Homework-3 Solutions

Question 1

Part 1

$$e_A = 0.5$$
$$e_B = 0.5$$

Part 2

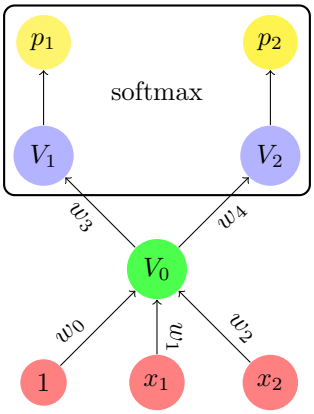
What are the computed errors in each case:

Permutations 1: $e_A = 5/6 \quad e_B = 5/6$
Permutations 2: $e_A = 4/6 \quad e_B = 0.5$
Permutations 3: $e_A = 5/6 \quad e_B = 5/6$

What are the estimates to the error and standard deviation for each algorithm:

Algorithm A: $e = 0.778 \quad \sigma = 0.096$
Algorithm B: $e = 0.722 \quad \sigma = 0.192$

Question 2



The above neural network has two inputs. It computes a selection between the two alternatives A, B in terms of two probability outputs. p_1 is the probability that A occurs, and p_2 is the probability that B occurs. The node V_0 is implemented with ReLU. The nodes V_1, V_2 are linear (ADALINE), and they are not connected to a bias. The probabilities p_1, p_2 are computed from the values of V_1, V_2 using softmax.

A.1: Compute the values of all nodes in forward propagation when the network is given the input $\boxed{x_1 = 2,\ x_2 = 7}$, the current weight values are: $\boxed{w_0 = 0,\ w_1 = 0.2,\ w_2 = 0.1,\ w_3 = 0.1,\ w_4 = 1}$, with $\boxed{\text{the desired selection being } \mathbf{A}}$. Use training rate $\boxed{\epsilon = 0.1}$. Your answer should be explicit numeric values for each node.

Answer

$$V_0 = w_0 + 2w_1 + 7w_2 = 0.4 + 0.7 = 1.1$$
$$V_1 = w_3 V_0 = 0.11$$
$$V_2 = w_4 V_0 = 1.1$$
$$p_1 = e^{V_1}/(e^{V_1} + e^{V_2}) = 0.27 \qquad (Z = 4.12)$$
$$p_2 = e^{V_2}/(e^{V_1} + e^{V_2}) = 0.73$$

Question 3

The Adam method uses a recursive method for computing running averages:

$$\overline{X}_0 = 0, \quad \overline{X_t} = \beta \overline{X}_{t-1} + (1-\beta) X_t, \quad \hat{X}_t = \frac{\overline{X}_t}{1-\beta^t}$$

1. Show that if $\beta = 0$ then $\hat{X}_t = X_t$ for all t.

Answer: From the recursive formula it follows that $\overline{X}_t = X_t$. Therefore $\hat{X}_t = \frac{X_t}{1-\beta^t} = X_t$.

2. Show that if $\beta \to 1$ then $\hat{X}_t \to \frac{1}{t}\sum_{i=1}^{t} X_i$ for all t.

Hint: Use the explicit formula for $\overline{X_t}$, form the limit, and solve it using, for example, L'Hopital's rule.

$$\text{The explicit formula:} \quad \overline{X}_t = (1-\beta)\sum_{i=1}^{t} \beta^{t-i} X_i$$

We need to evaluate:

$$\text{limit} = \lim_{\beta \to 1} \frac{(1-\beta)\sum_{i=1}^{t} \beta^{t-i} X_i}{1-\beta^t}$$

According to L'Hopital's rule we can take the derivatives of numerator and denominator. This gives:

$$\text{limit} = \lim_{\beta \to 1} \frac{-\sum_{i=1}^{t} \beta^{t-i} X_i + (1-\beta)\sum_{i=1}^{t} (t-i)\beta^{t-i-1} X_i}{-t\beta^{t-1}} = \frac{-\sum_{i=1}^{t} X_i + 0}{-t} = \frac{\sum_{i=1}^{t} X_i}{t}$$

Question 4

In the table below cases 3,4 are distributions, and cases 1, 2 can be converted into distributions.

case	A	B	C	D
1	1	-2	3	-4
2	1	2	-3	0
3	1	0	0	0
4	1/4	1/4	1/4	1/4

Converting 1 into a probability distribution using softmax:

$$V = \{1, -2, 3, -4\}$$
$$q = \{2.71828, 0.135335, 20.0855, 0.0183156\}$$
$$Z = 22.9575$$
$$p = \{0.118405, 0.00589504, 0.874902, 0.000797807\}$$

Converting 2 into a probability distribution using softmax:

$$V = \{1, 2, -3, 0\}$$
$$q = \{2.71828, 7.38906, 0.0497871, 1\}$$
$$Z = 11.1571$$
$$p = \{0.243636, 0.662272, 0.00446236, 0.0896288\}$$

1. Use cross entropy to determine which distribution among 1,2,3 is most similar to 4. **Show your computations.**

case	A	B	C	D	cross entropy of p_4 with candidate:
1	0.118405	0.00589504	0.874902	0.000797807	5.24224
2	0.243636	0.662272	0.00446236	0.0896288	3.47989
3	1	0	0	0	infinity
4	1/4	1/4	1/4	1/4	2

Answer: 1 / $\boxed{2}$ / 3

2. Use cross entropy to determine which distribution among 1,2,4 is most similar to 3. **Show your computations.**

case	A	B	C	D	cross entropy of p_3 with candidate:
1	0.118405	0.00589504	0.874902	0.000797807	3.0782
2	0.243636	0.662272	0.00446236	0.0896288	2.0372
3	1	0	0	0	0
4	1/4	1/4	1/4	1/4	2

Answer: 1 / 2 / $\boxed{4}$

Question 5

In this question, if you need to compute logarithms use natural basis logarithm (ln).

Consider a deep neural net applied to decide between the following three categories: A, B, C.

1.

What is the one-hot encoding of the category A?

Answer:

The one-hot encoding of the category A is: (1,0,0)

2.

Assuming that the true category is A, what is the cross entropy loss of the following estimate, given as the vector $z_1 = (1.0, 2.0, 3.0)$. You must show your computations.

Answer:

The cross entropy loss is

$$\ln(1/q_1) = \ln(\frac{e^1 + e^2 + e^3}{e^1}) \approx \ln(\frac{30.192}{2.718}) \approx \ln(11.10) \approx 2.4$$

3.

As in Part 2 assume that the true category is A. Consider an algorithm that updates the prediction by adding the value of ϵ to to each prediction value. (ϵ can also be negative.) The new prediction vector is $z_2 = (1 + \epsilon, 2 + \epsilon, 3 + \epsilon)$. Compute a value of ϵ that gives the best prediction according to cross entropy. What is the cross entropy loss for that value of ϵ? You must show your computations.

Answer:

ϵ = result is independent of ϵ

The cross entropy loss is

$$\ln(1/q_1) = \ln(\frac{e^{1+\epsilon} + e^{2+\epsilon} + e^{3+\epsilon}}{e^{1+\epsilon}}) = \ln(\frac{e^1 + e^2 + e^3}{e^1})$$

Answer should be the same as in Part 2.

4.

As in Part 2 assume that the true category is A. Consider an algorithm that updates the prediction by multiplying each value by ϵ. (ϵ can also be negative.) The new prediction vector is $z_3 = (1 \cdot \epsilon, 2 \cdot \epsilon, 3 \cdot \epsilon)$. Compute a value of ϵ that gives the best prediction according to cross entropy. What is the cross entropy loss for that value of ϵ? You must show your computations.

Answer:

$\epsilon = -\infty$

The cross entropy loss is

$$\ln(1/q_1) = \ln(\frac{e^{1\cdot\epsilon} + e^{2\cdot\epsilon} + e^{3\cdot\epsilon}}{e^{1\cdot\epsilon}}) = \ln(\frac{e^{\epsilon} + (e^{\epsilon})^2 + (e^{\epsilon})^3}{e^{\epsilon}}) = \ln(1 + e^{\epsilon} + (e^{\epsilon})^2) \xrightarrow[\epsilon \to -\infty]{} \ln(1) = 0$$