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.

b) Is $f(w)$ differentiable everywhere? If not, where not?

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.

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.

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.

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.

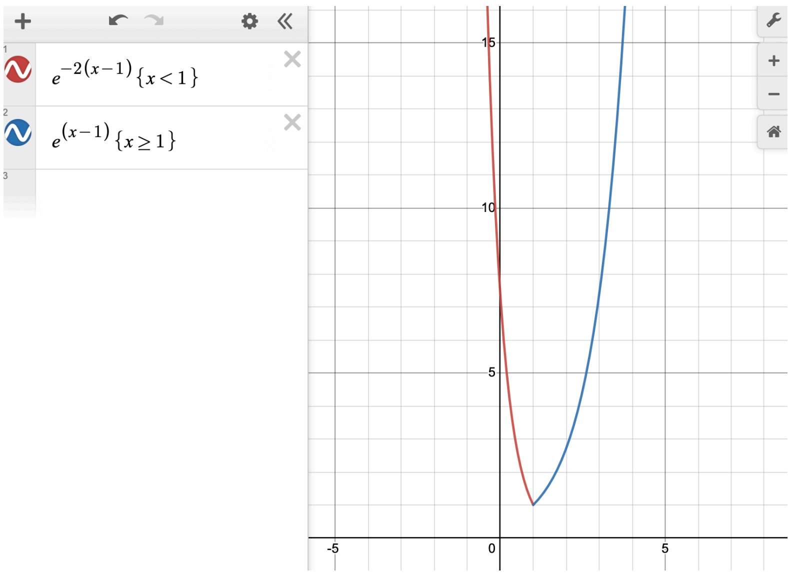
1. An exponential loss function $f(w)$ is defined as

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

a) Is $f(w)$ convex? Why? *Hint:* Graph the function.

b) Is $f(w)$ differentiable everywhere? If not, where not?

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.



1a.) Yes, $f(w)$ is convex by inspection of the graph.

1b.) Not differentiable at $x = 1$.

1c.) $\frac{d}{dx} e^{-2(w-1)}$ at $x = 1$: $\left[-2e^{-2w+2}\right]_{x=1} = -2$

$\frac{d}{dx} e^{(x-1)}$ at $x = 1$: $\left[e^{(x-1)}\right]_{x=1} = 1$

So the differential set at $w = 1$ is all lines:

$f'(w) = a(x-1) + 1$

where $a \in [-2, 1]$

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.

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.

a.) Initialize $\underline{w} = \underline{0}$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \tau \nabla_{\underline{w}} f(\underline{w}) \Big|_{\underline{w}^{(k)}}$$

- $\nabla_{\underline{w}} f(\underline{w}) \Big|_{\underline{w}^{(k)}}$ is computed as $\sum_i \left(-b_i \underline{x}_i \mathbb{1}\{ b_i \underline{x}_i^T \underline{w}^{(k)} < 1 \}\right.$

b.) The sum of the hinge loss term is 0 because all $\mathbb{1}\{ b_i \underline{x}_i^T \underline{w}^{(k)} < 1 \} = 0$

So the gradient descent doesn't move at all. $(\text{all } b_i \underline{x}_i^T \underline{w}^{(k)} \geq 1)$

$X w = y$

$4\times2 \quad 2\times1 \quad 4\times1$

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.

$i_k = 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, \ldots$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} + \tau\left(y_i^{(k)} - \underline{x}_i^{(k)T} \underline{w}^{(k)}\right) \underline{x}_i^{(k)} - \frac{\lambda \tau}{2N} \operatorname{sign}(\underline{w}^{(k)})$$

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

$$\underline{w}^{(1)} = \underline{0} + 1\left(1 - \begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)\begin{bmatrix} 1 \\ -1 \end{bmatrix} - \frac{2}{8} \operatorname{sign}(\underline{0})$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} - \begin{bmatrix} 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 3/4 \\ -5/4 \end{bmatrix}$$

$$\underline{w}^{(2)} = \begin{bmatrix} 3/4 \\ -5/4 \end{bmatrix} + 1\left(2 - \begin{bmatrix} 1 & -2 \end{bmatrix}\begin{bmatrix} 3/4 \\ -5/4 \end{bmatrix}\right)\begin{bmatrix} 1 \\ -2 \end{bmatrix} - \frac{2}{8}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$= \begin{bmatrix} 3/4 \\ -5/4 \end{bmatrix} + 1\left(2 - \left(3/4 + 10/4\right)\right)\begin{bmatrix} 1 \\ -2 \end{bmatrix} - \begin{bmatrix} 1/4 \\ -1/4 \end{bmatrix}$$

$$= \begin{bmatrix} 3/4 \\ -5/4 \end{bmatrix} + 1\left(-5/4\right)\begin{bmatrix} 1 \\ -2 \end{bmatrix} - \begin{bmatrix} 1/4 \\ -1/4 \end{bmatrix}$$

$$= \begin{bmatrix} 3/4 \\ 5/4 \end{bmatrix} + \begin{bmatrix} -5/4 \\ 5/2 \end{bmatrix} - \begin{bmatrix} 1/4 \\ -1/4 \end{bmatrix}$$

$$= \begin{bmatrix} -3/4 \\ 4 \end{bmatrix}$$