Midterm Review

CS 4277: Deep Learning

1. According to Tom Mitchell, machine learning is the study of algorithms that

   - improve their performance `P`
   - at some task `T`
   - with experience `E`.

   A well-defined learning task is given by `<P, T, E>`.

   Formulate the following problems according to Tom Mitchell's machine learning problem specification (see Machine Learning Slides) and the specification our textbook. For each of the following problems specify:

   - The task `T`,
   - The performance measure `P`,
   - The experience `E`,
   - The target function $f : \mathcal{X} \to \mathcal{Y}$, that is,
     - the input space $\mathcal{X}$, and
     - the output space $\mathcal{Y}$.

   Remember that a function maps a domain to a co-domain, and these domains are sets.

2. Medical diagnosis: A patient walks in with a medical history and some symptoms, and you want to identify the problem.

3. For a single neuron with a two-dimensional input $\boldsymbol{x}$ and single scalar output y, formulate the preactivation value using a linear algebra operation. Don't forget to account for the bias.

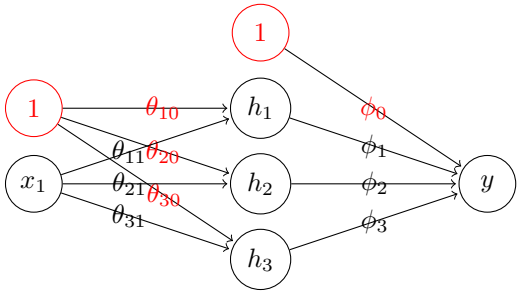
4. Write a mathematical definition of the ReLU activation function.

5. The squared error loss function for a regression model with a single scalar input and single scalar output is:

$$L(\boldsymbol{\phi}) = \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2 \tag{1}$$

   Write the squared error loss function for a regression model with a two-dimensional vector input, $\boldsymbol{x}$, and a single scalar output y.

6. Given the following shallow neural network:



Write a single equation for the network $y = f(x, \vec{\phi}$ with parameters $\boldsymbol{\phi} = \{\boldsymbol{\theta}, \boldsymbol{\phi}\}$.

7. State the universal approximation theorem.

8. A shallow network with $D > 2$ hidden units can create up to how many linear regions?

9. A deep network with K layers of $D > 2$ hidden units can create up to how many linear regions?

10. Explain the *depth efficiency* of deep networks vs shallow networks.

11. Explain the i.i.d. assumption.

12. What is a loss function?

13. How do machine learning algorithms use loss functions?

14. Write a concise general mathematical formulation of the goal of a machine learning algorithm that relates the model's parameters to the loss function.

15. Write the general 2-step gradient descent algorithm.

16. What are local minima? What is the global minimum?

17. What if the learning rate is set too high?

18. Is full-batch gradient descent guaranteed to find the global minimum? Why or why not?

19. Describe stochastic gradient descent and how it improves on full-batch gradient descent.

20. Describe SGD modified with momentum and the purpose of momentum.

21. Describe Adam and its purpose.

22. What are the components of the Hessian matrix?

23. What is useful about the Hessian matrix?

24. What is computed in the forward pass of the backpropagation algorithm?

25. What is computed in the backward pass of the backpropagation algorithm?

26. Describe the efficiency of the backpropagation algorithm with respect to time and memory.

27. How are the weights of a network typically initialized?

28. Describe the vanishing gradients problem and what initialization mistake leads to it.

29. Describe the exploding gradients problem and what initialization mistake leads to it.

30. What initialization method avoids the vanishing gradients and exploding gradients problem?