Siber Güvenlikte Yapay Zekâ

Yapay Zekâ ve Veri Dersleri Rehberi · Kısım VI: Sistem ve Uygulama

Saldırı tespiti | Zararlı yazılım | Dengesiz veri | Düşmanca saldırılar | YZ'nin güvenliği | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü:
https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_38_Siber_Guvenlikte_Yapay_Zeka.html

GİRİŞ

Ders Hangi Soruyu Sorar?

> **Yapay zekâ saldırıları nasıl yakalar ve kendisi nasıl saldırıya uğrar?**

Siber güvenlikte yapay zekâ iki soruyu birlikte sorar: makine öğrenmesi saldırıları, zararlı yazılımları ve oltalamayı nasıl yakalar; YZ sistemlerinin kendisi nasıl kandırılır? Alanı ayıran şey **düşmanca ortamdır**: karşıda modeli atlatmaya çalışan bir saldırgan vardır.

Bir bakışta

YAKLAŞIM güvenlik bilgisi ve MÖ	**GİRDİ** ağ trafiği, kayıtlar, dosyalar	**ÇIKTI** uyarı, sınıf, risk puanı
BAŞARI ÖLÇÜTÜ tespit oranı, yanlış alarm oranı	**MATEMATİK YÜKÜ** orta	**VERİ İHTİYACI** dengesiz ve sürekli değişen

Kısa Tarihçe

1987
Denning, saldırı tespitini olağandışı davranışı yakalayan bir model olarak tanımlar.

1999
KDD Cup 1999 saldırı tespiti veri seti yayımlanır; uzun yıllar ölçüt olarak kullanılır.

2002
Paul Graham, "A Plan for Spam" ile Bayesçi istenmeyen e-posta süzgecini yaygınlaştırır.

2014
Szegedy ve ark. düşmanca örnekleri gösterir: gözle görülmeyen değişiklik sınıflandırıcıyı yanıltır.

2016
DARPA Cyber Grand Challenge: makineler insan müdahalesi olmadan açık bulup yamalar.

2023
OWASP büyük dil modeli uygulamaları için ilk Top 10 listesini yayımlar; istem enjeksiyonu başta gelir.

DERS

Temel Kavramlar

Saldırı tespiti
İmza tabanlı ve anormallik tabanlı yaklaşımlar.

Zararlı yazılım
Dosyayı statik ya da davranışsal özniteliklerle sınıflama.

Oltalama tespiti
Sahte e-posta, alan adı ve web sayfalarını ayırmak.

Dengesiz veri
Saldırılar normal trafiğin yanında çok nadirdir.

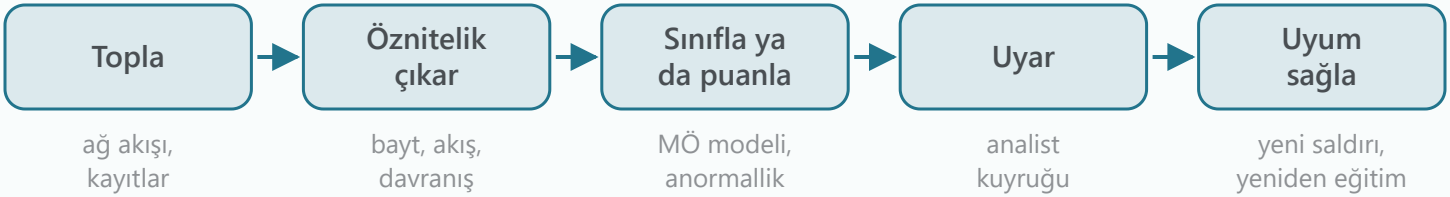
Yanlış alarm
Kesinlik, duyarlılık; analistin yorulması.

Düşmanca örnekler
Modeli yanıltmak için özenle bozulmuş girdiler.

Veri zehirleme
Eğitim verisine kötü niyetli örnek sokmak.

İstem enjeksiyonu
Dil modeline gizli talimatla amacını değiştirtmek.

Tipik İş Akışı



Örnek: Bir kurumun ağ geçidi bağlantı kayıtlarını toplar; her bağlantı için süre, bayt sayısı ve port gibi öznitelikler çıkarılır. Bir model şüpheli akışları puanlar ve yalnız en yüksek puanlılar analiste gider. Analistin kararları yeni etiketlere dönüşür ve model güncellenir.

DERS

Derste Neler İşlenir?

1. Tehdit modeli ve güvenlikte MÖ'nün sınırları
2. Ağ trafiği ve kayıtlardan öznitelik çıkarma
3. İmza ve anormallik tabanlı saldırı tespiti
4. Dengesiz veri, yanlış alarm ve uygun ölçütler
5. Zararlı yazılım: statik ve davranışsal analiz
6. Oltalama, spam ve kötü niyetli URL tespiti
7. Sahtekârlık ve hesap ele geçirme tespiti
8. Kavram kayması ve saldırganın uyum sağlaması
9. Düşmanca örnekler ve kaçınma saldırıları
10. Veri zehirleme ve arka kapılar
11. Model çalma, üyelik çıkarımı ve mahremiyet
12. Dil modeli güvenliği: istem enjeksiyonu

Kazandırdığı beceriler

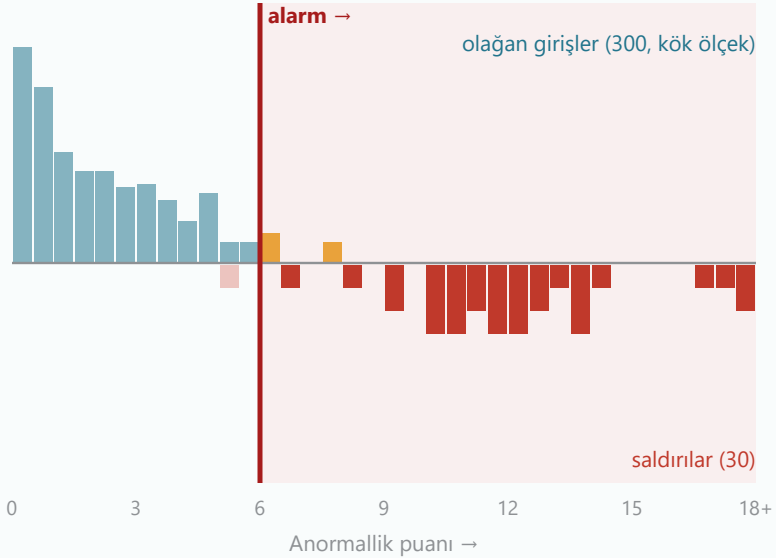
Wireshark ve Zeek ile trafik okumak · scikit-learn ile saldırı sınıflandırıcı · YARA kuralı yazmak · dengesiz veride doğru ölçütü seçmek · ART ile düşmanca saldırı denemek

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

Şüpheli Oturum Açmaları Yakala

Şirketin oturum açma kayıtlarında olağan girişlerin arasına parola deneme saldırıları karışmış. Model her girişe bir anormallik puanı verir. Eşiği kaydırın ya da grafiğe tıklayın; saldırgan uyum sağlayınca ne olur?



Alarm eşiği = **6,0**

Saldırgan uyum sağlar

Yakalanan saldırı: **29 / 30** (%97)
Yanlış alarm: **3** olağan giriş (%1,0)
Eşik iyi bir dengede.

Puan, girişin olağan davranıştan ne kadar şaşırtıcı olduğudur: başarısız deneme sayısı, saat, yeni cihaz, yabancı ülke. Model yalnız olağan girişlerden öğrendi.

Siber güvenlikte yapay zekâ **olağanı öğrenip sapmayı** yakalar; imzası bilinmeyen saldırıyı da görür. Ama karşıda öğrenen bir rakip var: model ve özellikler sürekli güncellenmeli.

KONUM

Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

Ders	Neden gerekir
Makine Öğrenmesi	saldırı tespiti, kötü amaçlı yazılım sınıflandırma
İşletim Sistemleri	süreçler, dosya sistemi ve izinler: saldırıların hedef aldığı katman
Bilgisayar Ağları	TCP/IP, port ve paketler: ağ trafiğini okuyabilmek
C ile Programlama	bellek ve işaretçiler: taşma açıklarının nasıl oluştuğunu görmek

KONUM

Bu Ders Sana Göre mi?

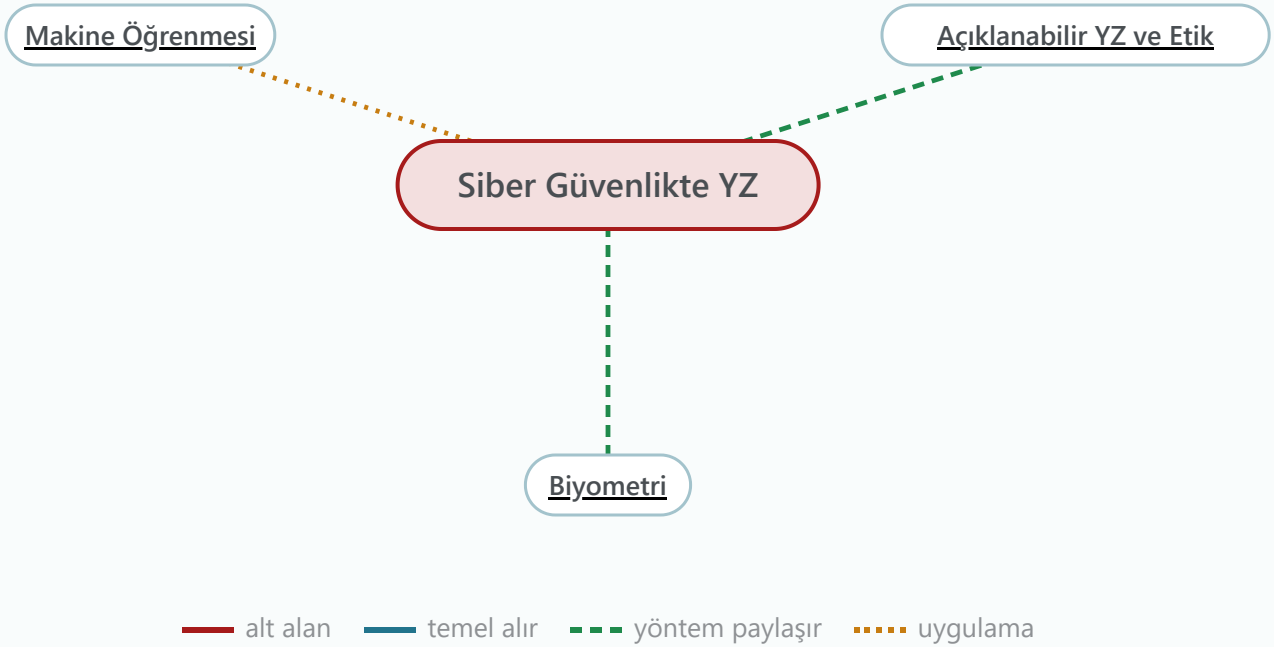
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

Kendine sor	Hayırsa önce
Dengesiz bir veri setinde kesinlik ve duyarlılığı yorumlayabiliyor musun?	Makine Öğrenmesi
Bir TCP bağlantısının nasıl kurulduğunu ve portun ne olduğunu açıklayabiliyor musun?	Bilgisayar ağları dersindeki TCP/IP konuları
Linux'ta süreçleri, dosya izinlerini ve sistem kayıtlarını inceleyebiliyor musun?	İşletim sistemleri dersi ve Linux komut satırı
Bir arabellek taşmasının bellekte nasıl oluştuğunu açıklayabiliyor musun?	C ile Programlama dersindeki işaretçi ve diziler

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

Dersin Komşuları



ÖRTÜŞME

Örtüştüğü Dersler

İlişki	Ders	Ortak konu
uygulama	Bu ders Makine Öğrenmesi yöntemlerini uygular	saldırı tespiti, kötü amaçlı yazılım sınıflandırma
yöntem paylaşır	Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği ile yöntem paylaşır	güvenilirlik, düşmanca örnekler
yöntem paylaşır	Biyometrik Sistemler ile yöntem paylaşır	kimlik doğrulama, sahtecilik tespiti

Ayrıştığı Noktalar (1/2)

Ölçüt	Siber Güvenlikte YZ	Makine Öğrenmesi	Açıklanabilir YZ ve Etik
Temel soru	Yapay zekâ saldırıları nasıl yakalar ve kendisi nasıl saldırıya uğrar?	Bir program örneklerden nasıl genelleme yapar?	Bir yapay zekâ sisteminin kararına nasıl güvenilir ve sonucundan kim sorumludur?
Yaklaşım	güvenlik bilgisi ve MÖ	istatistiksel öğrenme	insan merkezli değerlendirme
Girdi	ağ trafiği, kayıtlar, dosyalar	etiketli ya da etiketsiz örnekler	model, verisi ve kullanım bağlamı
Çıktı	uyarı, sınıf, risk puanı	tahmin yapan model	açıklama, denetim raporu, ilke
Başarı ölçütü	tespit oranı, yanlış alarm oranı	görülmemiş veride hata	anlaşılırlık, adalet, hesap verebilirlik
Matematik yükü	orta	yüksek: doğrusal cebir, olasılık, optimizasyon	düşük-orta
Veri ihtiyacı	dengesiz ve sürekli değişen	orta ve çok	modelin kendi verisi

- **Makine Öğrenmesi:** MÖ veri dağılımının sabit kaldığını varsayar; bu derste karşı taraf modeli atlatmak için girdiyi bilerek değiştirir.
- **Açıklanabilir YZ ve Etik:** Açıklanabilir YZ güvenilirliği genel olarak ele alır; bu ders saldırgana karşı dayanıklılığa ve saldırıyı yakalamaya odaklanır.

Ayrıştığı Noktalar (2/2)

Ölçüt	Siber Güvenlikte YZ	Biyometri	Veri Madenciliği
Temel soru	Yapay zekâ saldırıları nasıl yakalar ve kendisi nasıl saldırıya uğrar?	Bir kişi yüzünden, parmak izinden ya da sesinden nasıl tanınır?	Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur?
Yaklaşım	güvenlik bilgisi ve MÖ	eşleme ve doğrulama	keşfe dayalı çözümleme
Girdi	ağ trafiği, kayıtlar, dosyalar	yüz, parmak izi, iris, ses	büyük tablo ya da işlem verisi
Çıktı	uyarı, sınıf, risk puanı	kimlik kararı	kural, küme, aykırı değer
Başarı ölçütü	tespit oranı, yanlış alarm oranı	FAR, FRR, EER	ilginçlik, destek, güven
Matematik yükü	orta	orta	orta
Veri ihtiyacı	dengesiz ve sürekli değişen	kişi başına az örnek	çok

- **Biyometri:** Biyometri kişiyi tanıyarak kimlik doğrular; bu ders ağdan dosyaya, e-postadan modele kadar geniş saldırı yüzeyine bakar.
- **Veri Madenciliği:** Veri madenciliği olağandışı örüntüleri keşif için arar; bu ders onları saldırı olarak yorumlar ve alarma dönüştürür.

UYGULAMA

Uygulama Alanları

Saldırı tespiti
Ağ trafiğinde olağandışı bağlantıları bulmak.

Zararlı yazılım
Dosyayı çalıştırmadan zararlı mı diye kestirmek.

Oltalama ve spam
Sahte e-posta ve web sayfalarını ayıklamak.

Sahtekârlık
Şüpheli işlem ve hesap ele geçirme tespiti.

Güvenlik operasyonu
Binlerce uyarıyı önceliklendirip sunmak.

YZ'yi korumak
Düşmanca örnek, zehirleme, istem enjeksiyonu.

UYGULAMA

Riskler ve Etik

Çift kullanım
Tespit için öğrenilen teknikler, tespiti atlatan zararlı yazılım ya da ikna edici oltalama e-postası üretmekte de kullanılır.

İzinsiz test
Başkasının ağını ya da sistemini izinsiz taramak suçtur; deneyler yalnız yalıtılmış laboratuvarda ve yazılı izinle yapılır.

Çalışanların gözetlenmesi
İç tehdit tespiti çalışanların e-posta ve davranışlarını izler; orantılılık ve KVKK sınırları gözetilmelidir.

Otomatik müdahale
Modelin kararıyla bağlantıyı ya da hesabı kendiliğinden kesmek, yanlış alarmda meşru kullanıcıları ve işi durdurur.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

Araçlar ve Örnek Veriler

Araçlar ve kütüphaneler

scikit-learn · Zeek · Suricata · Wireshark · YARA · Elastic Stack · Adversarial Robustness Toolbox

Tanınmış veri setleri

KDD Cup 1999 / NSL-KDD · CICIDS2017 · UNSW-NB15 · EMBER (zararlı yazılım) · Enron-Spam · PhishTank

Eski veri setleri (KDD Cup 1999) hâlâ çok kullanılır ama bugünün trafiğini yansıtmaz; sonuçlar güncel veri setlerinde de sınanmalıdır.

UYGULAMA

Sık Karıştırılanlar

Yaygın düşünce	Doğrusu
%99 doğruluk iyi bir saldırı tespitidir.	Saldırılar çok nadirse her şeye normal diyen model de %99 doğrudur; kesinlik ve yanlış alarm oranı bakılır.
Model bir kez eğitilince saldırıları yakalar.	Saldırganlar yöntem değiştirir; model yeni örneklerle sürekli güncellenmelidir.
YZ güvenlik analistinin yerini alır.	YZ uyarıları süzer ve önceliklendirir; karar ve soruşturma hâlâ insandadır.
Derin ağlar saldırıya karşı dayanıklıdır.	Göze görünmeyen küçük değişiklikler (düşmanca örnekler) modelin kararını kolayca değiştirebilir.

SONRASI

Bu Dersten Sonra

Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

İlgili meslekler

Güvenlik analisti (SOC) | Tehdit avcısı | Zararlı yazılım analisti | Güvenlik veri bilimcisi | YZ güvenliği uzmanı

Bu ders iki yöne açılır: güvenlik tarafında ağ güvenliği, adli bilişim ve zararlı yazılım analizi; yapay zekâ tarafında düşmanca dayanıklılık, gizliliği koruyan öğrenme ve büyük dil modellerinin güvenliği. Açıklanabilir YZ ve etik dersi bu ikinci yolun doğal komşusudur.

SONRASI

Kaynaklar

- Chio ve Freeman, *Machine Learning and Security*, O'Reilly, 2018.
- Denning, "An Intrusion-Detection Model", *IEEE Transactions on Software Engineering* 13(2), 1987.
- Sommer ve Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection", *IEEE Symposium on Security and Privacy*, 2010.
- Goodfellow, Shlens ve Szegedy, "Explaining and Harnessing Adversarial Examples", *ICLR*, 2015.

Son

Bölüm 38: Siber Güvenlikte Yapay Zekâ

← Önceki bölüm · Alanın haritası · Sonraki bölüm →