ML HW6

B06303126 Lo Yun Chien

January 2020

Question 1

Correct answer: (b)
By Lecture 212 page 15, we have

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)} w_{jk}^{(l+1)} (tanh'(s_j^l))$$

Putting l = 2, we have 6 neurons, and k = 1 for there's 1 neuron in layer 3. Thus it counts for 6 * 1 operations. And for l = 1, we have 5 neurons, and k = 6 for there are 5 neurons in layer 2. Thus it counts for 5 * 6 operations.

Sum them up. The answer is 36

Question 2

Correct answer: (d)
First we starts from 1 hidden layer, and then it has 50 neurons. While creating the second layer, consider MR and MC

$$MR = 49 + 3$$
$$MC = 20 + 3$$

Which means, while we transfer 1 neuron to the second layer, we gain 49 + 3 connections and loss 20 + 3 connections. Since $MR > MC$, it's beneficial to do it. Until two layers each are 33, 17 neurons(MR = MC). Now we consider the third layer, transferring 1 neuron from first layer have $MR = 17 + 3$, $MC = 20 + 17$, transferring 1 neuron from second layer to third layer have $MR = 16 + 3$, $MC = 33 + 3$, both not worth it. Since the whole question is convex, then we sure the problem has reached its maximum, which is 1219

Question 3

Correct answer: (d)

$$\frac{\partial err}{\partial s_i^{(L)}} = -\sum_{k=1}^{K} v_k \frac{\frac{\partial q_k}{\partial s_i^{(L)}}}{q_k}$$

For $k \neq i$

$$\frac{\partial q_k}{\partial s_i^{(L)}} = \frac{-e^{s_k}}{(\sum_j e^{s_j})^2} e^{s_i} = -q_k q_i$$

For $k = i$

$$\frac{\partial q_k}{\partial s_i^{(L)}} = \frac{e^{s_i}}{\sum_j e^{s_j}} - (\frac{e^{s_i}}{\sum_j e^{s_j}})^2 = q_i - q_i^2$$

Summarized we get

$$\frac{\partial err}{\partial s_i^{(L)}} = -\sum_{k \neq i} v_k(-q_i) - v_i(1 - q_i) = q_i \sum_k v_k - v_i = q_i - v_i$$

Question 4

Correct answer: (a)
In the first run, $x_i^{(l)}$ are 0 for all i and l, thus $s_i^{(l)}$ are all 0 too. Then $\delta_1^{(2)} = -2$, $\delta_i^{(1)} = 0$ for all i, since it's multiplied the weight. When updating w, only the bias term in the second layer is updated, others since $x_i = 0$ then stay 0. The same induction works for the second and the third update. Therefore $w_{01}^{(1)}$ remain 0.

Question 5

Correct answer: (e)
Since V is a all 2 constant matrix, the hidden layer's value is 2. Running a regression per movie,

$$2 * w_m = r_{nm}$$

the best estimator is the average of r_{nm} among n. Since it has multiplied 2, then weight is half of the average rating.

Question 6

Correct answer: (b)
Derive the partial derivative,

$$\frac{\partial err}{\partial a_m} = -2(r_{nm} - w_m^T v_n - a_m - b_n)$$

Where $\eta = \frac{1}{2}$, and turn it to the negative direction.

$$a_m \longleftarrow a_m + \eta(r_{nm} - w_m^T v_n - a_m - b_n)$$

Thus the correct answer is (b)

Question 7

Correct answer: (d)
We calculate

$$E[y_n \neq G(x_n)] = E[y_n \neq sign(g_1(x_n) + g_2(x_n) + g_3(x_n))] = 0.2$$

It imply when G makes mistakes, at least two of g_i need make mistakes. Due to this probability framework, We define the below algorithm to check $E_{out}(g_i)$ is valid.

For $\max(E_{out}(g_i))$¿0.2 , if second $\max(E_{out}(g_i))$ is bigger than 0.2, then check whether the third one is less than 0.2, if it pass all the trial, it's valid.

if second $\max(E_{out}(g_i))$ is less than 0.2, check whether sum of the second one and the third one is bigger than 0.2, if it pass all the trial, it's valid.

The reason behind is trying the overlapping area is big enough, thus the answer is (d)

Question 8

Correct answer: (c)
The same as previous question, we need at least three g_i make mistakes. Since it's all independent, intersection can be calculated as product, which yield

$$\binom{5}{3}(0.6)^2(0.4)^3 + \binom{5}{4}(0.6)^1(0.4)^4 + (0.4)^5$$

The closest answer is (c)

Question 9

Correct answer: (b)
Calculate

$$(1 - \frac{1}{N})^{0.5N} = ((1 - \frac{1}{N})^N)^{0.5} = (\frac{1}{e})^{0.5}$$

The closest answer is (c)

Question 10

Correct answer: (e)
Consider the first component, denote as x_1 and x_1' and assume $x_1 \leq x_1'$. For any cutting point c less than x_1 or bigger than x_1', the decision stump yield the same sign and contribute 1 to ϕ_{ds}. For c between two points, the decision stump contribute -1 to ϕ_{ds}. To sum up, There are $2R-2L-|x_1-x_1'|$ $+1$ $|x_1 - x_1'|$ -1. Thus the first component contribute $2R - 2L - 2|x_1 - x_1'|$. The same induction can be expanded to other components, thus we yield $\phi_{ds}(x, x') = 2d(R - L) - 2\|x - x'\|_1$

Question 11

Correct answer: (a)
In Adaboost, incorrect example multiplies $1 - \epsilon_t$, correct example multiplies ϵ_t for they're sampling possibility. Where $\epsilon_t = 0.05$ and incorrect one is positive example. Thus the answer is $\frac{0.95}{0.05} = 19$

Question 12

Question 13

Correct answer: (c)
While μ_+ or μ_- is tends to 0, the maximum value tends to infinity, therefore, the one who dominant is the one who has the smallest value, therefore is $min(\mu_+, \mu_-)$.