1 Objetivos

- Comprender los ataques basados en extracción del conocimiento y evasión de clasificadores
- Utilizar el framework Adversarial Robustness ToolBox para atacar modelos
- Entender y aplicar la defensa de modelos ante ataques de extracción y evasión

2 Preámbulo

Seguridad en modelos de data science

Los modelos de machine learning y deep learning son activos que están sujetos a los ciberataques, como cualquier otro activo digital.

Los ataques varían según su propósito y clasificación. En los ataques de caja negra, el adversario no conoce los detalles de implementación del modelo, en tanto que, en los ataques de caja blanca, el adversario si conoce los detalles.

Robo de conocimiento

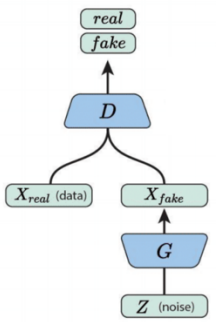
Las empresas que utilizan ML/DL para apoyar sus procesos de negocio invierten una gran cantidad de recursos en la investigación, desarrollo e implementación de sus modelos, y luego ofrecen un servicio pagado de clasificación a través de una API, por ejemplo.

Con esta información, se puede realizar un ataque de caja negra/blanca que consiste en utilizar un dataset y obtener la clasificación y confidencia a través de la API. Aun si utilizar la API tiene un costo, este será mínimo en comparación con los resultados. La idea es obtener las etiquetas y confidencia para cada una de las observaciones, y con ello, ¡entrenar un modelo propio! (Ataque de extracción).

Dado que el nuevo modelo se entrenará en la forma en que el modelo objetivo clasifica, este tendrá resultados muy similares, sin invertir la gran cantidad de recursos que el modelo original.

Engaño de modelos

Generative Adversarial Networks (GANs) son una forma de construir un modelo generativo al tener dos redes neuronales compitiendo una contra otra.



Una red toma el papel del generador (**G**), que convierte ruido aleatorio en imitaciones de data, intentando engañar al discriminador.

La otra red toma el papel del discriminador (**D**), que trata de distinguir data real de data falsa creada por el generador. Esto se puede aprovechar para realizar ataques con data que engañen a los modelos de clasificación.

3 Desarrollo

EL proyecto consiste en el desarrollo de dos ataques, utilizando el framework Adversarial Robustness ToolBox, originalmente desarrollador por IBM, y donado recientemente a The Linux Foundation.

https://adversarial-robustness-toolbox.org/

Este framework contiene módulos de ataque y defensa, métricas, etc; y soporta framewroks como TensorFlow, Keras, Scikit-Learn, PyTorch, etc., todo tipo de data (imágenes, tablas, video, etc.) y tareas de machine learning (clasificación, generación, etc.)

El modelo objetivo del ataque será el modelo desarrollado en el laboratorio #7 – Clasificación de malware con DL.

Ataque de extracción

1. Se debe mostrar un entrenamiento y predicción del modelo original
2. Se debe crear un modelo robado a partir del modelo original, entrenado con data completamente aleatoria, con los ataques Copycat CNN y KnockoffNets.
3. Mostrar en una gráfica el accuracy del modelo robado, contra el tamaño del dataset

4. Preparar una capa de defensa y crear un nuevo modelo protegido
5. Repetir el paso 2, pero con el modelo protegido
6. Mostrar en una gráfica el accuracy del modelo protegido, contra el tamaño del dataset
7. Describa lo que sucedió en el paso 6, detalle sus conclusiones

Ataque de evasión

1. Se debe mostrar un entrenamiento y predicción del modelo original
2. Utilizando el ataque de evasión Fast Gradient Method genere observaciones falsas
3. Muestre la cantidad de observaciones generadas que fueron clasificadas correcta e incorrectamente
4. Explique con detalle los resultados obtenidos. ¿Cómo podría proteger su modelo ante este tipo de ataques?
5. Realce las modificaciones que considere pertinentes para hacer su modelo robusto contra este tipo de ataques (puede modificar el modelo original, así como el dataset de entrenamiento). Repita los pasos 1 al 3 con el nuevo modelo robusto y demuestre que es menos susceptible a equivocarse con observaciones generadas

4 Calificación

- El proyecto será desarrollado de forma individual, parejas o tríos, con la restricción de que deben ser los mismos que en el laboratorio # 7 – Clasificación de malware con DL.
    - Se debe entregar el link al repositorio en Github del proyecto que debe incluir:
        - Jupyter Notebook
- La fecha de entrega máxima será el lunes **30 de mayo a las 17:20 horas.** Ese día se realizarán las presentaciones y calificaciones de los proyectos. Si un grupo termina antes de esa semana lo puede hacer, solamente debe notificar por correo para planificar la calificación.
- Plagio parcial o total anula el proyecto, y se elevará el caso a la Dirección para las sanciones administrativas.

5 Rúbrica

Aspectos	**Valor**
Ataque Extracción	40%
Defensa contra extracción	15%
Ataque Evasión	10%
Defensa contra evasión	15%
Conclusiones	20