Homework Assignment 1

SDS 385 Statistical Models for Big Data

Please upload the HW on canvas before class Oct 11th by 10am. Please type up your homework using latex. We will not accept handwritten homeworks.

1. (10 pts) **Convex functions:** Using the definition of convex function, i.e. $f(tx + (1-t)y) \leq tf(x) + (1-t)f(\beta')$ 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

$$y \sim Bernoulli\left(\frac{1}{1+e^{-\theta^T x}}\right)$$

      i. (1pt) Write down the log likelihood function.
      ii. (2pt) Show that this is concave. *Hint: for part d you can use first/second order conditions and properties of concave functions.*

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

$$\|\nabla f(\beta) - \nabla f(\beta')\| \leq L\|\beta - \beta'\|_2, \qquad \text{For any } \beta, \beta'$$

   In this problem we run GD for t iterations with a fixed step size $\alpha < 1/L$.
   (a) (1 pt) First show that for any β',

$$f(\beta') \leq f(\beta) + \nabla f(\beta)^T(\beta' - \beta) + \frac{L}{2}\|\beta' - \beta\|^2$$

   (b) (3 pts) Let $\beta_{t+1} = \beta_t - \alpha\nabla f(\beta_t)$. Now show:

$$f(\beta_{t+1}) \leq f(\beta_t) - \alpha\|\nabla f(\beta_t)\|^2/2$$

   (c) (3 pts) Now show that $f(\beta_{t+1}) - f(\beta^*) \leq \frac{1}{2\alpha}(\|\beta_t - \beta^*\|^2 - \|\beta_{t+1} - \beta^*\|^2)$
   (d) (3 pts) Using this, show that

$$f(\beta_t) - f(\beta^*) \leq \frac{\|\beta_0 - \beta^*\|^2}{2\alpha t}$$

3. (20 pts) **Programming question** Logistic regression is a simple statistical classification method which models the conditional distribution of the class variable y being equal to class c given an input $x \in \mathbb{R}^p$. We will examine two classification tasks, one classifying newsgroup posts, and the other classifying digits. In these tasks the input x is some description of the sample (e.g. word counts in the news case) and y is the category the sample belongs to (e.g. sports, politics). The Logistic Regression model assumes the class distribution conditioned on x is log-linear. For C classes, the goal is to learn $\beta_1, \ldots \beta_{C-1} \in \mathbb{R}^p$. We use the K^{th} class as a pivot.

$$\log \frac{p(Y = 1|X = x; \beta_1, \ldots, \beta_{C-1})}{p(Y = C|X = x; \beta_1, \ldots, \beta_{C-1})} = \beta_1^T x$$

Another way to think about this is to take β_C as all zeros. Thus,

$$P(Y = c|X = x, \beta_1, \ldots, \beta_C) = \frac{\exp(\beta_c^T x)}{\sum_{j=1}^{C} \exp(\beta_j^T x)}. \quad (1)$$

Once the model is learned, one can classify a new point by picking the class that maximizes the posterior probability of belonging to that class (see Eq 1). You can measure convergence by the relative error of the concantenated parameter vector $\beta = [\beta_1^T \ldots \beta_{K-1}^T] \in \mathbb{R}^{p(C-1)}$. You should write your loss function as an average, and you can use the regularization parameter to be $1/n$.

(a) Write down the likelihood of this model for n datapoints.

Extra credit Is the logarithm of this concave? Why?

(b) For the two datasets in the provided zip file, implement the following four methods. You will use ℓ_2 regularization.

i. Gradient descent

ii. Stochastic gradient descent

iii. Newton Raphson

(c) For each method, plot the loglikelihood as a function of number of iterations.

(d) For gradient descent try different step-sizes and provide a discussion on the effect of stepsize on the convergence.

(e) For SGD , how are you choosing your step-size?

(f) Finally compute the test set error and compare the GD and SGD methods on both of the datasets.

(g) Show the test error of NR on the small dataset. How does it perform compared to the other algorithms on the small dataset?