CMPS290C Machine Learning

Fall 17 Warmuth

handed out Tu 10-24-17

Homework Assignment 4

due Th 10-31-17 in class

Theoretical homework. All work should be done by yourself.

1. Consider 1-dimensional linear regression.

   First compute the optimum solution w^* for a batch of examples (x_i, y_i), $1 \leq i \leq n$, ie the weight that minimizes the total loss on all examples: $L(w) = \sum_{i=1}^{n}(w\,x_i - y_i)^2$.

   Assume labels are expensive (See Lecture 7). You are given only one of the label y_i. Compute the optimal solution w_i^* based on a single example (x_i, y_i).

   Show that if i is chosen wrt the distribution $\frac{x_i^2}{\sum_j x_j^2}$, then the expected loss of w_i^* on all examples is twice the optimum, ie

   $$\mathbb{E}\left[L(w_i^*)\right] = 2L(w^*),$$

   when all x_i are non-zero.

   Hint: First check the above equation in Octave or Matlab on some random data. Show that

   $$\mathbb{E}\left[L(w_i^*)\right] = 2\mathbf{y}^2 - \frac{2}{\mathbf{x}^2}\sum_i\sum_j y_i y_j x_i x_j,$$

   where $\mathbf{x}, \mathbf{y}$ denote the vectors of points and labels, respectively. Then show that $L(w^*)$ is half as much. Make your solution as simple as you can.

2. Compute all the derivatives using Backpropagation for a 3-layer neural net with an input layer, a hidden layer and one output node when the transfer function is the cumulative Gaussian density

   $$\Phi(a) := \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})\, dz$$

   and the output node is the square loss. Assume the node of the hidden layer as well as the output node each have a bias term, and all edges are present between the input layer and the hidden layer as well as between the hidden layer and the output node. Compute the derivatives of the loss wrt the weights between the hidden layer and the output node layer, the derivative of the loss wrt the weights between the input layer and hidden layer, as well as the derivatives of the loss wrt the bias terms.

   Hint: First produce a writeup when the transfer function is the sigmoid (as pretty much done in Lecture 8) and then modify it.

3. Derive the matching loss for the *rectifier* activation/transfer function $f(a) := \max(0, a)$. This function is also known as the ramp function.

   Hint: Review how the matching loss is computed when the transfer functions are the sigmoid function and the sign function: $f(a) := \text{sign}(a)$. (See material for Lecture 5.)