COMP9444 Neural Networks and Deep Learning

Quiz 1 Answers (Perceptron Learning and Backpropagation)

---

1. What class of functions can be learned by a Perceptron?

   Linearly Separable functions can be learned by a Perceptron.

2. Explain the difference between Perceptron Learning and Backpropagation.

   Perceptron Learning is only used by a Perceptron (one-layer neural network with step activation). Assume the function computed by the Perceptron is $g(w_0 + \Sigma_k\ w_k\ x_k)$ where $g()$ is the Heaviside step funtion, and $\eta$ is the learning rate. If the output is 0 but should have been 1, $\eta$ is added to the bias $w_0$, and $\eta x_k$ is added to each weight $w_k$. If the output is 1 but should have been 0, these values are instead subtracted rather than added.

   Backpropagation is a form of gradient descent, which can be applied to multi-layer neural networks provided the activation function is (mostly) differentiable. The derivative $\partial E/\partial w$ of the cost function $E$ with respect to each weight $w$ is calculated, and $\eta\ \partial E/\partial w$ is subtracted from $w$.

3. When training a Neural Network by Backpropagation, what happens if the Learning Rate is too low? What happens if it is too high?

   If the learning rate is too low, the training will be very slow. If it is too high, the training may become unstable and fail to learn the task successfully.

4. Explain why rescaling of inputs is sometimes necessary for Neural Networks.

   The differential of each weight in the first layer gets multiplied by the value of its corresponding input. Therefore, the network may give undue emphasis to inputs of larger magnitude. Rescaling encourages all inputs to be treated with equal importance.

5. What is the difference between Online Learning, Batch Learning, Mini-Batch Learning and Experience Replay? Which of these methods are referred to as "Stochastic Gradient Descent"?

   For online learning, each training item is presented to the network individually, and the weights are updated using the differentials computed for that item. For batch learning, the differentials for all training items are computed and aggregated, and the weights are updated simultaneously using these aggregated differentials. For mini-batch learning, the differentials for all items in a subset of the training data (called a mini-batch) are computed (perhaps in parallel) and these combined differentials are used to update the weights. For experience replay, items are generated by a separate process (for example, playing a video game) and stored in a database; minibatches are

then selected from that database and used to train the network in parallel. The term "Stochastic Gradient Descent" is sometimes used to refer to any method other than pure batch learning, because the order of the training items, or the choice of mini-batch, effectively adds some random noise to the true gradient.

---