**Title: Study of black box attacks on Machine Learning model and strategies to mitigate them**

**Project Partners:** (1) Shreeraj Shah (sbs577)

(2) Prachi Gupta (pg1647)

**Abstract:** Machine learning(ML) models today are vulnerable to various types of attacks. In this work, we will study a particular category of attack known as membership inference attack and show how ML models are susceptible to leaking secure information under such attacks. Given a data record and a black box access to a ML model, we present a framework to deduce whether the data record was part of the model’s training dataset or not. We achieve this objective by creating an attack ML model which learns to differentiate the target model’s predictions on its training data from target model’s predictions on data not part of its training data. In other words, we solve this membership inference problem by converting it into a binary classification problem. We also study mitigation strategies to defend the ML models against the attacks discussed in this work.

We evaluate our method on real world datasets: (1) CIFAR-10, (2) CIFAR-100 and (3) Purchases using classification as the task performed by the target ML models built on these datasets.

**Milestones:**

* Design and train target model: We will use neural networks as target models so that target models have sufficient sophistication
* Collect data for attack model: To train our attack ML model, we will create dummy models similar to the target model but we will know both the train and test data for these dummy models. So the output predictions from these dummy models will serve as training data to our attack model.
* Design and train dummy models: We will create dummy models in the same way as our target model but number of dummy models will be more than 1.
* Design and train attack model: A binary classification model
* Evaluate the attack model: We will randomly select equal number of samples from training and test data of target model. Then we will feed target model’s out on all these samples to the attack model to determine the efficacy of the attack.
* Implementing mitigation strategies
* Evaluating the effect of the mitigation techniques on success of the attack

**References:**

* Paper: <https://arxiv.org/pdf/1610.05820.pdf>
* CIFAR dataset: <https://www.cs.toronto.edu/~kriz/cifar.html>
* UCI Adult (Census Income): https://archive.ics.uci.edu/ml/datasets/adult