QBS 108: Homework 1 Key

April 14, 2020

1. From the L2-regularized cross-entropy loss function given below, write the formula for a single gradient descent update for $w_k \in \mathbf{w}$. Why would L1 be more difficult than calculating the formula for an L2 regularized model?

Cross entropy:

$$L = -\sum y_i$$ [redacted] $$ln(p(x_i)) + (1 - y_i)ln(1 - p(x_i)) + \lambda ||w||^2$$

Where;

$$p(x_i) = \frac{1}{1 + e^{-w'x}}$$

It is helpful to calculate the derivative of $p(x)$ separately.

$$p(x_i) = (1 + e^{-w'x})^{-1}$$

$$\frac{\partial p(x_i)}{\partial w_k} = -1\frac{1}{(1 + e^{-w'x})^2} * -xe^{-w'x}$$

$$\frac{\partial p(x_i)}{\partial w_k} = \frac{xe^{-w'x}}{(1 + e^{-w'x})^2}$$

$$\frac{\partial p(x_i)}{\partial w_k} = x\frac{e^{-w'x}}{(1 + e^{-w'x})^2}$$

This is 'the trick'- add and subtract 1 to rewrite the expression (alternatively, you can multiply by $\frac{1+e^{-x}}{1+e^{-x}}$ and reduce). This lets you reduce the squared term on one denominator.

$$\frac{\partial p(x_i)}{\partial w_k} = x[\frac{-1 + 1 + e^{-w'x}}{(1 + e^{-w'x})^2}]$$

$$\frac{\partial p(x_i)}{\partial w_k} = x[\frac{-1}{(1 + e^{-w'x})^2} + \frac{1}{(1 + e^{-w'x})}]$$

$$\frac{\partial p(x_i)}{\partial w_k} = \frac{x}{(1 + e^{-w'x})}[1 - \frac{1}{(1 + e^{-w'x})}]$$

Recognizing this as similar to $p(x_i)$ we can rewrite to yield:

$$\frac{\partial p(x_i)}{\partial w_k} = xp(x_i)(1-p(x_i))$$

Now we can derive the rest of the expression. From loss;

$$-L = \sum y_i ln(p(x_i)) + (1-y_i)ln(1-p(x_i)) + \lambda||w||^2$$

Plug in our derivative and use chain rule to calculate out the derivatives.

$$-\frac{\partial L}{\partial w_k} = \sum y_i \frac{\partial ln(p(x_i))}{\partial w_k} + (1-y_i)\frac{\partial ln(1-p(x_i))}{\partial w_k} + 2\lambda w_k$$

We calculated $\frac{\partial p(x_i)}{\partial w_k}$ above, and recall $\frac{\partial ln(x)}{\partial y}$ = $(\frac{1}{x})\frac{\partial x}{\partial y}$;

$$-\frac{\partial L}{\partial w_k} = \sum y_i \frac{1}{p(x_i)}\frac{\partial p(x_i)}{\partial w_k} + (1-y_i)\frac{1}{1-p(x_i)}\frac{\partial 1-p(x_i)}{\partial w_k} + 2\lambda w_k$$

Replace the term that we calculated above, and notice that $\frac{\partial 1-p(x_i)}{\partial w_k} = -\frac{\partial p(x_i)}{\partial w_k}$ as the constant drops off.

$$-\frac{\partial L}{\partial w_k} = \sum y_i(\frac{1}{p(x_i)})(xp(x_i)(1-p(x_i))) + (1-y_i)\frac{1}{(1-p(x_i))}(-xp(x_i)(1-p(x_i))) + 2\lambda w_k$$

Canceling out terms and combining yields;

$$-\frac{\partial L}{\partial w_k} = \sum y_i(x_i(1-p(x_i))) + (1-y_i)(-x_ip(x_i)) + 2\lambda w_k$$

$$-\frac{\partial L}{\partial w_k} = \sum y_i(x_i - x_ip(x_i))) + ((-x_ip(x_i)) - y_i(-x_i)p(x_i)) + 2\lambda w_k$$

$$-\frac{\partial L}{\partial w_k} = \sum (y_ix_i - y_ix_ip(x_i))) + ((-x_ip(x_i)) + (y_ix_ip(x_i)) + 2\lambda w_k$$

$$-\frac{\partial L}{\partial w_k} = \sum (y_ix_i - y_ix_ip(x_i))) + ((-x_ip(x_i)) + (y_ix_ip(x_i)) + 2\lambda w_k$$

$$-\frac{\partial L}{\partial w_k} = \sum (y_ix_i)) + ((-x_ip(x_i)) + 2\lambda w_k$$

$$-\frac{\partial L}{\partial w_k} = \sum_i^N (y_i - p(x_i))x + 2\lambda w_k$$

Making each update $w_k \to w_k'$;

$$\boxed{w_k' = w_k + \alpha[\frac{\partial L}{\partial w_k}] = w_k + \alpha[\sum_i^N (y_i - p(x_i))x_i + 2\lambda w_k]}$$

L1 is more difficult to derive than L2 because L1 penalty term is non-differentiable (continuity issues at 0).