CS260: Machine Learning Algorithms **Winter 2019**

Homework 2: Due Tuesday Feb. 26, 11:59 PM

Instructions: upload a PDF report using LaTeX containing your answers to CCLE (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) $2^{\sqrt{N}}$ (3) 1 (4) $N^2 - N + 2$ (5) none of the other choices

Problem 3. Proximal Gradient Descent

Consider solving the following problem

$$\min_{\boldsymbol{w}} \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1,$$

where $X \in R^{n\times d}$ is the feature matrix (each row is a feature vector), $\boldsymbol{y} \in R^n$ is the label vector, $\|\boldsymbol{w}\|_1 := \sum_i |w_i|$ and $\lambda > 0$ is a constant to balance loss and regularization. This is known as the Lasso regression problem and our goal is to derive the "proximal gradient method" for solving this.

- (10 pt) The gradient descent algorithm cannot be directly applied since the objective function is non-differentiable. Discuss why the objective function is non-differentiable.
- (30 pt) In the class we showed that gradient descent is based on the idea of function approximation. To form an approximation for non-differentiable function, we split the differentiable part and non-differentiable part. Let $g(\boldsymbol{w}) = \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2$, as discussed in the gradient descent lecture we approximate $g(\boldsymbol{w})$ by

  $$g(\boldsymbol{w}) \approx \hat{g}(\boldsymbol{w}) := g(\boldsymbol{w}_t) + \nabla g(\boldsymbol{w}_t)^T(\boldsymbol{w} - \boldsymbol{w}_t) + \frac{\eta}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2.$$

  In each iteration of proximal gradient descent, we obtain the next iterate ($\boldsymbol{w}_{t+1}$) by minimizing the following approximation function:

  $$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}} \hat{g}(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1.$$

  Derive the close form solution of $\boldsymbol{w}_{t+1}$ given $\boldsymbol{w}_t, \nabla g(\boldsymbol{w}_t), \eta, \lambda$. What's the time complexity for one proximal gradient descent iteration?