Labs
Machine Learning Course
Fall 2024

EPFL
School of Computer and Communication Sciences
Nicolas Flammarion & Martin Jaggi
www.epfl.ch/labs/mlo/machine-learning-cs-433

Problem Set 09, Nov 14, 2024 (Adversarial Robustness)

Security and robustness of machine learning models have been often overlooked, for the sake of greater performance and accuracies. However, it is usually quite easy for an adversary to create inputs (e.g., images) that fool an ML model into thinking they are something else, while preserving the semantic content of the original input.

The goal of this exercise is to better understand how to generate adversarial examples in practice, use them in adversarial training to get a more robust model, and to check what adversarial examples correspond to in the simple case of linear models.

Problem 1 (Adversarial training for linear models):

It can be often very insightful to analyze what a method corresponds to in a simple setting of linear models.

Assume we have input points $\boldsymbol{x}_i \in \mathbb{R}^d$ and binary labels $y_i \in \{-1, 1\}$. Let ℓ be a monotonically decreasing margin-based loss function, for example the hinge loss $\ell(z) = \max\{0, 1 - z\}$ or logistic loss $\ell(z) = \log(1 + \exp(-z))$ that you have seen before.

Consider the adversarial training objective for a linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ with respect to ℓ_2 adversarial perturbations:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2 \leq \varepsilon} \ell\left(y_i \cdot \boldsymbol{w}^\top \hat{\boldsymbol{x}}_i\right).$$

- Find a closed-form solution of the inner maximization problem $\max_{\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2 \leq \varepsilon} \ell\left(y_i \cdot \boldsymbol{w}^\top \hat{\boldsymbol{x}}_i\right)$ and the minimizer $\hat{\boldsymbol{x}}_i^\star$.
- In case of the hinge loss, $\ell(z) = \max\{0, 1 - z\}$, what is the connection between ℓ_2 adversarial training and the primal formulation of the soft-margin SVM?
- What if instead of ℓ_2 adversarial training, we performed ℓ_∞ adversarial training, how would the solution of the inner maximization problem change? Does the maximizer for ℓ_∞-perturbations resemble the Fast Gradient Sign Method (FGSM)?

Problem 2 (Adversarial training on MNIST):

In this problem you will:

1. Learn how to make small modifications in handwritten digit images that result in dramatic errors by ML models. However, humans can still recognize these adversarial examples.
2. Implement a simple defense against this attack.

Setup It is the easiest to run this notebook in Google Colab. You can make use of a free GPU there to train the models faster. If you want to run the notebook locally, you can also use `template/ex09.ipynb`. However, expect to have much longer running time if you don't have GPUs.

1. Open the colab link for the lab 09:
   `https://colab.research.google.com/github/epfml/ML_course/blob/master/labs/ex09/template/ex09.ipynb`
2. To save your progress, click on *"File > Save a Copy in Drive"* to get your own copy of the Notebook.
3. Click 'connect' on top right to make the notebook executable (or 'open in playground')
4. Start solving the missing parts.