

Isolation Forest ile sensör anomalisi: nasıl çalışır, nerede yanılır

Sensör tablosunda işaretlenen satırları izolasyon puanına göre okumayı, contamination değerinin ne anlama geldiğini ve yöntemin nerede yanıldığını öğrenin.

Kısaca

- Isolation Forest, az ve farklı noktaları rastgele bölmelerle daha az adımda ayırır ve mesafe hesabı gerektirmez.
- Puan 1'e yaklaştıkça nokta daha kolay izole edilmiş demektir; 0,5 civarındaki değerler belirgin bir ayırım olmadığını gösterir.
- Contamination bir varsayımdır: kesme noktasını yerleştirir ama sistemdeki gerçek arıza oranını ölçmez ve bu nedenle ölçülmüş bir değer gibi raporlanmamalıdır.

01 / Anomali neden kolay ayrılır

Isolation Forest (izolasyon ormanı), anomali tespiti için kullanılan bir topluluk yöntemidir. Temel fikir basittir: anomaliler azdır ve normal noktalardan farklıdır. Her izolasyon ağacı, rastgele seçilen bir sensörde rastgele bir kesme değeri kullanarak veriyi böler. Bu bölme, bir nokta kendi bölgesinde tek başına kalana kadar tekrarlanır. Bu noktaya kadar yapılan bölme sayısı, o noktanın yol uzunluğudur.

Normal bir nokta yoğun bir bölgededir ve ayrılması çok sayıda bölme gerektirir. Anormal bir nokta ise birkaç bölmede ayrılır. Yöntem mesafe ya da yoğunluk hesaplamaz. Her bölme yalnızca tek bir değişkenin aralığını kullandığı için birimlerdeki ölçek farkları bölmeleri değiştirmez; bu nedenle ölçekleme genellikle gerekmez. Bu sadelik bir doğruluk garantisi değildir. Yöntem bir noktanın ne kadar kolay ayrıldığını sıralar, ama nedenini kendisi açıklamaz.

02 / Yol uzunluğu ve puanın okunuşu

Model birçok ağaç kurar. Her ağaç, veriden rastgele seçilmiş küçük bir alt örnekle eğitilir; 256 noktalık bir alt örnek sık kullanılan bir varsayılandır, ancak kendi verinizle doğrulanmalıdır. Bir x noktası için tüm ağaçlardaki yol uzunluklarının ortalaması $E[h(x)]$ ile gösterilir. Anomali puanı şöyle tanımlanır:

$$s(x) = 2^{-(E[h(x)] / c(n))}$$

Burada $c(n)$, n noktalık bir veri kümesinde başarısız bir ikili arama ağacı araması için ortalama yol uzunluğudur. Puan 0 ile 1 arasında bir değer alır; 1'e yaklaştıkça nokta daha erken izole edilmiş demektir.

Hipotetik bir sayısal kontrol: noktanın ortalama yol uzunluğu $c(n)$ ise puan $2^{-1} = 0,5$ olur ve bu belirgin bir ayırım olmadığı anlamına gelir. Yol uzunluğu $c(n)$ 'nin yarısıysa puan $2^{-0,5} \approx 0,71$ olur; nokta tipik bir noktadan belirgin biçimde daha erken ayrılmıştır.

03 / Örnek: conta değişiminden sonra bir pompa

Aşağıdaki sayılar hipotetiktir; gerçek bir tesisten alınmamıştır. Beş sensörlü bir pompa tablosu düşünün: titreşim, yatak sıcaklığı, akış, basınç ve motor akımı. Model 4 000 normal satırla eğitilmiştir. Conta değişiminden sonra gelen 200 satırın 6'sı işaretlenmiş. Oran $6 / 200 = 0,03$, yani yüzde 3'tür. Eğitimde contamination 0,01 seçildiyse kesme noktası, eğitim satırlarının yaklaşık yüzde birinin, yani $4\ 000 \times 0,01 = 40$ satırın dışında kalacak şekilde yerleşir. Yeni verideki yüzde 3 oranı bu tasarım değerinin üç katıdır.

Operatörler işaretli altı satırı inceledi. Dördü, conta işinden sonra pompanın davranışında gerçek bir değişiklik olarak doğrulandı. İki pompa kalkışına denk gelen geçici satırlardı. Model gerçek bir işletme değişikliğini yakalamış görünüyor, ama her işaretin arıza olmadığı da açıktır. Dört gerçek satırda sensör açıklamalarının çoğunlukla akış ve motor akımını gösterdiğini varsayalım; ekip önce bu iki sensöre bakar.

04 / Ağaç sayısı, alt örnek ve contamination

Sonucu üç ayar şekillendirir. Ağaç sayısı arttıkça ortalama alınan puanlar daha kararlı hale gelir, ama hesap süresi de artar. Alt örnek boyutu her ağacı küçük ve birbirinden farklı tutar. 256 bir başlangıç değeridir, evrensel bir kural değildir.

Contamination kesme noktasının yerini belirler. Bu değer tesisteki gerçek arıza oranını ölçmez; ekibin kabul edebileceği bir tasarım seçimidir. Onu, bakım ekibinin bir vardiyada gerçekten işleyebileceği uyarı sayısı ile birlikte seçin. Ayrıca bu değer değiştikçe işaretlenen satır sayısının nasıl değiştiğini kaydedin. Tek bir eşik değeri, sonucun ne kadar hassas olduğunu gizler. Aynı eşiği tüm vardiyalarda ve tüm ürünlerde kullanmak farklı çalışma koşullarında farklı sonuçlar verebilir; bu yüzden sayıları vardiya ve ürün bazında da ayırın.

05 / Yöntemin sınırları

Isolation Forest satırları birbirinden bağımsız kabul eder. Zaman sırası ve mevsimsellik modele girmez. Bu yüzden her değeri tek tek normal görünen, ama haftalar boyunca yavaşça kayan bir sensör fark edilmeyebilir. Yavaş kaymalar için CUSUM ve EWMA gibi zamana duyarlı araçlar daha uygundur.

Her değişkeni tek tek normal olan, ama birlikte beklenmeyen bir kombinasyon da yöntem için zordur. Bu durumda çok değişkenli süreç izleme, örneğin Hotelling T^2 ve SPE, daha doğrudan bir yaklaşım sunar. Çalışma rejimi değiştiğinde, örneğin yeni bir ürün reçetesinde ya da farklı bir devirde, normal tanımı da değişir ve model yeniden eğitilmelidir. Eğitim verisine bir arıza dönemi karışmışsa model o arızayı normal davranış olarak öğrenebilir.

06 / Sahada nasıl başlanır

İlk adım, bilinen normal çalışmanın yaşandığı bir eğitim penceresi seçmektir. Bakım müdahalesi olmayan ve arıza kaydı bulunmayan, kararlı koşullardaki bir dönem uygundur. Modeli bu pencerede eğitin, sonra eğitimde hiç kullanılmamış daha geç bir dönemde test edin. Her işaretli satırı, makineyi tanıyan operatörlerle birlikte gözden geçirin ve nedenini kaydedin.

Haftalık yanlış alarm sayısını ayrıca sayın. Teknik olarak iyi görünen ama her hafta yığınla uyarı üreten bir model sahada kullanılmaz. Bu sayı ekibin kaldıramayacağı düzeydeyse eşiği ya da sensör setini yeniden düşünün. Son olarak her uyarıyı adlandırılmış sensörlerle açıklayıp açıklayamadığınızı kontrol edin. Açıklanamayan bir uyarı, kısa sürede güveni azaltır. Pilot boyunca eşiği sık sık değiştirmek yerine sabit tutun; değişiklik gerekiyorsa tarihini ve gerekçesini kaydedin.

Uygulayın

Sentinel’de örnek sensör tablosunu açın, işaretli bir satırı seçin ve açıklama katkılarında en çok öne çıkan iki sensörü bir cümleyle yazın.

Kaynaklar ve çevrimiçi rehber

Güncel çevrimiçi rehberi açın

scikit-learn: IsolationForest

Isolation forest (Wikipedia)

Isolation Forest (Liu, Ting, Zhou, ICDM 2008)

scikit-learn: novelty and outlier detection

Önerilen pilot adımları saha doğruluk iddiası içermez. Sentetik örnekler gerçek tesis ölçümü değildir.