CS/ECE/ME532 Period 20 Activity

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

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

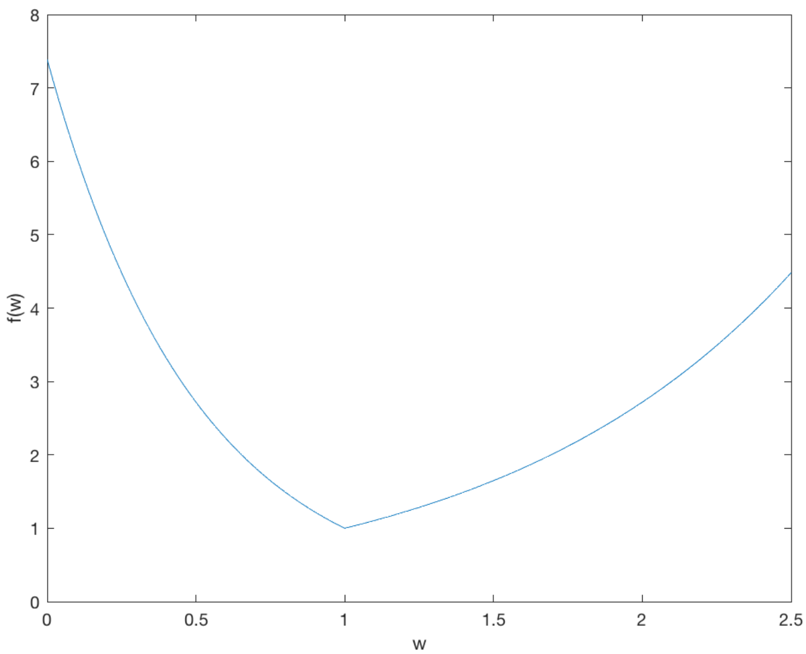
$$f(w) = \begin{cases} e^{-2(w-1)}, & w < 1 \\ e^{w-1}, & w \geq 1 \end{cases}$$

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.

SOLUTION:



a) Clearly $f(w)$ is convex as it is always above any tangent line.

b) $f(w)$ is differentiable everywhere except $w = 1$.

c) Note that $\frac{d}{dw}f(w) = \begin{cases} -2e^{-2(w-1)}, & w < 1 \\ e^{w-1}, & w > 1 \end{cases}$ At $w = 1$ the set of subgradients is $v \in [-2, 1]$. Hence we write differential set $v \in \partial f(w) = \begin{cases} -2e^{-2(w-1)}, & w < 1 \\ e^{w-1}, & w > 1 \\ \in [-2, 1], & w = 1 \end{cases}$

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.

SOLUTION:

a) Using the definition of hinge loss, we have:

$$(1 - b_i \boldsymbol{x}_i^T \boldsymbol{w})_+ = \begin{cases} 0 & \text{if } b_i \boldsymbol{x}_i^T \boldsymbol{w} > 1 \\ 1 - b_i \boldsymbol{x}_i^T \boldsymbol{w} & \text{if } b_i \boldsymbol{x}_i^T \boldsymbol{w} < 1 \end{cases}$$

Therefore the gradient is given by:

$$\nabla_{\boldsymbol{w}} (1 - b_i \boldsymbol{x}_i^T \boldsymbol{w})_+ = \begin{cases} 0 & \text{if } b_i \boldsymbol{x}_i^T \boldsymbol{w} > 1 \\ -b_i \boldsymbol{x}_i & \text{if } b_i \boldsymbol{x}_i^T \boldsymbol{w} < 1 \end{cases}$$

We can write this compactly as $\nabla_{\boldsymbol{w}}(1 - b_i \boldsymbol{x}_i^T \boldsymbol{w})_+ = -\frac{1}{2} b_i \left(1 + \text{Sign}(1 - b_i \boldsymbol{x}_i^T \boldsymbol{w})\right) \boldsymbol{x}_i$. A gradient descent algorithm involves the entire gradient and would look like:

1. initialize $\boldsymbol{w}^0$
2. compute $\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \frac{\tau}{2} \sum_{i=1}^{m} b_i \left(1 + \text{Sign}(1 - b_i \boldsymbol{x}_i^T \boldsymbol{w}^k)\right) \boldsymbol{x}_i$ for $k = 0, 1, \ldots$
3. If $||\boldsymbol{w}^{k+1} - \boldsymbol{w}^k||_2 < \text{tol}$, then stop

b) If classification is perfect, this means $b_i \boldsymbol{x}_i^T \boldsymbol{w} > 0$ for all i. If the margin is large enough so that $b_i \boldsymbol{x}_i^T \boldsymbol{w} > 1$ as well, then the gradient will be zero. So the gradient descent iterations stop.

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.

SOLUTION: We have

$$\hat{\boldsymbol{w}}^{(k+1)} = \hat{\boldsymbol{w}}^{(k)} + \tau \left(y_{i_k} - \boldsymbol{x}_{i_k}^T \hat{\boldsymbol{w}}^{(k)} \right) \boldsymbol{x}_{i_k} - \frac{\tau}{4} \mathrm{sign}\left(\hat{\boldsymbol{w}}^{(k)} \right)$$

Let $i_k = k, k = 1, 2, 3, 4$ and $i_5 = 1, i_6 = 2$ be the first six i_k. Then

$$\hat{\boldsymbol{w}}^{(1)} = y_1 \boldsymbol{x}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\hat{\boldsymbol{w}}^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \left(2 - \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ -2 \end{bmatrix} - \frac{1}{4} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -0.25 \\ 1.25 \end{bmatrix}$$