COMP5212: Machine Learning **Fall 2022**

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

Instructions: upload a PDF report using LaTeX 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.

(b) (10 pt) The VC dimension is always equal to the number of parameters in the model.

(c) (10 pt) For non-convex problems, gradient descent is guaranteed to converge to the global minimum.

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 $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$, L2-regularzied logistic regression estimates the model $\boldsymbol{w}$ by solving the following optimization problem:

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

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 $f(\boldsymbol{w})$.

(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?

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

$$\boldsymbol{w} \leftarrow w - \eta \boldsymbol{d}^* \tag{2}$$

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

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

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

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

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

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