Principled Approaches to Robust Machine Learning

September 25, 2019

- Tuesdays & Thursdays, 10:00 AM —11:30 AM. Room: G04.
- **Course website:** `https://jerryzli.github.io/robust-ml-fall19.html`
- **Mailing list:** `cse599m@cs.washington.edu`
- **Grading.** Grading will be 50% homework and 50% final project. But being a topics class, we intend to be pretty relaxed about this. You should be here because you are interested!
- **Topics.** We intend to have roughly 3 "units" on related topics:
  - **Learning in the presence of outliers.** Techniques for learning when our training dataset is corrupted by worst-case noise. This includes:
    - Robust statistics: Robust mean estimation, robust covariance estimation.
    - List learning: Learning when there is an overwhelming fraction of corrupted data.
    - Data poisoning attacks / defenses: Techniques for supervised learning with outliers.
    - Backdoor attacks: Watermarking attacks and defenses for neural networks.
  - **Adversarial examples.** Famously, neural network image classifiers can be fooled at test time by perturbing a test image by an imperceptible amount. We will discuss:
    - Empirical attacks: PGD and variants.
    - Empirical defenses: Adversarial training, pretraining, semi-supervision.
    - Theoretical models: The four worlds hypothesis.
    - Certified defenses: Exact certification, convex relaxations, and randomized smoothing.
  - **Model misspecification.** Understanding when algorithms designed for a specific generative model will still work when the true data may not come from something else.
    - Semi-random models: When "helpful" adversaries can hurt.
    - Truncated statistics: learning from a subset of the distribution.
    - Distributional shift: How well do models transfer from one distribution to another?

    Depending on time / how organized the instructor is, we may also have guest lecturers near the end of the quarter.
- **Prerequisites.** We will assume mathematical maturity and comfort with algorithms, probability, and linear algebra. Background in machine learning will be helpful but should not be necessary.