Problem Set 5

CS 4301

Due: 12/08/2020 by 11:59pm

Note: all answers should be accompanied by explanations and code for full credit. Late homeworks will not be accepted. **This homework is optional**: It can only be used to replace an existing homework grade.

Problem 1: Linear Regression (100 pts)

Suppose that we are given data points $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}^d$ with corresponding observed outcomes $y^{(1)}, \ldots, y^{(M)} \in \mathbb{R}$. In least squares regression, the aim is to find a function of the form $f : \mathbb{R}^d \to \mathbb{R}$ with $f(x) = w^T x + b$ for some $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that difference between $f(x^{(m)})$ and $y^{(m)}$ is as good as possible on average. Formally, we solve the following convex optimization problem.

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{M} \sum_{m=1}^{M} \left(y^{(m)} - w^T x^{(m)} - b \right)^2$$

Consider a penalized regression problem where we choose some $\lambda > 0$ and some regularizer h and solve the following optimization problem.

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \left[\frac{1}{M} \sum_{m=1}^{M} \left(y^{(m)} - w^T x^{(m)} - b \right)^2 + \lambda h(w) \right]$$

In this problem, we will consider the case $h(w) = \sum_{i=1}^{d} |w_i|$.

1. In Python, implement subgradient descent for the above optimization problem. That is, write a function that takes as input a matrix $X \in \mathbb{R}^{m \times n}$ whose columns are the data points a vector y of observed values, a $\lambda > 0$, an initial guess for w and b, and a number of iterations *its* and returns the result of performing subgradient descent for *its* iterations starting from the specified initial point.

2. In Python, implement proximal gradient descent (with h chosen as above) for the above optimization problem. That is, write a function that takes as input a matrix $X \in \mathbb{R}^{m \times n}$ whose columns are the data points a vector y of observed values, a $\lambda > 0$, an initial guess for w and b, and a number of iterations *its* and returns the result of performing proximal gradient descent for *its* iterations starting from the specified initial point.

3. Use the data set attached to this homework and different choices of λ to explain which method you think performs the best. Note, in the attached data, each row is of the form $({x^{(m)}}^T, y^{(m)})$.

4. What is the proximal update if $h(w) = \frac{1}{2} w^T A w + b^T w$ for some positive semidefinite matrix A?