CS280 Fall 2018 Assignment 2
Part A

CNNs

Due in class, Nov 02, 2018

Name: Zhijuan HU

Student ID: 67856754

1. Linear Regression(10 points)

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

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

  Proof. Since $L = \sum_i (w_0 + \mathbf{w}^T x_i - y_i)^2$, let $\nabla_{x_i} L = 0$,

  we have

$$\begin{aligned} nw_0 + \boldsymbol{w}^T\sum_i x_i &= \sum_i y_i \\ w_0 &= \frac{1}{n}\sum_i y_i - \frac{1}{n}\sum_i x_i^T\boldsymbol{w} \\ &= \overline{y} - \overline{x}^T\boldsymbol{w} \end{aligned}$$

  ■

- 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.
  Proof.
  First prove by contradiction that $|c| = \min_{a\geq 0} a$ where $a \geq c$ and $-a \geq c$. Suppose $|c| \neq \min_{a\geq 0}$ where $a \geq c$ and $-a \leq c$, then $\exists b$ where $b = |c|$ and rather $b < a$ or $b > a$. If $b > a$, then $b > a > c$, which can not happen and $b < a$, then $-b > -a \geq -c$, which can not happen as well. Let $a = (a_1, \ldots, a_m)$ where $a_i \geq < w, x > -y_i$ and $< w, x > -y_i \geq -a_i$. Thus $\min_w \sum_{i=1}^m |< w, x > -y_i| = \min_{a_i > 0} \sum_{i=1}^m a_i$. Let $c = [1, \ldots, 1] \in \mathbb{R}^m, v = [< w, x > -y_1, \ldots, < w, x > -y_m] \in \mathbf{R}^m$. Then $\min_{a_i>0} \sum_{i=1}^m < c, a >$ where $a \geq v$ and $-a \leq v$. ■

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.

  From the description above, denote T, H, W, C_1 are the temporal domain, height, width, channel input, for the video (since inputs are RGB, it have there channel); K_t, K_h, K_w, C_2 are the temporal, height, width, channel output, for the kernel. Hence,

$$
\begin{aligned}
inputs &= T \times H \times W \times C_1 \\
&= 30 \times 32 \times 32 \times 3 \\
Kernels &= K_t \times K_h \times K_w \times C_1 \times C_2 \\
&= 5 \times 5 \times 5 \times 3 \times 64
\end{aligned}
$$

  From the equation,

$$
\begin{aligned}
t &= (T - K_t + Pad_t)/stride + 1 \\
h &= (H - K_h + Pad_h)/stride + 1 \\
w &= (W - K_w + Pad_w)/stride + 1
\end{aligned}
$$

  Set, $Pad_t = 3, Pad_h = 1, Pad_w = 1$, we have

$$
\begin{aligned}
t &= \frac{(30 - 5 + 3)}{4} + 1 = 8 \\
h &= \frac{(32 - 5 + 1)}{2} + 1 = 15 \\
w &= \frac{(32 - 5 + 1)}{2} + 1 = 15
\end{aligned}
$$

  hence,

$$
\begin{aligned}
Outputs &= t \times h \times w \times C_2 \\
&= 8 \times 15 \times 15 \times 64
\end{aligned}
$$

- 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.

By using Dilated convolution, we have the equation,

$$
\begin{aligned}
\hat{K} &= 1 + rate \times (K - 1) \\
X_{out} &= (X_{in} - \hat{K} + Pad)/stride + 1
\end{aligned}
$$

Hence,

$$\hat{K}_t = 1 + 3 \times (5 - 1) = 13$$
$$\hat{K}_h = 1 + 3 \times (7 - 1) = 19$$
$$\hat{K}_w = 1 + 3 \times (7 - 1) = 19$$

Set $Pad_t = 12, Pad_h = 18, Pad_w = 18$, we have

$$t = \frac{(30 - 13 + Pad_t)}{1} + 1 = 30$$
$$h = \frac{(32 - 19 + Pad_h)}{1} + 1 = 32$$
$$w = \frac{(32 - 19 + Pad_w)}{1} + 1 = 32$$

Hence,

$$\begin{aligned} Outputs &= t \times h \times w \times C_1 \times C_2 \\ &= 30 \times 32 \times 32 \times 3 \times 1 \end{aligned}$$

the reason for last term 1 is for using one kernel.

If we want to compute the same-sized feature map without using dilation, the downsampling and upsampling stride we can set 1. Since

$$t = 1 + \frac{(30 + Pad_t - 5)}{1} = 30$$
$$h = 1 + \frac{(32 + Pad_h - 7)}{1} = 32$$
$$w = 1 + \frac{(32 + Pad_w - 7)}{1} = 32$$

where $Pad_t = 4, Pad_h = 6, Pad_w = 6$.

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.

  Proof. Since
$$\begin{aligned} BN((\alpha\mathbf{W})\mathbf{u}) &= \frac{(\alpha\mathbf{W})\mathbf{u} - \mathbb{E}[(\alpha\mathbf{W})\mathbf{u}]}{\sqrt{Var[(\alpha\mathbf{W})\mathbf{u}]}} \\ &= \frac{\alpha\mathbf{Wu} - \alpha\mathbb{E}[\mathbf{Wu}]}{\alpha\sqrt{Var[\mathbf{Wu}]}} \\ &= \frac{\mathbf{Wu} - \mathbb{E}[\mathbf{Wu}]}{\sqrt{Var[\mathbf{Wu}]}} \\ &= BN(\mathbf{Wu}) \end{aligned}$$

  ■

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

  Proof. Since
$$BN((\mathbf{Wu})) = BN((\alpha\mathbf{W})\mathbf{u}),$$
we have
$$\frac{\partial BN((a\mathbf{W})\mathbf{u})}{\partial \mathbf{u}} = \frac{\partial BN(\mathbf{Wu})}{\partial \mathbf{u}}$$

  ■