vanish trend

As the activation function is given, we can derive:

$$\frac{\partial z_{k+1}}{\partial z_k} = \frac{\partial \sigma(w_k z_k + b_k)}{\partial (w_k z_k + b_k)} \cdot \frac{\partial (w_k z_k + b_k)}{\partial z_k} \tag{17}$$

$$= \sigma(w_k z_k + b_k)(1 - \sigma(w_k z_k + b_k)) \cdot w_k \tag{18}$$

As we know, $\sigma(x)$ and $1 - \sigma(x)$ is smaller than 1, and $|w_k|$ is smaller than 1. Therefore, according what we derived from 1.3.1, we would konw that the $\frac{\partial L}{\partial b_1}$ tends to vanish when m is large.

1.3.2. (b) explain vanish trend given large $|w|$

Given the form of the $\frac{\partial z_{k+1}}{\partial z_k}$, we can simple consider this function:

$$f_{temp} = \frac{x \cdot exp(x)}{(1 + exp(x))^2} \tag{19}$$

which is the same "form" of the derivative. As we can see, with increase of x, the f_{temp} would become smaller and would be smaller than 1. The trend of $\frac{\partial z_{k+1}}{\partial z_k}$ is the same. When the $|w|$ is large, the total $\frac{\partial z_{k+1}}{\partial z_k}$ would still be smaller than 1, because the decreasing trend of $\sigma(w_k z_k + b_k)(1 - \sigma(w_k z_k + b_k))$ would be more influential than the increasing trend of w_k.

1.3.2. (b) explain vanish trend given large $|w|$

1.3.3. Explain the ReL

1.3.4. Prove the equation

From the equation, we can derive:

$$\frac{\partial logp(v)}{\partial\theta_i} = \frac{\partial log \sum_h p(v,h)}{\partial\theta_i} \tag{20}$$

$$log \sum_h p(v,h) = log \sum_h exp(\sum_i \theta_i\phi_i(v,h)) - log(\sum_{v,h} exp(\sum_i \theta_i\phi_i(v,h))) \tag{21}$$

For simplicity, we can use ϕ_i to replace $\phi_i(v,h)$, the first part and second part:

$$\frac{\partial log \sum_h p(v,h)}{\partial\theta_i} = \frac{1}{\sum_h exp(\sum_i \theta_i\phi_i)} \cdot \frac{\partial \sum_h exp(\sum_i \theta_i\phi_i)}{\partial\theta_i} \cdot \tag{22}$$

$$= \frac{1}{\sum_h exp(\sum_i \theta_i\phi_i)} \cdot \sum_h exp(\sum_i \theta_i\phi_i) \cdot \phi_i \tag{23}$$

$$= \sum_h \phi_i \frac{exp(\sum_i \theta_i\phi_i)}{\sum_h exp(\sum_i \theta_i\phi_i)} \tag{24}$$

Using bayes rule, we can derive the first part as:

$$p(h \mid v) = \frac{p(v,h)}{p(v)} = \frac{p(v,h)}{\sum_h p(v,h)} \tag{25}$$

$$= \frac{exp(\sum_i \theta_i\phi_i)}{\sum_h exp(\sum_i \theta_i\phi_i)} \tag{26}$$

$$\frac{\partial log \sum_h p(v,h)}{\partial\theta_i} = \sum_h \phi(v,h)p(h \mid v) \tag{27}$$

We can similarily derive the second part:

$$\frac{\partial \sum_{v,h} exp(\sum_i \theta_i\phi_i)}{\partial\theta_i} = \frac{1}{Z} \cdot \frac{\partial \sum_{v,h} exp(\sum_i \theta_i\phi_i)}{\partial\theta_i} \tag{28}$$

$$= \sum_{v,h} \phi_i \frac{exp(\sum_i \theta_i\phi_i)}{Z} \tag{29}$$

$$= \sum_{v,h} \phi_i p(v,h) \tag{30}$$

Combine the two parts, we can get the conclusion:

$$\frac{\partial logp(v)}{\partial\theta_i} = \sum_h \phi_i(v,h)p(h \mid v) - \sum_{v,h} \phi_i(v,h)p(v,h) \tag{31}$$

2 Regularized Linear Regression Using Lasso

The visualization is as follows:

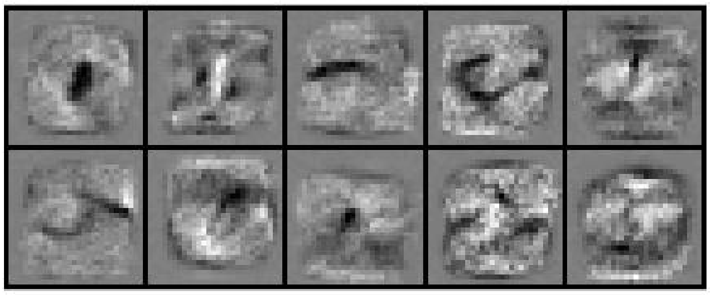

Figure 1: visualization

3 Collaboration

I dicussed with Zheng Chen with problem 2 on understanding finding the minimal for a quardratic function. And discussed with him on question 4 about using the cos value. And double checked the question 5 implementation.