Assignment 2

Rishabh Iyer

March 9, 2020

1 Assignment Policies for CS 6301

The following are the policies regarding this assignment.

1. This assignment needs be done individually by everyone.
2. You are expected to work on the assignments on your own. If I find the assignments of a group of (two or more) students very similar, the group will get zero points towards this assignment.
3. Please use Python for writing code. You can submit the code as a Jupyter notebook
4. For the theory questions, please use Latex
5. This Assignment is for 42 points.
6. This will be due on March 14th.

2 Questions

1. (12 points total) This question will focus on proving the convexity and in some cases, finding the (sub)gradients for gradient descent like optimization.
   - (4 Point) Define the Regularized Hinge/SVM Loss as: $L_H(w) = \sum_{i=1}^n \max\{0, 1-y_i w^T x_i\} + R(w)$. Here $y_i \in \{-1, +1\}$. and $R(w)$ is a Norm. Is $L_H(w)$ convex? Why? What about the Smooth Regularized SVM Loss: $L_S(w) = \sum_{i=1}^n \max\{0, 1 - y_i w^T x_i\}^2 + R(w)$? Is $L_S(w)$ convex? Why?
   - (2 Points) Consider a 2 Layer Function: $L(w_1, w_2, b) = \sum_{i=1}^n (y_i - w_1 \max(0, w_2^T x_i + b))^2 + R(w)$. Here $y_i \in \mathbf{R}$ and $R(w)$ is a Norm. Is $L(w_1, w_2, b)$ a convex Function? Why?
   - (4 Points) Recall the Logistic Loss: $L_{\log}(w) = \sum_{i=1}^n \log(1+\exp(-y_i w^T x_i))$. Denote P as a Polyhedron, and define a Polyhedral regularization as $R(w) = f_P(|w|)$ where $f_P(w) = \max_{y \in P} y^T w$. Is the function $L(w) = \sum_{i=1}^n \log(1 + \exp(-y_i w^T x_i)) + R(w)$ convex? Why? Compute the Gradient of $L(w)$.
   - (2 Points) Recall the Soft-Max Estimator for Contextual Bandits from Assignment 1 (and the Lecture notes). Is the function $\text{SM}(\theta) = \sum_{i=1}^n \frac{r_i}{p_i} \frac{\exp(\theta^T x_i^{a_i})}{\sum_{j=1}^k \exp(\theta^T x_i^j)}$ convex? Why?
2. (6 Points) Recall that the Prox operator of a function h is $\text{prox}_h(z) = \text{argmin}_x \frac{1}{2t}||x - z||^2 + h(x)$. Compute the Prox operator for the following functions (0.75 Points each). Assume $\lambda > 0$ wherever applicable.

(a) $g_1(x) = 0$ if $x \neq 0$ and $-\lambda$ if $x = 0$.

(b) $g_2(x) = 0$ if $x \neq 0$ and λ if $x = 0$

(c) $g_3(x) = \lambda x^3$ if $x \geq 0$ and ∞ otherwise.

(d) $g_4(x) = 0$ if $0 \leq x \leq \lambda$ and ∞ otherwise.

(e) $g_5(x) = -\log x$ if $x > 0$ and ∞ otherwise

(f) $g_6(x) = \lambda|x|$

(g) $g_7(x) = a^T x + b$

(h) $g_8(x) = \lambda|x|^3$.

3. (4 Points) Compute the Projection Operator $P_C(z) = \text{prox}_{I_C}(z) = \text{argmin}_x \frac{1}{2t}||x - z||^2 + I_C(x)$ for the following constraints:

   (a) $C = \{x \in \mathbf{R}^n : x \geq 0\}$

   (b) $C = \{x \in \mathbf{R}^n : ||x - c|| \leq R\}$

   (c) $C = \{x \in \mathbf{R}^n : a^T x \geq b\}$

   (d) $C = \{x \in \mathbf{R}^n : ||x||_1 \leq R\}$

4. (15 Points) Initialize the Following Loss Functions on the Datasets located here.[1] Please see the instructions to read the features and label file in the README. Make sure both the loss functions below are implemented efficiently.

   - Logistic Loss: $L(w) = \sum_{i=1}^n \log(1 + \exp(-y_i w^T x_i)) + \lambda||w||_2^2$
   - Hinge Loss/SVMs: $L(w) = \sum_{i=1}^n \max\{0, 1 - y_i w^T x_i\} + \lambda||w||_2^2$. Here $y_i \in \{-1, +1\}$

   Implement the following variants of Gradient Descent and test them out on the Loss functions above. Compare the convergence properties (loss function value over iterations vs iterations) for each of the algorithms. Run these for a fixed number of iterations numIter = 250.

   - Gradient Descent with Fixed Learning rate α = 1e-05.
   - Gradient Descent with Armijo Line Search (v4 version in the Slides). Use parameter γ = 1e-04
   - Implement the Accelerated Gradient Descent
   - Implement Conjugate Gradient (any one version of CG - please specify which one you use)
   - Implement the Barzelia-Borwein step Gradient Descent (again use one of the two versions)

5. (5 Points) Implement the L1 Regularized Logistic Loss: $L(w) = \sum_{i=1}^n \log(1+\exp(-y_i w^T x_i)) + \lambda||w||_1$ with a value of $\lambda = 1000$. Compare Proximal Gradient Descent with standard Gradient Descent,

[1] https://github.com/rishabhk108/OptimizationML/tree/master/Assignments/data