CSE 417T: Homework 3

Due: March 25 (Wednesday), 2020

Notes:

- Please submit your homework via Gradescope and check the submission instructions.
- Make sure you **specify the pages for each problem correctly**. You **will not get points** for problems that are not correctly connected to the corresponding pages.
- Homework is due **by 11:59 PM on the due date.** Remember that you may not use more than 2 late days on any one homework, and you only have a budget of 5 in total.
- Please keep in mind the collaboration policy as specified in the course syllabus. If you discuss questions with others you **must** write their names on your submission, and if you use any outside resources you **must** reference them. **Do not look at each others' writeups, including code.**
- There are 5 problems on 2 pages in this homework.

Problems:

1. (30 points) The weight decay regularizer is also called L_2 regularizer, since $\vec{w}^T\vec{w}$ is the square of the 2-norm of the weight vector $\|\vec{w}\|_2 = \sqrt{\sum_{i=0}^{d} w_i^2}$. Another common regularizer is called L_1 regularizer, since 1-norm ($\|\vec{w}\|_1 = \sum_{i=0}^{d} |w_i|$) is used as the regularizer.

   Below are the definitions of the two regularizations [1]

   - L_1 regularization: $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda\|\vec{w}\|_1$
   - L_2 regularization: $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda\vec{w}^T\vec{w}$

   (a) Answer LFD Problem 4.8.

   (b) Similar to Problem 4.8, we now aim to derive the update rule of gradient descent for minimizing the augmented error with L_1 regularizer. Note that the gradient of 1-norm is not well-defined at 0. To address this issue, we can utilize the *subgradient* idea defined as follows:

   $$\frac{\partial}{\partial w_i}\|\vec{w}\|_1 = \begin{cases} +1 & \text{if } w_i > 0 \\ \text{any value in } [-1, 1] & \text{if } w_i = 0 \\ -1 & \text{if } w_i < 0 \end{cases}$$

   To simplify the discussion, we let $\frac{\partial}{\partial w_i}\|\vec{w}\|_1 = 0$ when $w_i = 0$. Please write down the update rule of gradient descent for L_1 regularization. (You can define a $sign()$ function that returns $+1, 0, -1$ when the input is positive, zero, negative, respectively).

[1] When applying these regulerizations to linear regression, they are called Ridge Regression (L_2 regularizer) and Lasso Regression (L_1 regularizer) respectively.

Note that, we can use *truncated* gradient [1] in practical implementation for L_1 regularizer. Let $\vec{w}'(t+1) \leftarrow \vec{w}(t) - \eta \nabla E_{\text{in}}(\vec{w}(t))$ be the update rule of gradient descent without regularization. The update rule for L_1 regularization should be in the form of

$$\vec{w}(t+1) \leftarrow \vec{w}'(t+1) + \text{additional term}$$

Truncated gradient works as follows: At each step t, you first perform the update as what you derive above. Then for each i, if $w_i(t+1)$ and $w_i'(t+1)$ have different signs and when $w_i'(t+1) \neq 0$, we set the update $w_i(t+1)$ to 0 (i.e., we *truncate* the update if the additional term makes the new weight change signs).

(c) Update your implementation of logistic regression in HW2 to include the L_1 and L_2 regularizers (use truncated gradient for L_1 regularizer). Examine different regularization strengths $\lambda = 0.001, 0.01, 0.05, 0.1$ (please feel free to try more choices of λ). Train your models on `clevelandtrain.csv`. For each trained model, report (1) the classification error on `clevelandtest.csv` and (2) the number of 0s in your learned weight vector. Describe your observations on the property of the L_1 regularizer.

For the other training parameters, please use the following. Normalize input data. Set learning rate $\eta = 0.01$. The maximum number of iterations is 10^6. Terminate learning if the magnitude of every element of the gradient (of E_{in}) is less than 10^{-6}. In calculating classification error, classify the data using a cutoff probability of 0.5.

You do not need to submit your code for this question.

2. (10 points) LFD **Exercise** (not Problem) 4.5

3. (25 points) LFD Problem 4.25 (a) to (c)

4. (15 points) LFD Problem 5.4. Note that the problem makes a simplifying definition: a stock is called "profitable" if it went up half of the days, and whether the stock goes up or down is a random draw from an unknown distribution that associates with how good the stock is. While this is not accurate in practice, please use this as the definition for your discussion.

5. (20 points) You have been hired by a biologist to learn a decision tree to determine whether a mushroom is poisonous. You have been given the following data:

Color	Stripes	Texture	Poisonous?
Purple	No	Smooth	No
Purple	No	Rough	No
Red	Yes	Smooth	No
Purple	Yes	Rough	Yes
Purple	Yes	Smooth	Yes

Use ID3 to learn a decision tree from the data (this is a written exercise – no need to code it up):

(a) What is the root attribute of the tree? Show the computations.

(b) Draw the decision tree obtained using ID3.

References

[1] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777801, 2009.