

Unidad 6: Aprendizaje automático adversarial y retos de seguridad

Adarsh KUMAR

Universitat Politècnica de Catalunya
Department of Computer Science

Coordinador del proyecto:
Prof. Ilker Demirkol

Proyecto MERiT
3 de septiembre de 2025

Co-funded by the European Union

Diapositiva 2

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística

Esquema

1. Ataques adversariales
2. Detección de anomalías
3. Medidas de divergencia estadística

Diapositiva 3

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Ataques adversariales)

Aprendizaje automático adversarial: visión general

- Estudia las vulnerabilidades de los modelos de ML ante entradas manipuladas intencionadamente.
- Los ejemplos adversariales son entradas modificadas para engañar a los modelos de ML.
- Las modificaciones suelen ser imperceptibles, pero provocan predicciones incorrectas.
- Riesgo clave en los sistemas de ML para ciberseguridad.
- Los atacantes explotan las debilidades del modelo para evadir la detección.
- Requiere defensas y un diseño de modelos robustos.

Diapositiva 4

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Ataques adversariales)

Tipos de ataques adversariales: perspectiva del conocimiento

- Caja blanca: el atacante tiene acceso completo a la arquitectura, los pesos y los gradientes del modelo.
- Caja negra: sin conocimiento interno del modelo; el atacante consulta el modelo para obtener salidas.
- Caja gris: conocimiento parcial (datos de entrenamiento, pero no parámetros).
- Los distintos niveles de conocimiento determinan las estrategias de ataque.
- La caja blanca permite ataques más precisos.
- La caja negra suele recurrir a aproximaciones basadas en consultas.

Diapositiva 5

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Ataques adversariales)

Tipos de ataques adversariales: objetivos

Co-funded by the European Union

- Ataques de evasión: modifican las entradas durante la inferencia para evitar la detección.
- Ataques de envenenamiento: inyectan datos maliciosos en el entrenamiento para corromper el modelo.
- Inversión/extracción de modelos: inferir o replicar datos confidenciales o el propio modelo.
- La evasión es lo más práctico en la detección de malware e intrusiones.
- El envenenamiento amenaza la integridad del modelo durante el reentrenamiento.
- La extracción pone en riesgo la propiedad intelectual y la privacidad.

Diapositiva 6

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Ataques adversariales)

Técnicas para generar ejemplos adversariales

- FGSM (Método del signo del gradiente rápido): utiliza el gradiente de la pérdida para perturbar la entrada.
- PGD (Descenso de gradiente proyectado): método iterativo más potente.
- C&W y DeepFool: basados en optimización, minimizan la magnitud de la perturbación mientras provocan la clasificación errónea.
- Una mayor sofisticación aumenta el éxito del ataque.
- Las técnicas explotan la diferenciabilidad del modelo.
- La defensa requiere contrarrestar estos métodos de ataque.

Diapositiva 7

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Ataques adversariales)

Implicaciones para la ciberseguridad

- Los clasificadores de malware son vulnerables a binarios modificados.
- La detección de intrusiones en red puede pasar por alto tráfico alterado sutilmente.
- Los sistemas biométricos pueden ser engañados con entradas adversariales.
- Los sistemas de reconocimiento facial son susceptibles a imágenes adversariales.
- Amenaza la fiabilidad de las herramientas de seguridad basadas en IA.
- Exige diseños de ML robustos frente a ataques adversariales.

Diapositiva 8

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Detección de anomalías)

Detección de anomalías y análisis estadístico de valores atípicos

- Fundamental para identificar ataques novedosos o sigilosos.

Co-funded by the European Union

- El comportamiento malicioso se desvía estadísticamente de los patrones normales.
- Detecta ataques que no figuran en los datos de entrenamiento etiquetados.
- Ofrece una defensa dinámica y adaptable.
- Ayuda a detectar intentos de evasión o envenenamiento.
- Mejora la calidad de las alertas en los sistemas de ciberseguridad.

Diapositiva 9

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Detección de anomalías)

Tipos de detección de anomalías

- Anomalías puntuales: datos individuales poco frecuentes (p. ej., una transferencia de datos de gran volumen).
- Anomalías contextuales: anómalas solo en contextos específicos (p. ej., la hora de inicio de sesión de un administrador).
- Anomalías colectivas: grupo de puntos con anormalidad colectiva (p. ej., un escaneo lento de puertos).
- Detecta comportamientos de amenaza diversos.
- Admite el análisis multidimensional.
- Las estrategias de detección a medida mejoran la eficacia.

Diapositiva 10

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Detección de anomalías)

Técnicas estadísticas para la detección de anomalías

- Análisis de puntuación Z (Z-score): marca los datos que superan los umbrales de desviación estándar.
- Las estadísticas robustas (desviación absoluta mediana, MAD) resisten los efectos de datos sesgados.
- Los métodos basados en densidad (Local Outlier Factor, LOF) identifican puntos de datos dispersos.
- Adecuadas para espacios de datos de ciberseguridad de alta dimensionalidad.
- Aumentan la resiliencia frente al ruido y los ataques.
- Permiten una clasificación eficaz de anomalías.

Diapositiva 11

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Detección de anomalías)

Papel en el aprendizaje automático adversarial

Co-funded by the European Union

- Detectan ejemplos adversariales que se desvían estadísticamente de los datos naturales.
- Monitorizan patrones inusuales de acceso o consulta que indiquen ataques de extracción.
- Proporcionan una capa adicional de defensa en la seguridad del ML.
- Complementan los métodos de robustez basados en el modelo.
- Mejoran la detección de ataques sutiles y novedosos.
- Apoyan el conocimiento continuo de la situación de seguridad.

Diapositiva 12

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Medidas de divergencia estadística)

Medidas de divergencia estadística en la detección adversarial

- Cuantifican las diferencias entre distribuciones de probabilidad (normal frente a anómala).
- Detectan cambios distribucionales causados por actividad adversarial o envenenamiento de datos.
- Evalúan la eficacia de los modelos generativos y de los detectores de anomalías.
- Apoyan la monitorización de desviaciones en el comportamiento de los usuarios.
- Esenciales para una monitorización de ciberseguridad robusta.
- Mejoran la fidelidad de la detección de amenazas en tiempo real.

Diapositiva 13

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Medidas de divergencia estadística)

Medidas de divergencia comunes

- Divergencia de Kullback–Leibler (KL): medida asimétrica, sensible a las colas.
- Divergencia de Jensen–Shannon (JS): versión simétrica, más suave y acotada.
- Distancia de variación total y distancia de Wasserstein: métricas prácticas adicionales.
- La elección depende de la tarea de detección y de las propiedades de los datos.
- Se utilizan para cuantificar cambios distribucionales anómalos.
- Mejoran la robustez adversarial en ML.

Diapositiva 14

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Medidas de divergencia estadística)

Aplicaciones en ciberseguridad

- Miden cambios en el comportamiento de los usuarios que indiquen amenazas internas.
- Comparan los perfiles de tráfico de red con el comportamiento normal esperado.

Co-funded by the European Union

- Evalúan la robustez y la fiabilidad de los clasificadores de amenazas.
- Detectan cambios sospechosos que señalan ataques.
- Impulsan actualizaciones defensivas adaptativas.
- Mejoran la resiliencia del sistema frente a la manipulación adversarial.

Diapositiva 15

Ataques adversariales | Detección de anomalías | Medidas de divergencia estadística (sección activa: Medidas de divergencia estadística)

Conclusión

- El ML adversarial expone vulnerabilidades de seguridad críticas.
- Comprender y detectar los ataques (evasión, envenenamiento, extracción de modelos) es esencial.
- La detección estadística de anomalías y las medidas de divergencia ofrecen capas de defensa robustas.
- La evaluación y el fortalecimiento continuos de los modelos de ML mejoran la resiliencia en ciberseguridad.
- La investigación y la defensa deben evolucionar al ritmo de la sofisticación de los adversarios.
- Construir sistemas de ML seguros es clave para mitigar las amenazas futuras.

Co-funded by the European Union