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 L.

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

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

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

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

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

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, n_2, and n_3 all should use S. 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:

$$n_3 = \frac{1}{1 + exp(n_1 w_5 + n_2 w_6)}, \quad y = sign(\alpha_1 Y_1 + \alpha_2 Y_2) \tag{13}$$

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

1.2 Convolutional Neural Networks

1.2.1 Parameters

For the convolution layers, we have:

6 * (25 + 1) + (6 * 25 + 1) * 16 = 2572

For the fully connected layers, we have:

(1 + 16 * 5 * 5) * 120 + (120 + 1) * 84 + (84 + 1) * 10 = 59134

1.2.2 Derive the gradients

The gradient is:

$$\frac{\partial L}{\partial w_i^{(q)}} = \sum_{i=1}^{J^{(q)}} \frac{\partial L}{\partial z_j^{(q)}} \frac{\partial z_j^{(q)}}{\partial w_i^{(q)}} \tag{15}$$

$$= \sum_{i=1}^{J^{(q)}} \frac{\partial L}{\partial z_j^{(q)}} \sigma'(\sum_i w_i^{(q)} x_{ji}^{(q)}) x_{ji}^{(q)} \tag{16}$$

Here σ is the activation function.

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:

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_m} \frac{\partial z_m}{\partial z_{m-1}} \cdots \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial b_1} \tag{17}$$

$$= \frac{\partial L}{\partial z_m} \cdot \sigma'(w_1 x + b_1) \cdot \prod_{k=1}^{m} \frac{\partial z_{k+1}}{\partial z_k} \tag{18}$$

1.3.2. (a) explain vanish trend

As the activation function is given, we can derive:

$$\frac{\partial z_{k+1}}{\partial z_k} = \frac{\partial \sigma(w_k z_k + b_k)}{\partial (w_k z_k + b_k)} \cdot \frac{\partial (w_k z_k + b_k)}{\partial z_k} \tag{19}$$

$$= \sigma(w_k z_k + b_k)(1 - \sigma(w_k z_k + b_k)) \cdot w_k \tag{20}$$

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:

$$f_{temp} = \frac{x \cdot exp(x)}{(1 + exp(x))^2} \tag{21}$$

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.

1.3.3. Explain the ReL

We could find that when $x > 0$:

$$\frac{\partial \sigma(x)}{\partial x} = 1 \tag{22}$$

Then in this case, we know that the parameters would vanish as the sigmoid function.

1.3.4. Prove the equation

From the equation, we can derive:

$$\frac{\partial log p(v)}{\partial \theta_i} = \frac{\partial log \sum_h p(v,h)}{\partial \theta_i} \tag{23}$$

$$log \sum_h p(v,h) = log \sum_h exp(\sum_i \theta_i \phi_i(v,h)) - log(\sum_{v,h} exp(\sum_i \theta_i \phi_i(v,h))) \tag{24}$$

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

$$\frac{\partial log \sum_h p(v,h)}{\partial \theta_i} = \frac{1}{\sum_h exp(\sum_i \theta_i \phi_i)} \cdot \frac{\partial \sum_h exp(\sum_i \theta_i \phi_i)}{\partial \theta_i} \cdot \tag{25}$$

$$= \frac{1}{\sum_h exp(\sum_i \theta_i \phi_i)} \cdot \sum_h exp(\sum_i \theta_i \phi_i) \cdot \phi_i \tag{26}$$

$$= \sum_h \phi_i \frac{exp(\sum_i \theta_i \phi_i)}{\sum_h exp(\sum_i \theta_i \phi_i)} \tag{27}$$

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

$$p(h \mid v) = \frac{p(v,h)}{p(v)} = \frac{p(v,h)}{\sum_h p(v,h)} \tag{28}$$

$$= \frac{exp(\sum_i \theta_i \phi_i)}{\sum_h exp(\sum_i \theta_i \phi_i)} \tag{29}$$

$$\frac{\partial log \sum_h p(v,h)}{\partial \theta_i} = \sum_h \phi(v,h) p(h \mid v) \tag{30}$$

We can similarily derive the second part:

$$\frac{\partial \sum_{v,h} exp(\sum_i \theta_i \phi_i)}{\partial \theta_i} = \frac{1}{Z} \cdot \frac{\partial \sum_{v,h} exp(\sum_i \theta_i \phi_i)}{\partial \theta_i} \tag{31}$$

$$= \sum_{v,h} \phi_i \frac{exp(\sum_i \theta_i \phi_i)}{Z} \tag{32}$$

$$= \sum_{v,h} \phi_i p(v,h) \tag{33}$$

Combine the two parts, we can get the conclusion:

$$\frac{\partial log p(v)}{\partial \theta_i} = \sum_h \phi_i(v,h) p(h \mid v) - \sum_{v,h} \phi_i(v,h) p(v,h) \tag{34}$$

2 Support Vector Machines

2.1 Support Vector Regression

2.1.1. write the dual problem

Suppose we have ξ_i ofr each x_i:

$$|f(x) - y| < \epsilon \tag{35}$$

$$-\epsilon < w^T x_i - y_i < \epsilon \tag{36}$$

Let's set:

$$\left| w^T x_i - y_i \right| - \epsilon = \xi_i \tag{37}$$

$$L = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1} \xi_i - \sum_i \alpha_i (\epsilon + \xi_i - y_i + w x_i) - \sum_i \alpha_i^\star (\epsilon + \xi_i + y_i - w x_i) - \sum_i \beta_i \xi_i \tag{38}$$

$$\alpha_i >= 0, \alpha_i^\star >= 0, \beta_i >= 0 \tag{39}$$

We have to satisfy saddle constraints condition, so

$$\partial_w L = w - \sum_i (\alpha_i - \alpha_i^\star) x_i = 0 \tag{40}$$

$$\partial_{\xi_i} L = C - \alpha_i - \alpha_i^\star - \beta_i = 0 \tag{41}$$

Take the the conditions back and we can get the form of dual problem of SVR:

$$maximize: -\frac{1}{2}\sum_{i,j=1}(\alpha_i - \alpha_i^\star)(\alpha_j - \alpha_j^\star)x_i^T x_j - \epsilon\sum_i(\alpha_i + \alpha_i^\star) + \sum_i y_i(\alpha_i - \alpha_i^\star) \tag{42}$$

$$subject \quad to: \sum_i(\alpha_i - \alpha_i^\star) = 0, \qquad \alpha_i, \alpha_i^\star \in [0, C] \tag{43}$$

2.1.2. KKT Conditions

The KKT condition for this problem is:

$$\partial_w L = w - \sum_i(\alpha_i - \alpha_i^\star)x_i = 0 \tag{44}$$

$$\partial_{\xi_i} L = C - \alpha_i - \alpha_i^\star - \beta_i = 0 \tag{45}$$

$$\epsilon + \xi_i - y_i + w^T x_i >= 0 \tag{46}$$

$$\epsilon + \xi_i^\star + y_i - w^T x_i >= 0 \tag{47}$$

$$\alpha_i >= 0, \alpha_i^\star >= 0, \beta_i >= 0 \tag{48}$$

$$\alpha_i(\epsilon + \xi_i - y_i + w^T x_i) = 0 \tag{49}$$

$$\alpha_i^\star(\epsilon + \xi_i^\star + y_i - w^T x_i) = 0 \tag{50}$$

2.1.3. Kernelized SVR

The prediction rule is:

$$w = \sum_i(\alpha_i - \alpha_i^\star)x_i \tag{51}$$

$$f(x) = \sum_i(\alpha_i - \alpha_i^\star)x_i^T x \tag{52}$$

Add we can use kernel to transfer the x, therefore:

$$f(x_j) = \sum_i(\alpha_i - \alpha_i^\star)K(x_i, x_j) \tag{53}$$

2.1.4. Why using SVR primal

We are using the primal of SVM and the dual of SVR, this is because we want to compose the term:

$$\sum_i x_i x \tag{54}$$

Therefore, we could use the kernel trick to implicitly transfer the feature of the data to other spaces.

2.2 Support Kernel Machines

2.2.1.(a) The minimum of the Lagrangian function

Given L, the Lagrangian function, we could derive to optimize the w:

$$\frac{\partial L}{\partial w_j} = \frac{1}{2}\frac{\partial(\sum_j d_j \left\|w_i\right\|_2)^2}{\partial w_j} - \frac{\partial \alpha_i y_i \sum_j w_j^T x_{ji}}{\partial w_j} \tag{55}$$

We can have $\frac{\partial L}{\partial w_j} = 0$. Then we can both times the w_j^T:

$$\frac{w_j d_j \sum_j \left\|w_j\right\|_2 d_j}{\left\|w_j\right\|_2} = \sum_i \alpha_i y_i x_{ji} \tag{56}$$

$$\left\|w_j\right\|_2 d_j \gamma = w_j^T \sum_i \alpha_i y_i x_{ji} \tag{57}$$

2.2.1.(b) Show that $\left\|\sum_i \alpha_i y_i x_{ji}\right\|_2 < d_j\gamma$

In the last section, we have:

$$\frac{w_j d_j \sum_j \left\|w_j\right\|_2 d_j}{\left\|w_j\right\|_2} = \sum_i \alpha_i y_i x_{ji} \tag{58}$$

Therefore, when the w is not 0, we can take the l2-norm of the this two vectors and get:

$$\left\|\frac{w_j d_j \sum_j \left\|w_j\right\|_2 d_j}{\left\|w_j\right\|_2}\right\|_2 = \left\|\sum_i \alpha_i y_i x_{ji}\right\|_2 \tag{59}$$

Because w is a vector, we know that:

$$\left\|\frac{w_j}{\left\|w_j\right\|_2}\right\|_2 = 1 \tag{60}$$

So we have:

$$\left\|w_j\right\|_2 d_j \gamma = w_j^T \sum_i \alpha_i y_i x_{ji} \tag{61}$$

2.2.1.(c) Show that

For the first point, from the previous problem, we would know that:

$$\left\|\sum_i \alpha_i y_i x_{ji}\right\|_2 <= d_j\gamma \tag{62}$$

And we have also proven that, when w is not zero:

$$\left\|\sum_i \alpha_i y_i x_{ji}\right\|_2 = d_j\gamma \tag{63}$$

Then with the $w_j = 0$, we should have:

$$\left\| \sum_i \alpha_i y_i x_{ji} \right\|_2 < d_j \gamma \tag{64}$$

For the second point, we would know that:

$$\frac{w_j d_j \sum_j \|w_j\|_2 d_j}{\|w_j\|_2} = \sum_i \alpha_i y_i x_{ji} \tag{65}$$

$$w_j = \frac{\|w_j\|_2}{d_j \gamma} \sum_i \alpha_i y_i x_j i \tag{66}$$

We could know that $\|w_j\|_2 > 0, d_j > 0, \gamma > 0$, therefore, we could always have a η_i:

$$w_j = \eta_j \sum_i \alpha_i y_i x_{ji}, \eta_j > 0 \tag{67}$$

When $w_j = 0$, we could set $\eta_j = 0$. Therefore, we always have:

$$\eta_j >= 0 \tag{68}$$

2.2.2. Encourage Sparsity From the 2.2.1.(c), we would know that:

$$if \left\| \sum_i \alpha_i y_i x_{ji} \right\|_2 < d_j \gamma, \quad then \quad w_j = 0 \tag{69}$$

We can see that if the result of $y_i x_i \alpha_i$ is smaller than a certain number, which is given by the regularization terms, then the parameter w_j would tend to vanish to minimize the total loss.

2.2.3. Kernelized version For the kernelized version, We could derive the graient for the minimal in the same way:

$$\frac{\partial L}{\partial w_j} = \frac{1}{2} \frac{\partial (\sum_j d_j \|w_i\|_2)^2}{\partial w_j} - \frac{\partial \alpha_i y_i \sum_j w_j^T \phi(x_{ji})}{\partial w_j} \tag{70}$$

$$\frac{w_j d_j \sum_j \|w_j\|_2 d_j}{\|w_j\|_2} = \sum_i \alpha_i y_i \phi(x_{ji)} \tag{71}$$

$$w_j = \frac{\|w_j\|_2}{d_j \gamma} \sum_i \alpha_i y_i \phi(x_j i) \tag{72}$$

Similar with previous problem, we could have:

$$w_j = \eta_j \sum_i \alpha_i y_i \phi(x_{ji}), \eta_j > 0 \tag{73}$$

When $w_j = 0$, we could set $\eta_j = 0$. Therefore, we always have:

$$\eta_j >= 0 \tag{74}$$

3 Collaboration

I dicussed with Zheng Chen with problem 1.2 on how to calculate the parameters, and he also informed me about 2.2.1(b).