ADVERSARIAL MACHINE LEARNING

TAXONOMY OF ADVERSARIAL ATTACKS

9

- Based on the intention of the adversary
  - Targeted
    - Multi class (Goal of adversary to have an adversarial example getting classified into a particular class)
      - e.g. Classify stop sign as speed up
    - Fooling anomaly detection (special case - binary classification)
  - Non-Targeted
    - Goal of adversary is to just break the classifier

TAXONOMY OF ADVERSARIAL ATTACKS

- Based on the knowledge of the adversary
    - White Box Attack
        - Adversary has access to the trained model and weights
    - Black Box Attack
        - Adversary has only access to the output of the model e.g. API

TAXONOMY OF ADVERSARIAL ATTACKS

- Based on the intention of the adversary
  - Targeted
    - Multi class (Goal of adversary to have an adversarial example getting classified into a particular class)
      - e.g. Classify stop sign as speed up
    - Fooling anomaly detection (special case - binary classification)
  - Non-Targeted
    - Goal of adversary is to just break the classifier