CSE 417T: Homework 3 Solution Sketches

March 5, 2022

Note: These are not intended to be comprehensive, just to help you see what the answers should be.

Regularizations For part (a) First note that

$$\nabla E_{\text{aug}}(\vec{w}) = \nabla(E_{\text{in}}(\vec{w} + \lambda \vec{w}^T \vec{w}) = \nabla E_{\text{in}}(\vec{w}) + 2\lambda \vec{w}$$

Thus, the weight update for gradient descent ($-\eta \nabla E_{\text{aug}}(\vec{w})$) becomes:

$$\vec{w} - \eta \nabla E_{\text{in}}(\vec{w}) + 2\eta\lambda\vec{w} = \vec{w}(1 - 2\eta\lambda) - \eta \nabla E_{\text{in}}(\vec{w})$$

Similarly for part (b), let $sign(\vec{w})$ be a vector denoting the sign of each element (let the sign of 0 be 0). We get that

$$\vec{w}(t+1) \leftarrow \vec{w}(t) - \eta\lambda sign(\vec{w}(t)) - \eta \nabla E_{\text{in}}(\vec{w}(t))$$

For part (c), below are the reference results.

	L_1 Regularizer		L_2 Regularizer	
	Binary Error on Test Set	# 0 in weights	Binary Error on Test Set	# 0 in weights
$\lambda = 0$	-	-	0.1028	8
$\lambda = 0.0001$	0.0981	8	0.1028	8
$\lambda = 0.001$	0.0935	15	0.0935	8
$\lambda = 0.005$	0.0888	26	0.0981	8
$\lambda = 0.01$	0.0794	36	0.0981	8
$\lambda = 0.05$	0.1028	52	0.1168	8
$\lambda = 0.1$	0.1355	57	0.1215	8

Main takeaways: [There might be differences on the number of 0s depending whether students give some tolerance range to define 0. Please focus on the trend during grading.] L_1 regularizer tends to generate weights vectors with 0s. This is helpful in reducing dimensions in high-dimensional learning.

Exercise 4.5 (a) Γ is the identify matrix.
(b) Γ is a row vector with every element to be 1.

Problem 4.25 (a)-(c) For part (a), not necessarily. For example, a learner with low validation error on a small validation set may just have gotten lucky compared with a learner with a larger validation error on a larger validation set.

For part (b), yes, because all the validation sets are the same.

For part (c), let's index the error terms by the number of the learning model. Then the event that $E_{\text{out}}(m^*) > E_{\text{val}}(m^*) + \epsilon$ implies that either $E_{\text{out}}(1) > E_{\text{val}}(1) + \epsilon$ or $E_{\text{out}}(2) > E_{\text{val}}(2) + \epsilon$, …, or $E_{\text{out}}(M) > E_{\text{val}}(M) + \epsilon$. Applying the union bound, similarly to what we did for the multiple hypotheses version of the Hoeffding inequality, we get:

$$\Pr[E_{\text{out}}(m^*) > E_{\text{val}}(m^*) + \epsilon] \leq \sum_{m=1}^{M} \Pr[E_{\text{out}}(m) > E_{\text{val}}(m) + \epsilon] \leq \sum_{m=1}^{M} e^{-2\epsilon^2 K_m}$$

(where the second inequality comes from the one-sided version of Hoeffding's inequality). From the definition of $\kappa(\epsilon)$, $e^{-2\epsilon^2\kappa(\epsilon)} = \frac{1}{M}\sum_{m=1}^{M} e^{-2\epsilon^2 K_m}$. Therefore:

$$\Pr[E_{\text{out}}(m^*) > E_{\text{val}}(m^*) + \epsilon] \leq Me^{-2\epsilon^2\kappa(\epsilon)}$$

5.4 For (a), note that you didn't pick the best out of 500, you effectively picked the best out of 50,000 because the S&P first picked the 500 for you. Therefore, you should use $M = 50,000$ in the bound, which would make the bound basically useless.

For (b), again we are getting sample selection bias (in this case by snooping). We can't use the set of stocks currently in the S&P 500, we have to use those that were in there when the buy and hold strategy would have started, 50 years ago. There is really nothing we can say about buy and hold for general stock trading based on the current S&P 500, we'd have to be sampling from the right set to start with.