

Homework 2: Due Friday Nov. 3, 11:59 PM

Instructions: upload a PDF report using L^AT_EX containing your answers to Canvas (remember to include your name and ID number).

Problem 1. True or False

Decide whether the following statements are true or false. **Justify your answers.**

- (a) (10 pt) If classifier A has smaller training error than classifier B , then classifier A will have smaller generalization (test) error than classifier B . *F. overfitting.*
- (b) (10 pt) It is not always good to use model with high complexity. *T. depends on what kinda data we are dealing with.*
- (c) (10 pt) Gradient descent needs to decrease the learning rate (step size) in order to converge to the optima. *False. if we are given a relatively small learning rate, w.o. decaying it, it can still converge. The rationale to decrease the learning rate is that: we want to have a higher l.r. to train faster. yet - large learning rates or constant learning rates could lead to divergent behavior. We would want to examine quickly down initial parameter, then explore deeper & narrower peaks of the loss func. also, large learning rates could lead to "jumping over" the optimal region.*

Problem 2. Multiple choice questions

Choose the correct answer and **justify your answer.**

- (a) (20 pt) Which of the following is not a possible growth function $m_{\mathcal{H}}(N)$ for some hypothesis set? (1) 2^N (2) $2^{\lfloor \sqrt{N} \rfloor}$ (3) 1 (4) $N^2 - N + 2$ (5) none of the other choices *-*

Problem 3. L2-Regularized Logistic Regression

Given a set of instance-label pairs (\mathbf{x}_i, y_i) , $i = 1, \dots, n$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$, L2-regularized logistic regression estimates the model \mathbf{w} by solving the following optimization problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) := \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) \right\} \quad (1)$$

We assume data matrix $X \in \mathbb{R}^{n \times d}$ is sparse, each column of X has n_j nonzero elements, and each row of X has d_i nonzero elements. The whole training dataset has $\text{nnz}(X) := \sum_{j=1}^d n_j = \sum_{i=1}^n d_i$ nonzero elements.

- (a) (20 pt) Derive the gradient and Hessian of $f(\mathbf{w})$.
- (b) (5 pt) What is the update rule of gradient descent (using a fixed step size η)

- (c) (5 pt) What is the time complexity of one gradient descent update?

$\mathcal{O}(nd)$, as we need to go thru gradient calculation of d dimension with n features

Newton method is a classical second order method for minimizing $f(\mathbf{w})$. The update rule for Newton method is:

$$\mathbf{w} \leftarrow \mathbf{w} + \eta \mathbf{d}^* \quad (2)$$

where $\mathbf{d}^* = -\nabla^2 f(\mathbf{w})^{-1} \nabla f(\mathbf{w})$

- (d) (5 pt) Assume we first form the Hessian matrix $\nabla^2 f(\mathbf{w})$ and then compute the Newton direction $(\nabla^2 f(\mathbf{w}))^{-1} \nabla f(\mathbf{w})$. What is the time complexity of one Newton update (eq. (2)) for L2-regularized logistic regression? (Assume n is close to d).

$$n \quad F + \frac{1}{3} n^3 = n^2 d + \frac{1}{3} n^3 = n^3 + \frac{1}{3} n^3 = \frac{4}{3} n^3$$

- (e) (5 pt) The update rule in eq. (2) can also be written as solving the following optimization problem:

$$\mathbf{d}^* = \underset{\mathbf{d}}{\operatorname{argmin}} \left\{ \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{w}) \mathbf{d} + \nabla f(\mathbf{w})^T \mathbf{d} \right\} := J(\mathbf{d}) \quad (3)$$

Proof the optimal solution of (3) is $-(\nabla^2 f(\mathbf{w}))^{-1} \nabla f(\mathbf{w})$.

$$\nabla^2 f(\mathbf{w}) \mathbf{d} + \nabla f(\mathbf{w})^T = 0$$

- (f) (10 pt) Since the matrix inversion would be numerically unstable in certain condition, what is the alternative solution to get $(\nabla^2 f(\mathbf{w}))^{-1} \nabla f(\mathbf{w})$ without matrix inversion?

$$A\mathbf{x} = \mathbf{b}$$

pseudo-inverse

$$(a) \quad \nabla = W + C \sum_{i=1}^n \frac{-y_i}{1 + \exp(-y_i W^T x_i)} x_i \quad x_i \in \mathbb{R}^d$$

$$\therefore \nabla = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_d \end{bmatrix} + C \sum_{i=1}^n$$

$$\begin{bmatrix} \frac{\exp(-y_i W^T x_i) (-y_i)}{[1 + \exp(-y_i W^T x_i)]} x_{i,1} \\ \frac{\exp(-y_i W^T x_i) (-y_i)}{[1 + \exp(-y_i W^T x_i)]} x_{i,2} \\ \vdots \\ \frac{\exp(-y_i W^T x_i) (-y_i)}{[1 + \exp(-y_i W^T x_i)]} x_{i,d} \end{bmatrix}$$

$$\nabla^2 = \begin{bmatrix} \frac{\partial f_1}{\partial W_1} & \frac{\partial f_1}{\partial W_2} & \dots & \frac{\partial f_1}{\partial W_d} \\ \frac{\partial f_2}{\partial W_1} & \frac{\partial f_2}{\partial W_2} & \dots & \frac{\partial f_2}{\partial W_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_d}{\partial W_1} & \frac{\partial f_d}{\partial W_2} & \dots & \frac{\partial f_d}{\partial W_d} \end{bmatrix}$$

$$\frac{\exp(-y_i W^T x_i) (-y_i)}{[1 + \exp(-y_i W^T x_i)]} x_{i,1}$$

$$= J + C \sum_{i=1}^n \begin{bmatrix} \left\{ \frac{-y_i x_{i,1} \exp(-y_i W^T x_i)}{[1 + \exp(-y_i W^T x_i)]} (-y_i x_i) + \frac{\exp(-y_i W^T x_i) (-y_i) x_{i,1}}{[1 + \exp(-y_i W^T x_i)]^2} \exp(-y_i W^T x_i) (-y_i x_i) \right\}^T \\ \vdots \\ \left\{ \frac{-y_i x_{i,d} \exp(-y_i W^T x_i)}{[1 + \exp(-y_i W^T x_i)]} (-y_i x_i) + \frac{\exp(-y_i W^T x_i) (-y_i) x_{i,d}}{[1 + \exp(-y_i W^T x_i)]^2} \exp(-y_i W^T x_i) (-y_i x_i) \right\}^T \end{bmatrix}$$

$$(b) \quad w \leftarrow w + \eta d$$

where

$$d = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} + C \sum_{i=1}^n$$

$$\begin{bmatrix} \frac{\exp(-y_i w^T x_i) (-y_i)}{[1 + \exp(-y_i w^T x_i)]} x_{i1} \\ \frac{\exp(-y_i w^T x_i) (-y_i)}{[1 + \exp(-y_i w^T x_i)]} x_{i2} \\ \vdots \\ \frac{\exp(-y_i w^T x_i) (-y_i)}{[1 + \exp(-y_i w^T x_i)]} x_{id} \end{bmatrix}$$

(c)

$$(Ax - b)^T (Ax - b)$$

$$\times A^T A x - \cancel{A^T b} - \cancel{b^T A x} + \cancel{b^T b}$$

$$(A^T A) x = A^T b$$