Homework Assignment 1

SDS 385 Statistical Models for Big Data

Please upload the HW on canvas by noon Sept 27th. Please type up your homework using latex. We will not accept handwritten homeworks. Each group should write the names and EID's of the members clearly on the submission and there should be one submission for each group.

1. (10 pts) **Convex functions:** Using the definition of convex function, i.e. $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ show that the following functions are convex.
   (a) (3pts) e^x
   (b) (2pts) If $f(x)$ is convex for $x \in \Re^p$, show that so is $f(Ax+b)$ for $A \in \Re^{p\times p}$ and $b \in \Re^p$.
   (c) (2pts) If $f_i(x), i \in [k]$ are convex functions, show that the pointwise maximum, i.e. $g(x) = \max_{i\in[k]} f_i(x)$ is also convex.
   (d) (3 pts) Consider the logistic regression problem. For $x \in \Re^p$, You have
   $$y_i = \frac{1}{1+e^{-\theta^T x}}$$
      i. (1pt) Write down the log likelihood function.
      ii. (2pt) Show that this is convex. *Hint: In class we showed that the sigmoid function is convex when the argument is a scalar. You will have to extend that to allow multivariate arguments.*

2. (10 pts) **Convergence of gradient descent:** In class, we used strong convexity to show convergence of GD. In this homework we will revisit this for Lipschitz functions. To be concrete, suppose the function f is convex and differentiable and its gradient is Lipschitz condition with constant $L > 0$, i.e. we have
   $$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|_2, \qquad \text{For any } x, y$$
   In this problem we run GD for k iterations with a fixed step size $t < 1/L$.
   (a) (1 pt) First show that for any y,
   $$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$$
   (b) (3 pts) Let $y' = x - t\nabla f(x)$. Now show:
   $$f(y') \leq f(x) - t\|\nabla f(x)\|^2/2$$
   (c) (3 pts) Now show that $f(y') - f(x^*) \leq \frac{1}{2t}(\|x-x^*\|^2 - \|y'-x^*\|^2)$

(d) (3 pts) Using this, show that

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|^2}{2tk}$$

3. (20 pts) **Programming question** Read the paper "Large Scale Online Learning" (LSOL) in `http://yann.lecun.com/exdb/publis/pdf/bottou-lecun-04b.pdf` by Bottou and Le Cun et al. Follow the data generation procedure sketched in Section 5. In particular, draw two separate sets for training and testing with 1M examples each. Now generate one hundred permutations of the first set. Train each learning algorithm (given below) using various number of examples taken sequentially from the beginning of the permuted sets. Measure the resulting performance on the testing set and average over the one hundred permutations.

   (a) (5 pts) Implement the batch newton algorithm with the Gauss-Newton approximation sketched in "Efficient Back-prop" by Le Cun, Y., Bottou, L., Orr, G. B., and Muller, K.-R in 1998.

   (b) (5 pts) Implement the Online-Kalman algorithm as sketched in the LSOL paper.

   (c) (5 pts) Reproduce Figures 1,2, 3, and 4 in the LSOL paper.

   (d) (5 pts) Write a one page discussion of your findings and how that aligns with the methodological and theoretical results shown in this paper.