CS280 Fall 2018 Assignment 2
Part A

CNNs

Due in class, Nov 02, 2018

Name: Yingying Ma

Student ID: 88678580

1. Linear Regression(10 points)

- Linear regression has the form $E[y|x] = w_0 + \boldsymbol{w}^{\boldsymbol{T}} x$. It is possible to solve for $\boldsymbol{w}$ and w_0 seperately. Show that

$$w_0 = \frac{1}{n}\sum_i y_i - \frac{1}{n}\sum_i x_i^T \boldsymbol{w} = \overline{y} - \overline{x}^T \boldsymbol{w}$$

- Show how to cast the problem of linear regression with respect to the absolute value loss function, $l(h, x, y) = |h(x) - y|$, as a linear program.

Solution.

- Assume the loss function is MSE:

$$L = \sum_i [y_i - (w_0 + \mathbf{w}^T x_i)]^2$$

Compute the derivative of w_0,

$$\frac{\partial L}{\partial w_0} = \sum_i [y_i - (w_0 + \mathbf{w}^T x_i)]$$

Let the derivative be 0, then we get

$$w_0 = \frac{1}{n}\sum_i y_i - \frac{1}{n}\sum_i x_i^T \boldsymbol{w} = \overline{y} - \overline{x}^T \boldsymbol{w}$$

- We need to convert the formula to a linear program problem:

$$\min l(h, x, y) = |h(x) - y| = |w^T x - y|$$

Define a vector $s = (s_1, \ldots, s_m)$. If $|h(x_i) - y_i)| = |w^T x_i - y_i| \geq 0$, let $s_i \geq w^T x_i - y_i$, which is equal to

$$w^T x_i - s_i \leq y_i$$

If $|h(x_i) - y_i)| = |w^T x_i - y_i| \leq 0$, let $s_i \geq -w^T x_i + y_i$, which is equal to

$$-w^T x_i - s_i \leq -y_i$$

The LP problem is:

$$\begin{aligned} \min \quad & s_i \\ \textbf{s.t.} \ & w^T x_i - s_i \leq y_i \\ & -w^T x_i - s_i \leq -y_i \end{aligned}$$

Let $A = [X - I_m; -X - I_m] \in \mathbb{R}^{2m \times (m+d)}$, $v = (w_1, \ldots, w_d, s_1, \ldots, s_m) \in R^{d+m}$, $b = (y_1, \ldots, y_m, -y_1, \ldots, y_m)^T \in R^{2m}$, $c = (0_d; 1_m)$. The LP problem is:

$$\begin{aligned} \min \quad & c^T v \\ \textbf{s.t.} \ & Av \leq b \end{aligned}$$

2. Convolution Layers (5 points)

We have a video sequence and we would like to design a 3D convolutional neural network to recognize events in the video. The frame size is 32x32 and each video has 30 frames. Let's consider the first convolutional layer.

- We use a set of $5 \times 5 \times 5$ convolutional kernels. Assume we have 64 kernels and apply stride 2 in spatial domain and 4 in temporal domain, what is the size of output feature map? Use proper padding if needed and clarify your notation.
- We want to keep the resolution of the feature map and decide to use the dilated convolution. Assume we have one kernel only with size $7 \times 7 \times 5$ and apply a dilated convolution of rate 3. What is the size of the output feature map? What are the downsampling and upsampling strides if you want to compute the same-sized feature map without using dilation?

Note: You need to write down the derivation of your results.

Solution.

- Use padding 2:

$$H_{out} = \frac{32-5}{2} + 1 = 14$$
$$W_{out} = \frac{32-5}{2} + 1 = 14$$
$$T_{out} = \frac{30-5}{4} + 1 = 7$$

- The size of the output feature map is:

$$H_{out} = \frac{32 - 3 \times (7-1) - 1}{1} + 1 = 14$$
$$W_{out} = \frac{32 - 3 \times (7-1) - 1}{1} + 1 = 14$$
$$T_{out} = \frac{30 - 3 \times (5-1) - 1}{1} + 1 = 18$$

Let down-sampling kernel size be $3 \times 3 \times 2$, stride be 4, after down-sampling, the size of the output is:

$$H_{out} = \frac{32+6-7}{4} + 1 = 8$$
$$W_{out} = \frac{32+6-7}{4} + 1 = 8$$
$$T_{out} = \frac{30+4-5}{4} + 1 = 8$$

For up-sampling, the kernel size need to be $6 \times 6 \times 7$

$$H_{out} = \frac{8+12-7}{1} + 1 = 14$$
$$W_{out} = \frac{8+12-7}{1} + 1 = 14$$
$$T_{out} = \frac{8+14-5}{1} + 1 = 18$$

3. Batch Normalization (5 points)

With Batch Normalization (BN), show that backpropagation through a layer is unaffected by the scale of its parameters.

- Show that
$$BN(\mathbf{Wu}) = BN((a\mathbf{W})\mathbf{u})$$
where $\mathbf{u}$ is the input vector and $\mathbf{W}$ is the weight matrix, a is a scalar.

- (Bonus: 5 pts) Show that
$$\frac{\partial BN((a\mathbf{W})\mathbf{u})}{\partial \mathbf{u}} = \frac{\partial BN(\mathbf{Wu})}{\partial \mathbf{u}}$$

Solution.

- $$BN(\mathbf{Wu}) = \frac{\mathbf{Wu} - \mu}{\sqrt{\sigma^2}}$$
where μ is the mean of $\mathbf{Wu}$ and σ^2 is the variation of $\mathbf{Wu}$.
$$\begin{aligned} BN((a\mathbf{W})\mathbf{u}) &= \frac{a\mathbf{Wu} - a\mu}{\sqrt{a^2\sigma^2}} \\ &= BN(\mathbf{Wu}) \end{aligned}$$

- $$\begin{aligned} \frac{\partial BN(\mathbf{Wu})}{\partial \mathbf{u}} &= \frac{\mathbf{W}}{\sqrt{\sigma^2}} \\ \frac{\partial BN((a\mathbf{W})\mathbf{u})}{\partial \mathbf{u}} &= \frac{a\mathbf{W}}{\sqrt{a^2\sigma^2}} \\ &= \frac{\mathbf{W}}{\sqrt{\sigma^2}} \\ &= \frac{\partial BN(\mathbf{Wu})}{\partial \mathbf{u}} \end{aligned}$$