CSE 250B Quiz 8

Tuesday March 4, 2014

Instructions. Do this quiz in partnership with exactly one other student. Write both your names at the top of this page. Discuss the answer to the question with each other, and then write your joint answer below the question. Use the back of the page if necessary. It is fine if you overhear what other students say, because you still need to decide if they are right or wrong. You have ten minutes.

Question. In the approach to semantics presented in class, the meaning vector $x_w \in \mathbb{R}^d$ of node w with child nodes u and v is

$$x_w = h(W[x_u; x_v] + b)$$

where W and b are parameters to be learned, and h is a pointwise function. Assume that x_u and x_v are fixed.

1. Work out $(\partial/\partial W_{ij})x_{wk}$ as explicitly as possible, for every component x_{wk} of x_w and all i and j.

2. Now suppose h is the sigmoid function, with $h(a) = \frac{1}{1+e^{-a}}$ and suppose $x_u = x_v =$ the all ones vector. Suppose $W = [I_d; I_d]$ is two identity matrices horizontally stacked. That is, W is a matrix with $w_{i,i} = w_{i,i+d} = 1$ $\forall\ i = 1, 2, \ldots, d$ and all other entries zero. Assume $b = 0$ and $d = 10$ (possibly reasonable values in practice). What would this derivative be ? (a reasonable approximate answer is fine). What does this say about xs and learning Ws ?

Answer. Let $W_{k\cdot}$ be row number k of the parameter matrix W. The components of the vector x_w are

$$x_{wk} = [h(W[x_u; x_v] + b)]_k = h([W[x_u; x_v] + b]_k) = h(W_{k\cdot}[x_u; x_v] + b_k)$$

using the same notation h for the scalar and pointwise sigmoid functions. Write $a = W_{k\cdot}[x_u; x_v] + b_k$. Then

$$\frac{\partial}{\partial W_{ij}} x_{wk} = h'(a) \frac{\partial a}{\partial W_{ij}}.$$

We cannot simplify $h'(a)$ because h is not specified. For the other factor,

$$\frac{\partial a}{\partial W_{ij}} = \begin{cases} 0 & \text{if } i \neq k \\ x_{uj} & \text{if } i = k \text{ and } j \leq d \\ x_{v(j-d)} & \text{if } i = k \text{ and } j > d. \end{cases}$$

For part 2. Recognize that $\frac{\partial h}{\partial a} = h(a)[1 - h(a)]$.

For any k $a_k = 2$. $h(a_k) = \frac{1}{1+\frac{1}{e^2}} = \frac{e^2}{1+e2}$.

$\frac{\partial x_{wk}}{\partial a} = \left(\frac{e}{1+e^2}\right)^2 \approx \frac{1}{10}$

Here any approximation of $\frac{1}{10}$ is reasonable.

$\frac{\partial x_{wk}}{\partial w_{ij}} = \frac{\partial h}{\partial a} \frac{\partial a}{\partial w_{ij}} =\approx \frac{1}{10}$ or 0 as in part 1.

which is quite small. Thus each x has a small influence on its own $(h(a_k))$ output.

Improvements include - making W small/normalized or xs normalized by $O(\frac{1}{\sqrt{d}})$ factors so that a_k stays small.