Ethics of Data Science – Part III

Adversarial learning – evasion attacks and impact on safety

Dr. Chris Anagnostopoulos, Hon. Senior Lecturer

Approaches to create adversarial examples

Evasion techniques were common in early ML systems for fraud and spam detection (e.g., keeping fraudulent transactions below a certain threshold, or personalising emails to evade spam filters). Modern-day approaches use ML, rather than intuition, to build adversarial examples. Key methods include:

- Carlini and Wagner (C&W)[1]
- Fast Gradient Sign Method (FGSM)[2]
- DeepFool (DF)[3]
- Projected Gradient Descent (PGD)[4]

1. Carlini, Nicholas; Wagner, David (2017-03-22). "Towards Evaluating the Robustness of Neural Networks". arXiv:1608.04644
2. Goodfellow, Ian J.; Shlens, Jonathon; Szegedy, Christian (2015-03-20). "Explaining and Harnessing Adversarial Examples". arXiv:1412.6572
3. Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
4. Madry, Aleksander; Makelov, Aleksandar; Schmidt, Ludwig; Tsipras, Dimitris; Vladu, Adrian (2019-09-04). "Towards Deep Learning Models Resistant to Adversarial Attacks". arXiv:1706.06083

Fast Gradient Sign Method is a seminal paper

A large part of the literature was inspired by FGSM, which perturbs each example in the direction of the gradient evaluated at the current model parameters (i.e., in the direction of maximal change).

FGSM has also been applied in iterative manners to increasingly improve evasive ability, with Projected Gradient Descent falling in this broader family of methods. The method by C&W is simple and very powerful, and it assumes access just to the probability scores for arbitrary input queries, and tries to maximise the confidence with which the classifier misclassifies each example perturbation subject to minimising the distance of the perturbed example from the original.

FGSM attack ← `model_obj.predict_proba(X)` → C&W attack

Relationship to statistical robustness

Adversarial examples are a form of worst-case inliers/outliers or dataset shift, so that statistical robustness will be to a smaller or larger extent protective. Conversely, defending against adversarial examples is also likely to lead to ML models that are also more robust to chance outliers/dataset shift.

1. Adversarial training: the ML model is exposed to cycles of training followed by adversarial example generation, followed by training on an augmented dataset that includes the adversarial examples.[1]
2. Anomaly detection: although adversarial examples are meant to be as close as possible to genuine examples, they can still be revealed to be outliers in a suitable projection of the input space.[2] Suspected adversarial examples may still be naturally occurring outliers, however.
3. Distillation[3]: distillation relies on the observation that transfer learning by design creates models that can generalize in the presence of dataset shift, and are hence also robust to adversarial examples.

1. Shafahi, Ali, et al. "Adversarial training for free!." *Advances in Neural Information Processing Systems* 32 (2019).
2. Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." *international conference on machine learning*. PMLR, 2019.
3. Papernot, Nicolas, et al. "Distillation as a defense to adversarial perturbations against deep neural networks." *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016.

Try to preserve the black box advantage

In general, white box evasion is easier than black box evasion. This means that federated learning is also an important defence against evasion attacks in that it conceals aspects of the trained model.

However, adversaries are able to generate synthetic datasets by collecting independently training examples and asking the ML model to classify them, then training an auxillary model on this set that will have a local gradient similar to the target model, which can then be exploited for adversarial examples. In a sense, this then becomes a combined extraction and evasion attack.

Several techniques exist to make black box models resilient to this type fall under *gradient masking*[1], which uses models that are locally smooth or piecewise constant (e.g., nearest neighbours).

1. Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017.