107061112 王昊文 2021/11/07

COM526000 Deep Learning Assignment 1

1. Show the following properties of the sigmoid and tanh activation functions (denoted by $\Phi(\cdot)$ in each case):

   (a) Sigmoid activation: $\Phi(-\mathrm{v}) = 1 - \Phi(\mathrm{v})$

   (b) Tanh activation: $\Phi(-\mathrm{v}) = -\Phi(\mathrm{v})$

   (c) Hard tanh activation: $\Phi(-\mathrm{v}) = -\Phi(\mathrm{v})$

Solution.

(a) Sigmoid activation is defined as:

$$\Phi(v) = \frac{1}{1+e^{-v}}$$

So,

$$\Phi(-v) = \frac{1}{1+e^{v}} = \frac{e^{-v}}{e^{-v}+1}$$

Multiplying the numerator and denominator by e^{-v}, we get:

$$\frac{1}{1+e^{v}} = \frac{e^{-v}}{e^{-v}+1}$$

We know that:

$$\frac{e^{-v}}{e^{-v}+1} = \frac{1+e^{-v}}{1+e^{-v}} - \frac{1}{1+e^{-v}}$$

that is,

$$\Phi(-v) = 1 - \Phi(v)$$

(b) tanh activation is defined as:

$$\Phi(v) = \frac{e^{v}-e^{-v}}{e^{v}+e^{-v}}$$

So,

$$\Phi(-v) = \frac{e^{-v}-e^{v}}{e^{-v}+e^{v}}$$

Multiplying the numerator and denominator of $\Phi(-v)$ by e^{2v}

$$\Phi(-v) = \frac{e^{-v} - e^{v}}{e^{-v} + e^{v}} = \frac{e^{v} - e^{3v}}{e^{v} + e^{3v}}$$

Also multiplying the numerator and denominator of $\Phi(v)$ by e^{2v}

$$\Phi(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} = \frac{e^{3v} - e^{v}}{e^{3v} + e^{v}}$$

We know that:

$$\frac{e^{v} - e^{3v}}{e^{v} + e^{3v}} = -(\frac{e^{3v} - e^{v}}{e^{3v} + e^{v}})$$

that is,

$$\Phi(-v) = -\Phi(v)$$

(c) Hard **tanh** activation is defined as:

$$\Phi(v) = \mathsf{max}\left\{\mathsf{min}\left[v, 1\right], -1\right\}$$

That is:

$$\Phi(v) = \begin{cases} -1 & \text{for } v < -1 \\ \text{v} & \text{for } -1 \leq v \leq 1 \\ 1 & \text{for } v > 1 \end{cases}$$

So

$$\Phi(-v) = \begin{cases} -1 & \text{for } v > 1 \\ \text{-v} & \text{for } -1 \leq v \leq 1 \\ 1 & \text{for } v < -1 \end{cases}$$

Also,

$$-\Phi(v) = \begin{cases} 1 & \text{for } v < -1 \\ \text{-v} & \text{for } -1 \leq v \leq 1 \\ -1 & \text{for } v > 1 \end{cases}$$

That is:

$$\Phi(-v) = -\Phi(v)$$

2. Consider a network with two inputs x_1 and x_2. It has two hidden layers, each of which contain two units. Assume that the weights in each layer are set so that top unit in

each layer applies sigmoid activation to the sum of its inputs and the bottom unit in each layer applies tanh activation to the sum of its inputs. Finally, the single output node applies ReLU activation to the sum of its two inputs. Write the output of this neural network *in closed form* as a function of x_1 and x_2. This exercise should give you an idea of the complexity of functions computed by neural networks.

Solution. The closed form should be as following:

$$o = \max\{0, \sigma(\sigma(x_1 + x_2) + \tanh(x_1 + x_2)) + \tanh(\sigma(x_1 + x_2) + \tanh(x_1 + x_2))\}$$

Where o denotes the output, σ denotes the sigmoid activation.

3. Consider the following loss function for training pair $(\overline{X}, y)$:

$$L = \max(0,\ a - y(\overline{W} \cdot \overline{X}))$$

The test instances are predicted as $\hat{y} = \text{sign}(\overline{W} \cdot \overline{X})$. A value of $a = 0$ corresponds to the perceptron criterion and a value of $a = 1$ corresponds to the SVM. Show that any value of $a > 0$ leads to the SVM with an unchanged optimal solution when no regularization is used. What happens when regularization is used?

Solution. Assuming the optimal weight matrix $\overline{W}_o$, at $a = 1$ that acheives the minimum loss is L_o, that is:

$$L_o = \max(0,\ 1 - y(\overline{W}_o \cdot \overline{X}))$$

Note, optimizing $L = \max(0,\ 1 - y(\overline{W} \cdot \overline{X}))$ and $L' = a \cdot \max(0,\ 1 - y(\overline{W} \cdot \overline{X}))$ is the same thing when doing gradient decent w.r.t weights, the formula for weight update is the same. We should end up in the same minimum points.
For any a, we will result in the a corresponding solution $a\overline{W}_o$, that satisfies the following:

$$\hat{y} = \text{sign}(a\overline{W}_o \cdot \overline{X})$$

If we plug in the loss function, for $a > 0$:

$$L = \max(0,\ a - y(a \cdot \overline{W}_o \cdot \overline{X})) = a \cdot \max(0,\ 1 - y(\overline{W}_o \cdot \overline{X}))$$

This suggests that we are optimizing the same loss function as we use $a = 1$ and using the solution $\overline{W_o}$. Hence, $\overline{W_o}$ does not change for any a.
Now, if we are using regularization, then the loss function should look something like:

$$L = \max(0,\ 1 - y(\overline{W}_o \cdot \overline{X})) + \lambda||W_o||^n$$

which is also identical to optimizing:

$$a\max(0,\ 1 - y(\overline{W}_o \cdot \overline{X})) + a\lambda||W_o||^n = \max(0, a - y(a\overline{W}_0 \cdot \overline{X})) + a\lambda||W_o||^n$$

However, we need to plug in aW_o for the regularization term. This suggests that if we are using regularization, the new regularization parameter λ':

$$a\lambda' = a^{n+1}\lambda$$

That is

$$\lambda' = \frac{\lambda}{a^n}$$

So depending on the regularization we are using, the regularization parameter must be scaled down.

4. Based on exercise 3, formulate a generalized objective for the Weston-Watkins SVM.

Solution. For the ith training instance, the generalized objective function is defined as:

$$J_i = \sum_{r:r\neq c(i)} \max(\overline{W_r} \cdot \overline{X_i} - \overline{W}_{c(i)} \cdot \overline{X}_i + a,\ 0)\ , \forall a > 0$$

Where the ith training instance is denoted as $(\overline{X_i}, c(i))$, $\overline{X_i}$ contains the d-dimensional feature vaiables, and $c(i)$ contains the class index drawn from {1, ... k}.

5. Consider a two-input neuron that multiplies its two inputs x_1 and x_2 to obtain the output o. Let L be the loss function that is computed at o. Suppose that you know that $\frac{\partial L}{\partial o} = 5$, $x_1 = 2$, $x_2 = 3$. Compute the values of $\frac{\partial L}{\partial x_1}$ and $\frac{\partial L}{\partial x_2}$.

Solution. First, we have that

$$x_1 \times x_2 = o$$

So, we have

$$\frac{\partial o}{\partial x_1} = x_2 \quad \frac{\partial o}{\partial x_2} = x_1$$

By chain rule,

$$\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial o} \times \frac{\partial o}{\partial x_1} \tag{1}$$

$$\frac{\partial L}{\partial x_2} = \frac{\partial L}{\partial o} \times \frac{\partial o}{\partial x_2} \tag{2}$$

Plug in values for (1) and (2), we get

$$\frac{\partial L}{\partial x_1} = 5 \times 3 = 15$$

and

$$\frac{\partial L}{\partial x_2} = 5 \times 2 = 10$$

6. Show that if the dot product of a d-dimensional vector $\overline{v}$ with d linearly independent vectors is 0, then $\overline{v}$ must be the zero vector.

Solution. First, let us define the d linearly independent vectors as $x_1, x_2, \ldots x_d$. Since they are linearly independent, these d vectors form a basis in a d-dimensional space. Hence, the d-dimensional vector $\overline{v}$ can be written as a linear combination of $x_1, x_2, \ldots x_d$. That is:

$$\overline{v} = \sum_{i=1}^{d} \alpha_i \overline{x_i}$$

If we calculate $\|\overline{v}\|^2$:

$$\|\overline{v}\|^2 = \sum_{i=1}^{d} \alpha_i (\overline{v} \cdot \overline{x_i})$$

and that the dot product of $\overline{v}$ with $x_1, x_2, \ldots x_d$ is 0, that is:

$$v \cdot x_i = 0, \ \forall\ 0 < i \leq d, \ i \in \mathbb{Z}$$

So,

$$\|\overline{v}\|^2 = \sum_{i=1}^{d} \alpha_i (\overline{v} \cdot \overline{x_i}) = \sum_{i=1}^{d} \alpha_i (0) = 0$$

That is:

$$\|\overline{v}\| = 0$$

Thus, $\overline{v}$ must be a zero vector.

7. Consider two neural networks used for regression modeling with identical structure of an input layer and 10 hidden layers containing 100 units each. In both cases, the output node is a single unit with linear activation. The only difference is that one of them uses linear activations in the hidden layers and the other uses sigmoid activations. Which model will have higher variance in prediction?

Solution. The neural network with linear activations will result in a linear model, which can be expressed as a closed-form linear equation. Linear models are one of the simplest models, hence it will have higher bias and lower vairiance.
On the other hand, the neural network with sigmoid activaions adds non-linearity to the model, resulting in a more complex model. Hence it will result in a higher variance in predection.
To sum up, the neural network with sigmoid activaions will have higher variance in predection.

8. Consider a network with a single input layer, two hidden layers, and a single output predicting a binary label. All hidden layers use the sigmoid activation function and no regularization is used. The input layer contains d units, and each hidden layer contains p units. Suppose that you add an additional hidden layer between the two current hidden layers, and this additional hidden layer contains q linear units.

   (a) Even though the number of parameters have increased by adding the hidden layer, discuss why the capacity of this model will decrease when $q < p$.

   (b) Does the capacity of the model increase when $q > p$?

Solution.

   (a) Since $q < p$, this makes the q dimensional layer a *low-rank approximation* of the p dimensional information. The This technique is widely used in autoencoders, to extract information to a lower dimension.

   (b) If $q > p$, the capacity will not increase, since there is a maximum informaion

you can extract from linear layers, the additional linear layer can be collapsed with the existing weight matrix without changing the prediction.