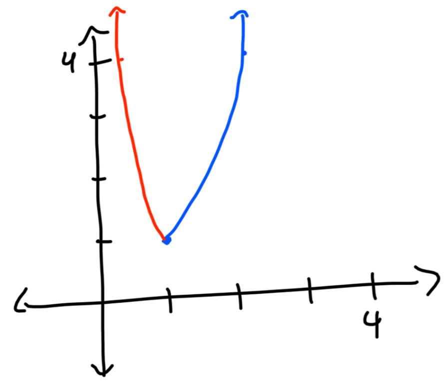

CS/ECE/ME532 Activity 20

Estimated time: 15 min for P1, 20 min for P2, 15 min for P3

1. An exponential loss function $f(w)$ is defined as

$$f(w) = \begin{cases} e^{-2(w-1)}, & w < 1 \\ e^{w-1}, & w \geq 1 \end{cases}$$

a) Is $f(w)$ convex? Why? *Hint:* Graph the function.
Yes the function is convex, it will be above any tangent line to it

b) Is $f(w)$ differentiable everywhere? If not, where not?
No the function has a sharp point at $w=1$ where it isn't differentiable

c) The "differential set" $\partial f(\boldsymbol{w})$ is the set of subgradients $\boldsymbol{v} \in \partial f(\boldsymbol{w})$ for which $f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + (\boldsymbol{u} - \boldsymbol{w})^T \boldsymbol{v}$. Find the differential set for $f(w)$ as a function of w.

$$\partial f(w) = \begin{cases} -2e^{-2(w-1)}, & w < 1 \\ e^{w-1}, & w > 1 \end{cases} \quad 0, \; w = 1$$

2. We are trying to predict whether a certain chemical reaction will take place as a function of our experimental conditions: temperature, pressure, concentration of catalyst, and several other factors. For each experiment $i = 1, \ldots, m$ we record the experimental conditions in the vector $\boldsymbol{x}_i \in \mathbb{R}^n$ and the outcome in the scalar $b_i \in \{-1, 1\}$ (+1 if the reaction occurred and −1 if it did not). We will train our linear classifier to minimize hinge loss. Namely, we solve:

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \sum_{i=1}^{m} (1 - b_i \boldsymbol{x}_i^T \boldsymbol{w})_+ \qquad \text{where } (u)_+ = \max(0, u) \text{ is the hinge loss operator}$$

a) Derive a gradient descent method for solving this problem. Explicitly give the computations required at each step. *Note:* you may ignore points where the function is non-differentiable. Initialize a start point → calculate gradient which is $g = \sum_{i=1}^{N} -d_i x_i$ (ignore non-differentiable) → update the weights using $w^{(k+1)} = w^{(k)} - \tau g$

b) Explain what happens to the algorithm if you land at a $\boldsymbol{w}^k$ that classifies all the points perfectly, and by a substantial margin.
The solution converges and will stay at that w^k.

3. You have four training samples $y_1 = 1, \boldsymbol{x}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $y_2 = 2, \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$, $y_3 = -1, \boldsymbol{x}_3 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$, and $y_4 = -2, \boldsymbol{x}_4 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$. Use cyclic stochastic gradient descent to find the first two updates for the LASSO problem

$$\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + 2||\boldsymbol{w}||_1$$

assuming a step size of $\tau = 1$ and $\boldsymbol{w}^{(0)} = 0$. Also indicate the data used for the first six updates.

$$w^{(0)} = 0$$

$$w^{(1)} = w^{(0)} + \tau\left(d_1 - x_1^T w^{(0)}\right) x_1 - \frac{2\tau}{2N}\,\mathrm{sign}(w^{(0)})$$

$$= 0 + 1\left(1 - [1\ \ -1]\, 0\right)\begin{bmatrix}1\\-1\end{bmatrix} - \frac{1}{2}(0)$$

$$= \begin{bmatrix}1\\-1\end{bmatrix}$$

$$w^{(2)} = \begin{bmatrix}1\\-1\end{bmatrix} + 1\left(2 - [1\ \ -2]\begin{bmatrix}1\\-1\end{bmatrix}\right)\begin{bmatrix}1\\-2\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1\\-1\end{bmatrix}$$

$$= \begin{bmatrix}1\\-1\end{bmatrix} + (2-3)\begin{bmatrix}1\\-2\end{bmatrix} - \begin{bmatrix}\frac{1}{2}\\-\frac{1}{2}\end{bmatrix}$$

$$= \begin{bmatrix}1\\-1\end{bmatrix} + \begin{bmatrix}-1\\2\end{bmatrix} - \begin{bmatrix}\frac{1}{2}\\-\frac{1}{2}\end{bmatrix}$$

$$= \begin{bmatrix}-\frac{1}{2}\\ \frac{3}{2}\end{bmatrix}$$

For the first six updates the data will be:

$\underline{x_1, x_2, x_3, x_4, x_1, x_2}$