Optimization and Machine Learning, Fall 2023
Homework 5

(Due Thursday, Jan 11 at 11:59pm (CST))

1. [10 points] [Deep Learning Model]

   (a) Consider a 2D convolution layer. Suppose the input size is $4 \times 64 \times 64 \times$ (channel, width, height) and we use **ten** 3×3 (width, height) kernels with 4 channels input and 4 channels output to convolve with it. Set stride = 1 and pad = 1. What is the output size? Let the bias for each kernel be a scalar, how many parameters do we have in this layer? [5 points]

   (b) The convolution layer is followed by a max pooling layer with 2×2 (width, height) filter and stride = 2. What is the output size of the pooling layer? How many parameters do we have in the pooling layer? [5 points]

(a) output size $= \frac{64 - 3 + 2\times 1}{1} + 1 = 64$

\# of parameters $= (3\times3\times4+1) \times 4 \times 10 = 1480$

(b) output size $= 4 \times (\frac{64-2}{2}+1) \times (\frac{64-2}{2}+1) \times 4 = 4\times32\times32\times4$

\# of parameters $= 0$

2. [10 points] Use the k-means++ algorithm and Euclidean distance to cluster the 8 data points into $K = 3$ clusters. The coordinates of the data points are:

$$x^{(1)} = (2,8),\ x^{(2)} = (2,5),\ x^{(3)} = (1,2),\ x^{(4)} = (5,8),$$
$$x^{(5)} = (7,3),\ x^{(6)} = (6,4),\ x^{(7)} = (8,4),\ x^{(8)} = (4,7).$$

Suppose that initially the first cluster centers is $x^{(1)}$.

(a) Perform the k-means++ algorithm to initialize other centers and report the coordinates of the resulting centroids. [3 points]

(b) Calculate the loss function

$$Q(r,c) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} r_{ij}||x^{(i)} - c_j||^2, \tag{1}$$

where $r_{ij} = 1$ if $x^{(i)}$ belongs to the j-th cluster and 0 otherwise. [2 points]

(c) How many more iterations are needed to converge? [3 points] Calculate the loss after it converged. [2 points]

(a) Distance to $X^{(1)}$: $X^{(2)} = \sqrt{(2-2)^2+(5-8)^2} = 3$ $\quad X^{(3)} = \sqrt{(1-2)^2+(2-8)^2} = \sqrt{37}$

$X^{(4)} = \sqrt{(5-2)^2+(8-8)^2} = 3$ $\quad X^{(5)} = \sqrt{(7-2)^2+(3-8)^2} = 5\sqrt{2}$

$X^{(6)} = \sqrt{(6-2)^2+(4-8)^2} = 4\sqrt{2}$ $\quad X^{(7)} = \sqrt{(8-2)^2+(4-8)^2} = 2\sqrt{13}$

$X^{(8)} = \sqrt{(4-2)^2+(7-8)^2} = \sqrt{5}$

since $X^{(7)}$ is the biggest, it has the biggest probabilities of being selected as the next centroid.

Then calculate the distance of each point to the nearest centroid $X^{(1)}$ and $X^{(7)}$

$X^{(3)}$ is the biggest.

So the points for $K=3$ cluster are $X^{(1)}, X^{(3)}, X^{(7)}$

(b) for $X^{(1)} = (2,8)$ $\quad ||X^{(1)} - C_1||^2 = ||(2,8)-(2,8)||^2 = 0$

$||X^{(1)} - C_2||^2 = ||(2,8)-(8,4)||^2 = 52$ $\quad ||X^{(1)} - C_3||^2 = ||(2,8)-(7,3)||^2 = 50$

Repeat the steps for other points

$Q(r,c) = \frac{1}{8}(0+52+50+18+0+6+0+18) = \frac{71}{4}$

(c) 3 more iterations.

3. [10 points] Name 2 deep generation networks. [2 points] Briefly describe the training procedure of a GAN model.(What's the objective function? How to update the parameters in each stage?) [8 points]

Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE)

(1) Initialize the generator and discriminator

(2) Training on generator and discriminator:

① generate fake data samples by passing random noise

② these samples are fed into the discriminator will real data samples

③ calculate the loss

④ update the generator's weights to minimize the loss

(3) Adversarial training: repeat (2) steps iteratively to optimize the generator and discriminator.

(4) Converges when the generator generates indistinguishably fake data and the discriminator can't differentiate between real and fake samples.

The minimax objective function is:

$$\min_{\theta_g} \max_{\theta_d} \left[E_{x \sim p_{data}} \log D_{\theta_d}(x) + E_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$