CS 188 Summer 2020 Introduction to Artificial Intelligence

Final Review

NN and Logistic Regression Solutions

Q1. Optimization

We would like to classify some data. We have N samples, where each sample consists of a feature vector $\mathbf{x} = \{x_1, \cdots, x_k\}$ and a label $y = \{0, 1\}$.

We introduce a new type of classifier called logistic regression, which produces predictions as follows:

$$P(Y = 1|X) = h(\mathbf{x}) = s\left(\sum_i w_i x_i\right) = \frac{1}{1 + \exp(-(\sum_i w_i x_i))}$$

$$s(\gamma) = \frac{1}{1 + \exp(-\gamma)}$$

where $s(\gamma)$ is the logistic function, $\exp x = e^x$, and $\mathbf{w} = \{w_1, \cdots, w_k\}$ are the learned weights.

Let's find the weights w_j for logistic regression using stochastic gradient descent. We would like to minimize the following loss function for each sample:

$$L = -[y \ln h(\mathbf{x}) + (1 - y) \ln(1 - h(\mathbf{x}))]$$

(a) Find dL/dw_i. Hint: $s'(\gamma) = s(\gamma)(1 - s(\gamma))$.

Use chain rule:

$$\frac{dL}{dw_i} = -\left[\frac{y}{h(\mathbf{x})} s'(\sum_i w_i x_i) x_i - \frac{1 - y}{1 - h(\mathbf{x})} s'(\sum_i w_i x_i) x_i\right]$$

Use hint:

$$\frac{dL}{dw_i} = -\left[\frac{y}{h(\mathbf{x})} h(\mathbf{x})(1 - h(\mathbf{x})) x_i - \frac{1 - y}{1 - h(\mathbf{x})} h(\mathbf{x})(1 - h(\mathbf{x})) x_i\right]$$

Simplify:

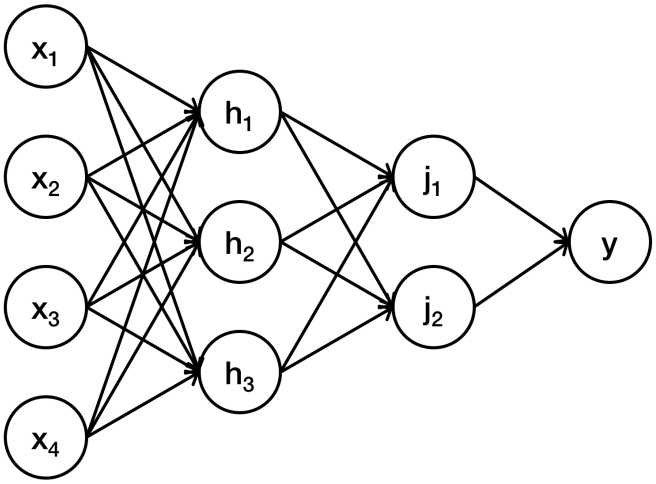
$$\begin{aligned}\frac{dL}{dw_i} &= -\left[y(1 - h(\mathbf{x})) x_i - (1 - y) h(\mathbf{x}) x_i\right] \\ &= -x_i[y - y h(\mathbf{x}) - h(\mathbf{x}) + y h(\mathbf{x})] \\ &= -x_i(y - h(\mathbf{x}))\end{aligned}$$

(b) Write the stochastic gradient descent update for w_i. Our step size is η.

$$w_i \leftarrow w_i + \eta x_i (y - h(\mathbf{x}))$$

Q2. Neural Network Data Sufficiency

The next few problems use the below neural network as a reference. Neurons h_{1-3} and j_{1-2} all use ReLU activation functions. Neuron y uses the identity activation function: $f(x) = x$. In the questions below, let $w_{a,b}$ denote the weight that connects neurons a and b. Also, let o_a denote the value that neuron a outputs to its next layer.



Given this network, in the following few problems, you have to decide whether the data given are sufficient for answering the question.

(a) Given the above neural network, what is the value of o_y?

Data item 1: the values of all weights in the network and the values o_{h_1}, o_{h_2}, o_{h_3}
Data item 2: the values of all weights in the network and the values o_{j_1}, o_{j_2}

- ○ Data item (1) alone is sufficient, but data item (2) alone is not sufficient to answer the question.
- ○ Data item (2) alone is sufficient, but data item (1) alone is not sufficient to answer the question.
- ○ Both statements taken together are sufficient, but neither data item alone is sufficient.
- ● Each data item alone is sufficient to answer the question.
- ○ Statements (1) and (2) together are not sufficient, and additional data is needed to answer the question.

(b) Given the above neural network, what is the value of o_{h_1}?

Data item 1: the neuron input values, i.e., o_{x_1} through o_{x_4}
Data item 2: the values o_{j_1}, o_{j_2}

- ○ Data item (1) alone is sufficient, but data item (2) alone is not sufficient to answer the question.
- ○ Data item (2) alone is sufficient, but data item (1) alone is not sufficient to answer the question.
- ○ Both statements taken together are sufficient, but neither data item alone is sufficient.
- ○ Each data item alone is sufficient to answer the question.
- ● Statements (1) and (2) together are not sufficient, and additional data is needed to answer the question.

(c) Given the above neural network, what is the value of o_{j_1}?

Data item 1: the values of all weights connecting neurons h_1, h_2, h_3 to j_1, j_2
Data item 2: the values o_{h_1}, o_{h_2}, o_{h_3}

- ○ Data item (1) alone is sufficient, but data item (2) alone is not sufficient to answer the question.
- ○ Data item (2) alone is sufficient, but data item (1) alone is not sufficient to answer the question.
- ● Both statements taken together are sufficient, but neither data item alone is sufficient.
- ○ Each data item alone is sufficient to answer the question.
- ○ Statements (1) and (2) together are not sufficient, and additional data is needed to answer the question.

(d) Given the above neural network, what is the value of $\partial o_y / \partial w_{j_2,y}$?

Data item 1: the value of o_{j_2}
Data item 2: all weights in the network and the neuron input values, i.e., o_{x_1} through o_{x_4}

- ○ Data item (1) alone is sufficient, but data item (2) alone is not sufficient to answer the question.
- ○ Data item (2) alone is sufficient, but data item (1) alone is not sufficient to answer the question.
- ○ Both statements taken together are sufficient, but neither data item alone is sufficient.
- ● Each data item alone is sufficient to answer the question.
- ○ Statements (1) and (2) together are not sufficient, and additional data is needed to answer the question.

(e) Given the above neural network, what is the value of $\partial o_y / \partial w_{h_2,j_2}$?

Data item 1: the value of $w_{j_2,y}$
Data item 2: the value of $\partial o_{j_2} / \partial w_{h_2,j_2}$

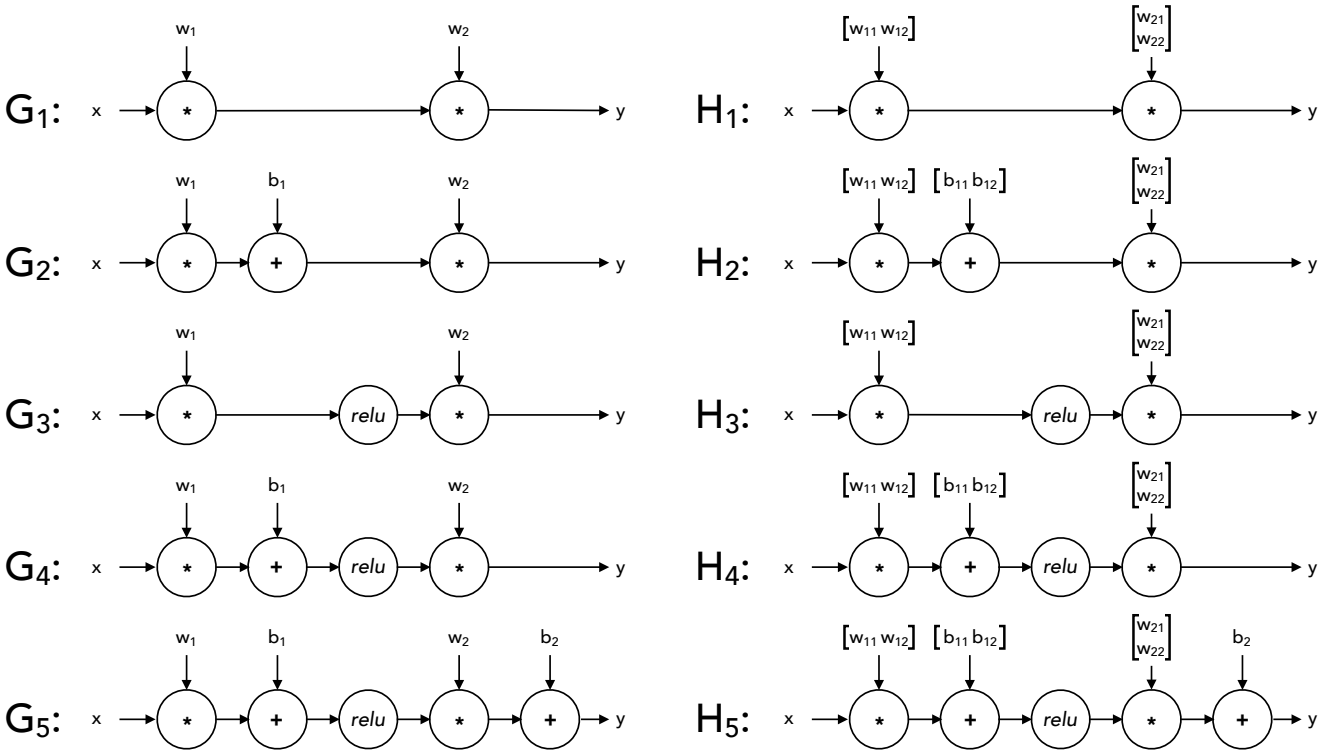
- ○ Data item (1) alone is sufficient, but data item (2) alone is not sufficient to answer the question.
- ○ Data item (2) alone is sufficient, but data item (1) alone is not sufficient to answer the question.
- ● Both statements taken together are sufficient, but neither data item alone is sufficient.
- ○ Each data item alone is sufficient to answer the question.
- ○ Statements (1) and (2) together are not sufficient, and additional data is needed to answer the question.

(f) Given the above neural network, what is the value of $\partial o_y / \partial w_{x_1,h_3}$?

Data item 1: the value of all weights in the network and the neuron input values, i.e., o_{x_1} through o_{x_4}
Data item 2: the value of w_{x_1,h_3}

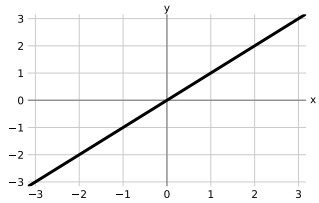
- ● Data item (1) alone is sufficient, but data item (2) alone is not sufficient to answer the question.
- ○ Data item (2) alone is sufficient, but data item (1) alone is not sufficient to answer the question.
- ○ Both statements taken together are sufficient, but neither data item alone is sufficient.
- ○ Each data item alone is sufficient to answer the question.
- ○ Statements (1) and (2) together are not sufficient, and additional data is needed to answer the question.

Q3. Neural Networks: Representation

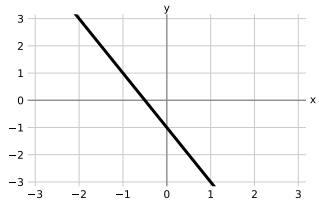


For each of the piecewise-linear functions below, mark all networks from the list above that can represent the function **exactly** on the range $x \in (-\infty, \infty)$. In the networks above, *relu* denotes the element-wise ReLU nonlinearity: $relu(z) = max(0, z)$. The networks G_i use 1-dimensional layers, while the networks H_i have some 2-dimensional intermediate layers.

(a)



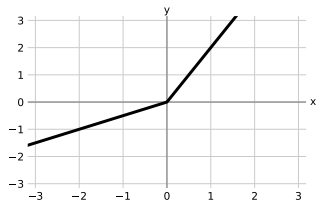
☐ G_1 ☐ H_1
☐ G_2 ☐ H_2
☐ G_3 ☐ H_3
☐ G_4 ☐ H_4
☐ G_5 ☐ H_5
○ None of the above



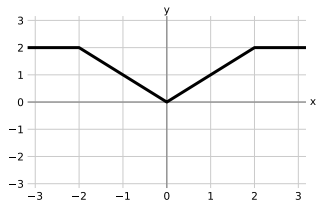
☐ G_1 ☐ H_1
☐ G_2 ☐ H_2
☐ G_3 ☐ H_3
☐ G_4 ☐ H_4
☐ G_5 ☐ H_5
○ None of the above

The networks G_3, G_4, G_5 include a ReLU nonlinearity on a scalar quantity, so it is impossible for their output to represent a non-horizontal straight line. On the other hand, H_3, H_4, H_5 have a 2-dimensional hidden layer, which allows two ReLU elements facing in opposite directions to be added together to form a straight line. The second subpart requires a bias term because the line does not pass through the origin.

(b)



☐ G_1 ☐ H_1
☐ G_2 ☐ H_2
☐ G_3 ☐ H_3
☐ G_4 ☐ H_4
☐ G_5 ☐ H_5
○ None of the above



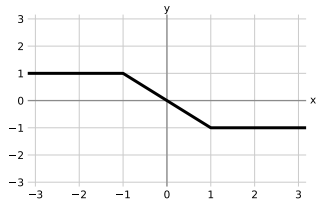
☐ G_1 ☐ H_1
☐ G_2 ☐ H_2
☐ G_3 ☐ H_3
☐ G_4 ☐ H_4
☐ G_5 ☐ H_5
● None of the above

These functions include multiple non-horizontal linear regions, so they cannot be represented by any of the networks G_i which apply ReLU no more than once to a scalar quantity.

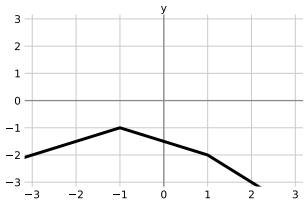
The first subpart can be represented by any of the networks with 2-dimensional ReLU nodes. The point of nonlinearity occurs at the origin, so nonzero bias terms are not required.

The second subpart has 3 points where the slope changes, but the networks H_i only have a single 2-dimensional ReLU node. Each application of ReLU to one element can only introduce a change of slope for a single value of x.

(c)



☐ G_1 ☐ H_1
☐ G_2 ☐ H_2
☐ G_3 ☐ H_3
☐ G_4 ☐ H_4
☐ G_5 ■ H_5
○ None of the above



☐ G_1 ☐ H_1
☐ G_2 ☐ H_2
☐ G_3 ☐ H_3
☐ G_4 ■ H_4
☐ G_5 ■ H_5
○ None of the above

Both functions have two points where the slope changes, so none of the networks G_i; H_1, H_2 can represent them.

An output bias term is required for the first subpart because one of the flat regions must be generated by the flat part of a ReLU function, but neither one of them is at $y = 0$.

The second subpart doesn't require a bias term at the output: it can be represented as $-relu(\frac{-x+1}{2}) - relu(x + 1)$. Note how if the segment at $x > 2$ were to be extended to cross the x axis, it would cross exactly at $x = -1$, the location of the other slope change. A similar statement is true for the segment at $x < -1$.