CS 329T: Homework 1 Written Answers

Trustworthy Machine Learning Spring 2021

Name: <Name> **SUNet Id:** <SUNet>

Written Exercise 1. *Derive the gradient of loss in terms of* $\boldsymbol{W}$ *and* $\boldsymbol{b}$*:* $\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{W}}$ *and* $\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{b}}$*. Show your work and make sure the dimensions of your vectors are consistent with the ones in the problem description.*

Solution

Written Exercise 3. *Given a pre-softmax logistic regression model* $f : \boldsymbol{x} \mapsto \left(\boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{b}\right)$*, an input* $\boldsymbol{x}$ *and, class index c, define an attribution* $\boldsymbol{a}$ *for* $f(\boldsymbol{x})_c = y$ *that is complete for all baselines.*

Solution

Written Exercise 6. *Is it possible to implement the attack in the prior exercise given access to post-softmax probabilities? If no, how would you adjust the exercise to make it possible while still being able to call it a "model stealing" attack?*

Solution

Written Exercise 8. *We can use* $L_*(\boldsymbol{x} - \boldsymbol{x'})$*, for various bases* $*$*, to measure how close the adversarial example is to the original. Pick a base from* $* \in 0, 1, 2, \infty$ *and describe a pair of images which are different according to the* L_* *but are actually close when it comes to human perception (i.e. they are close to indistinguishable).*

Solution