Problem Set 1

CS 6375

Due: 2/10/2022 by 11:59pm

Note: all answers should be accompanied by explanations for full credit. All code used as part of your solutions should be included for partial credit. Late homeworks will not be accepted.

Warm-Up: Subgradients & More (15 pts)

1. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, $\lambda f(x) + (1 - \lambda)f(y) \geq f(\lambda x + (1 - \lambda)y)$. Using this definition, show that

   (a) $f(x) = wf_1(x)$ is a convex function for $x \in \mathbb{R}^n$ whenever $f_1 : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $w \geq 0$

   (b) $f(x) = f_1(x) + f_2(x)$ is a convex function for $x \in \mathbb{R}^n$ whenever $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$ are convex functions

   (c) $f(x) = \max\{f_1(x), f_2(x)\}$ is a convex function for $x \in \mathbb{R}^n$ whenever $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$ are convex functions

2. Compute a subgradient at the specified points for each of the following convex functions.

   (a) $f(x) = \max\{x^2 - 2x, |x|\}$ at $x = 0$, $x = -2$, and $x = 1$.

   (b) $g(x) = \max\{(x - 1)^2, (x - 2)^2\}$ at $x = 1.5$ and $x = 0$.

Problem 1: Perceptron Learning (45 pts)

Consider the data set (perceptron.data) attached to this homework. This data file consists of M data elements of the form $(x_1^{(m)}, x_2^{(m)}, x_3^{(m)}, x_4^{(m)}, y^{(m)})$ where $x_1^{(m)}, \ldots, x_4^{(m)} \in \mathbb{R}$ define a data point in $\mathbb{R}^4$ and $y^{(m)} \in \{-1, 1\}$ is the corresponding class label.

1. In class, we saw how to use the perceptron algorithm to minimize the following loss function.

$$\frac{1}{M} \sum_{m=1}^{M} \max\{0, -y^{(m)} \cdot (w^T x^{(m)} + b)\}$$

   What is the smallest, in terms of number of data points, two-dimensional data set containing both class labels on which the perceptron algorithm, with step size one, fails to converge? Use this example to explain why the method may fail to converge more generally.

2. Consider the following alternative loss function.

$$\frac{1}{M}\sum_{m=1}^{M}\max\{0, 1 - y^{(m)} \cdot (w^T x^{(m)} + b)\}^2$$

(a) If the data is linearly separable, does this loss function have any local optima that are not global optima?

(b) For each optimization strategy below, report the number of iterations that it takes to find a perfect classifier for the data, the values of w and b for the first three iterations, and the final weights and bias. Each descent procedure should start from the initial point

$$w^0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad b^0 = 0.$$

i. Standard subgradient descent with the step size $\gamma_t = 1$ for each iteration.

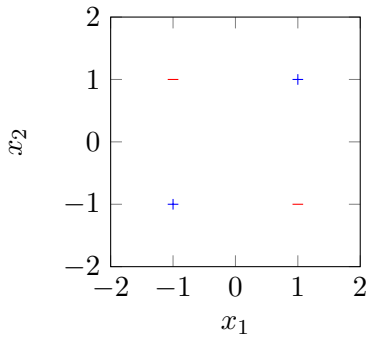
ii. Stochastic subgradient descent where exactly one component of the sum is chosen to approximate the gradient at each iteration. Instead of picking a random component at each iteration, you should iterate through the data set starting with the first element, then the second, and so on until the M^{th} element, at which point you should start back at the beginning again. Again, use the step size $\gamma_t = 1$.

iii. How does the rate of convergence change as you change the step size? Provide some example step sizes to back up your statements.

(c) Does your subgradient descent implementation with step size one always converge using this loss? Explain.

Problem 2: Separability & Feature Vectors (15 pts)

1. Consider the following data set.



Under which of the following feature vectors is the data linearly separable? For full credit, you must justify your answer by either providing a linear separator or explaining why such a

separator does not exist.

$$\text{(a) } \phi(x_1, x_2) = \begin{bmatrix} x_1 + x_2 \\ x_1 - x_2 \end{bmatrix} \qquad \text{(c) } \phi(x_1, x_2) = \begin{bmatrix} \exp(x_1) \\ \exp(x_2) \end{bmatrix}$$

$$\text{(b) } \phi(x_1, x_2) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix} \qquad \text{(d) } \phi(x_1, x_2) = \begin{bmatrix} x_1 \cdot \sin(x_2) \\ x_1 \end{bmatrix}$$

2. Suppose that you wanted to perform polynomial regression for 2-dimensional data points using gradient descent, i.e., you want to fit a polynomial of degree k to your data. Explain how to do this using feature vectors. What is the per iteration complexity of gradient descent as a function of the size of your feature representation and the number of training data points?

Problem 3: Support Vector Machines (25 pts)

For this problem, consider the data set (mystery.data) attached to this homework that, like Problem 1, contains four numeric attributes per row and the fifth entry is the class variable (either + or -). Find a perfect classifier for this data set using support vector machines. Your solution should explain the optimization problem that you solved and provide the learned parameters, the optimal margin, and the support vectors. Note, for full credit, your solution should only make use of a quadratic programming solver, i.e., you should not rely on existing SVM toolkits.