ECS 170: Problem Set 6

March 13, 2016

This assignment will not be graded! We're providing it to practice for the final.

1. What condition must hold on the training data so that the perceptron algorithm can learn a classifier that makes perfect predictions on the training data?

   The data must be linearly separable (i.e. there exists a hyperplane that separates the two classes).

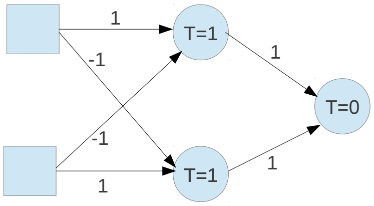
2. Name two desirable properties of the linear unit objective function (shown below). D is the set of training data, t_d is the label of the dth training instance and $o_d = w_0 + w_1 x_1 + \ldots + w_m x_m$.

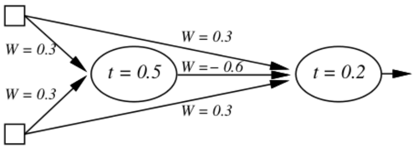
   $$E[w] = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

   - a single local (global) minimum, i.e. function is convex: this is nice since we only need to train our model once.
   - differentiable: desirable so that gradient-based optimization algorithms can be used

3. Construct a network of linear units that is capable of representing the XOR function of two inputs. For this problem, let false be encoded as -1 and true as 1.

   Two possible solutions are shown below, where the left one uses two perceptrons, and the one on the right uses only one perceptron but with slightly more complicated network structure.
   T (or t) in the figures indicate the activation thresholds for the perceptrons.





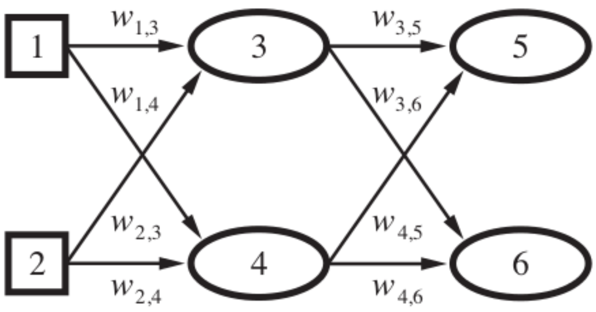


Figure 1

4. Suppose the inputs are given by x_1 and x_2, and the activation functions at each unit is given by the function g. Write out the values o_5 and o_6 at the output nodes (nodes 5 and 6) of figure 1 in terms of the weights $w_{i,j}$ and the inputs x_k. In the above figure, nodes 1 and 2 correspond to x_1 and x_2, nodes 5 and 6 correspond to o_5 and o_6 and nodes 3 and 4 correspond to intermediate values (say x_3 and x_4).

   Recall the activation function is the function used at a unit that takes in the input and outputs the value of this unit. For example, the hard threshold activation $g(\vec{w}, \vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{x} > 0 \\ -1 & \text{otherwiese} \end{cases}$ is used in a perceptron, and the logistic activation function is used in a sigmoid unit.

$$\begin{aligned} x_3 &= g(w_{1,3}x_1 + w_{2,3}x_2) \\ x_4 &= g(w_{1,4}x_1 + w_{2,4}x_2) \\ o_5 &= g(w_{3,5}x_3 + w_{4,5}x_4) \\ o_6 &= g(w_{3,6}x_3 + w_{4,6}x_4) \end{aligned}$$

5. Explain how overfitting can occur for a network of sigmoid units.

   Here are two ways: if the training set is used too many times in the backpropagation algorithm or if the network is too complicated.

6. Why is overfitting a bad in practice?

   It can lead to poor performance on test data.

7. It turns out that larger networks can sometimes perform worse than smaller networks. Give a reason for why this is the case.

   It is easier for larger networks to overfit.

8. When tuning parameters such as the number of iterations to run backpropagation or the number of hidden layers in the network, why is it important to measure performance using cross validation or on a validation set rather than on the test data?

   If you measure performance on the test data then you are implicitly training on your test data, which means your final performance on the test data will not reflect how well the network will perform on future data.