CLASS Exercises: WORD CLASSIFICATION

Exercise 1

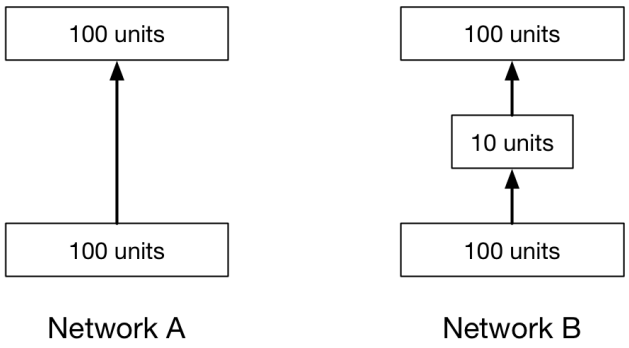
We have a multilayer perceptron for classification with a single hidden layer with the hard threshold activation function. The output layer uses the softmax activation function with cross-entropy loss. What will happen when training with gradient descent?

SOLUTION:

For the hard threshold activation function, the derivative is zero almost everywhere, so the gradient for the first-layer weights will be zero, and the weights will never be updated.

Exercise 2

We have two alternative networks with multilayer perceptrons. Consider that all of the layers use linear activation functions.



2a. Mention and argue one advantage of Network A over Network B.
2b. Mention and argue one advantage of Network B over Network A.

SOLUTION

2a.
*A is more expressive than B.
*A has fewer units or is easier to implement.

2b
*B has fewer connections, so it's less prone to overfitting
*B has fewer connections, so backprop requires fewer operations
*B has a bottleneck layer, so the network is forced to learn a compact representation (like an autoencoder)

Exercise 3

Let's imagine that you are training a neural network for a classification task. In this process, you get a much lower training loss than the validation loss. Mention what problem are you facing and how can you solve it.

SOLUTION

This is overfitting. You can use a network with fewer layers, Increase L2 regularization weight

Exercise 4

The task to perform is to classify movie review text as either positive or negative sentiment, and either action, comedy or romance movie genre. To perform these two related classification tasks, we use a neural network that shares the first layer, but branches into two separate layers to compute the two classifications. The loss is a weighted sum of the two cross-entropy losses.

4a-Given the previous description, rewrite the equations by completing the information marked with ? on the following equations:

$$h = RELU(W_0X + b_0), h \in \mathbb{R}^{10}, W_0 \in \mathbb{R}^{10x?},$$
$$\hat{y}_1 = softmax(W_1? + b_1), \hat{y}_1 \in \mathbb{R}^{?}, W_1 \in \mathbb{R}^{?x10}$$
$$\hat{y}_2 = softmax(W_2? + b_2), \hat{y}_2 \in \mathbb{R}^{?}, W_1 \in \mathbb{R}^{?x10}$$
$$J = \alpha CE(y_1, ?) + \beta(y_2, ?)$$

4b-When training the model, we see that the model is underfitting? What does it mean? Provide solutions for this.

SOLUTION

4a

$$h = RELU(W_0X + b_0), h \in \mathbb{R}^{10}, W_0 \in \mathbb{R}^{10x10},$$
$$\hat{y}_1 = softmax(W_1h + b_1), \hat{y}_1 \in \mathbb{R}^{2}, W_1 \in \mathbb{R}^{2x10}$$
$$\hat{y}_2 = softmax(W_2h + b_2), \hat{y}_2 \in \mathbb{R}^{3}, W_1 \in \mathbb{R}^{3x10}$$
$$J = \alpha CE(y_1, \hat{y}_1) + \beta(y_2, \hat{y}_2)$$

4b

The model is too simple (just two layers, hidden layers' dimension is only 10, input feature dimension is only 10).
Anything increasing the complexity of the model would be accepted, including:
_ Increasing dimensions of the hidden layer
_Adding more layers
_ Splitting the model into two (with more overall parameters)