19 Aralık 2023 (24 Aralık 2024'te güncellendi)

Açıklanabilir Yapay Zeka

10. Hafta: Yerel düzeyde açıklayıcılar | Karşı-olgusal açıklamalar

Mustafa Cavus, Ph.D.
🏢 Eskişehir Teknik Üniversitesi - İstatistik Bölümü
📧 mustafacavus@eskisehir.edu.tr
🌐 linktr.ee/mustafacavus

Giriş

“Kahveyi sıcakken içmeseydim, dilim yanmazdı.”

Giriş

A

“Kahveyi sıcakken içmeseydim, dilim yanmazdı.”

B

Örnek Olay

Kredi başvurusu yapan ve bankacılık karar destek sistemi tarafından başvurusu reddedilen bir müşterinin durumunu ele alalım. Doğal olarak müşteri, başvurusunun neden reddedildiğini ve kredi alma şansını nasıl yükseltebileceğini merak eder.

"Neden" sorusu, karşı-olgusal bir ifade olarak formüle edilebilir: Müşterinin değişkenlerde (gelir, kredi kartı sayısı, yaş, vb.) yapılacak hangi değişiklik, başvurunun sistem tarafından kabul etmesini sağlar?

- Müşterinin yılda 100.000 TL daha fazla kazanması olabilir.
- Müşterinin daha az sayıda kredi kartı kullanıyor olması ve beş yıl önce aldığı krediyi zamanında ödemiş olması olabilirdi.

Örnek Olay

Bir ev sahibinin dairesini kiralamak istediğini, ancak ne kadar ücret talep etmesi gerektiğinden emin olmadığı durumu ele alalım:

Bu konuda ev sahibi, emlakçısının kullandığı karar destek sisteminden yardım alabilir. Sistem dairenin büyüklük, konum, evcil hayvan kabul edilip edilmediği gibi tüm detayları girdikten sonra, 9.000 TL talep etmesini önerir. Ev sahibi beklentisi bu tutarın üzerinde ise dairenin kira bedelini nasıl artırabileceğini merak edebilir.

- Dairenin 15 m2 daha büyük olması durumunda dairenin 10.000 TL'den fazla kiralanabileceği bilgisine ulaşabilir. Ancak böyle bir öneri hayata geçirilemez, çünkü daireyi genişletemez.
- Sonunda, yalnızca kontrolü altındaki değişkenlerin değerlerini değiştirerek (ankastre mutfak eklemek, evcil hayvan kabul etmek ya da zemini yenilemek gibi), evcil hayvanlara izin verirse ve daha iyi izolasyona sahip pencereler takarsa, 10.000 TL talep edebilir.

Karşı-olgusal açıklamalar

Giriş

Bir tahmin değerinin karşı-olgusal açıklaması, değişken değerlerinde yapılacak en küçük değişiklik ile model tahminin istenen değere ulaşmasını sağlar.

Olgusal durum: $Y = f(X)$

Karşı-olgusal durum: $Y' = f(X')$

Karşı-olgusal açıklamalar ve Rashomon etkisi

Rashomon etkisi

Karşı-olgusal açıklamalar, mevcut duruma karşı zıtlık oluşturdukları ve genellikle birkaç değişken değerinin değişikliğine odaklandıkları için kullanıcı dostu açıklamalardır.

Ancak karşı-olgusal açıklamalar, '**Rashomon Etkisi**'nden olumsuz etkilenirler. Rashomon, bir Samuray'ın cinayetinin farklı kişiler tarafından nasıl anlatıldığı konu alan Japon yapımı bir filmdir. Hikayelerin tümü aynı sonucu anlatır, ancak hikayeler birbirinden farklıdır.

Karşı-olgusal açıklamalarda da benzer durumla karşılaşılabilir çünkü genellikle birden fazla farklı karşı-olgusal açıklama bulunur. Her bir karşı-olgusal açıklama, belirli bir sonuca nasıl ulaşıldığına dair farklı bir "hikaye" anlatır.

Bir karşı-olgusal açıklama A değişkenin aldığı değeri değiştirmeyi söylerken, diğeri A değişkeninin aynı kalması, ancak B değişkeninin aldığı değerin değişmesini önerebilir, ki bu bir çelişkidir. Bu sorunu, ya tüm karşı-olgusal açıklamaları raporlayarak ya da karşı-olgusal açıklamaları değerlendirmek ve en iyisini seçmek için bir kriter kullanarak ele alınabilir.

İyi bir karşı-olgusal açıklamanın sahip olması gereken özellikler

- Bir karşı-olgusal açıklama, önceden belirlenmiş tahminle mümkün olduğunca yakın olmalıdır.

- Karşı-olgusal, gözlenen değişken değerlerine mümkün olduğunca benzer olmalı ve az sayıda değişken değerini değiştirmelidir. Örneğin, Öklid uzaklığı veya L0 normu kullanarak bu benzerliği ölçebiliriz.

- Birden fazla çeşitli karşı-olgusal açıklama sunulmalıdır. Bu, farklı sonuçları elde etmek isteyen bir kişinin çeşitli seçeneklere erişimini sağlar. Çeşitlilik, karar alanın tercih ettiği değişkenlerin aldığı değerleri değiştirmesine olanak tanır.

- Karşı-olgusal bir açıklama, gerçekçi değişken değerlerini almalıdır. Örneğin, daire boyutunun negatif veya gerçekçi olmayan bir değer alması mantıklı değildir. Olası değişken değerleri, veri setinin ortak dağılımına uygun olmalıdır.

Karşı-olgusal açıklamalar oluşturma yöntemleri

Karşı-olgusal açıklamaların oluşturulması

Karşı-olgusal açıklamaları oluşturmak için basit bir yaklaşım, deneme yanılma yöntemiyle arama yapmaktır. Bu yaklaşım, ilgilenilen gözlem değerinin değişken değerlerini rastgele değiştirme ve istenen çıktı tahmin edildiğinde durma işlemi içerir.

Örneğin, bir ev sahibinin dairesi için daha yüksek kira talep edebileceği bir gözlem setini bulmaya çalıştığı durumda olduğu gibi. Ancak deneme yanılma yönteminden daha iyi, kayıp fonksiyonuna dayalı yaklaşımlar bulunmaktadır.

Bu yöntemlerde kayıp fonksiyonu, gözlem ve istenen sonucu girdi olarak kullanır. Ardından, bu kaybı en aza indiren karşı-olgusal açıklamayı bir optimizasyon algoritması kullanarak bulabiliriz. Birçok yöntem bu şekilde ilerler, ancak kayıp fonksiyonu tanımı ve optimizasyon yöntemi açısından farklılık gösterir.

Watcher et al. (2017)

Watcher et al. aşağıdaki kayıp fonksiyonunu minimize etmeyi önermişlerdir:

$$L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')$$

İlk terim, karşı-olgusal x' için model tahmini ile kullanıcının önceden tanımlaması gereken istenen sonuç y' arasındaki ikinci dereceden mesafedir. İkinci terim, açıklanacak x gözlemi ile karşıolgusal x' arasındaki d mesafesidir. Kayıp, karşı-olgusalın tahmin edilen sonucunun önceden tanımlanmış sonuçtan ne kadar uzakta olduğunu ve karşı-olgusalın ilgilenilen durumdan ne kadar uzakta olduğunu ölçer. Mesafe fonksiyonu d, her bir özelliğin ters medyan mutlak sapması (MAD) ile ağırlıklandırılmış Manhattan mesafesi olarak tanımlanır.

$$d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j}$$

Watcher et al. (2017)

Toplam uzaklık, p değişken için uzaklıkların toplamıdır, yani x örneği ile karşı-olgusal x' arasındaki değişken değerlerinin mutlak farklarıdır. Değişken bazında uzaklıklar, j özelliğinin veri kümesi üzerindeki medyan mutlak sapmasının tersiyle ölçeklendirilir ve şu şekilde tanımlanır:

$$MAD_j = \text{median}_{i \in \{1,\ldots,n\}}(|x_{i,j} - \text{median}_{l \in \{1,\ldots,n\}}(x_{l,j})|)$$

MAD, bir özelliğin varyansının eşdeğeridir, ancak ortalamayı merkez olarak kullanmak ve uzaklıkların kareleri üzerinden toplamak yerine, ortancayı merkez olarak kullanırız ve mutlak uzaklıkların toplamını kullanırız. Önerilen uzaklık fonksiyonunun Öklid uzaklığına göre aykırı değerlere karşı daha dayanıklı olması avantajı vardır. Tüm değişkenleri aynı ölçeğe getirmek için MAD ile ölçeklendirme gereklidir; bir dairenin boyutunu metrekare veya metrekare olarak ölçmek önemli değildir.

Watcher et al. (2017)

λ parametresinin yüksek değerleri istenilen tahmin değerine yakın, küçük değerleri ise gözlem değerine yakın (benzeyen) karşı-olgusal açıklamalar elde edilmesini sağlar. Kullanıcının bu deneyi sağlaması için parametrenin değerini öncesinde belirlemesi gerekir. Bunu belirlemek yerine aşağıdaki kısıtı da kullanabilir:

$$|\hat{f}(x') - y'| \leq \epsilon$$

Nelder-Mead gibi optimizasyon algoritmalarıyla istenilen sonuca yakınsayana kadar kayıp fonksiyonu minimize edilir:

$$\arg\min_{x'} \max_{\lambda} L(x, x', y', \lambda).$$

Watcher et al. (2017)

Algoritma:

1. Açıklanacak gözlem değeri x, istenilen yanıt değeri y', tolerans değeri ϵ ve λ değerini belirleyin.
2. Başlangıç için rastgele bir karşı-olgusal seçiniz.
3. Başlangıç noktası olarak seçilen karşı olgusal ile kaybı optimize edin.

4. $|\hat{f}(x') - y'| > \epsilon$ koşulu sağlandığı sürece:
   λ değerini arttır
   Kayıp fonksiyonunu başlangıç noktası olarak seçilen karşı-olgusala göre optimize et.
   Kayıp fonksiyonunu minimize eden karşı-olgusalı döndür.

5. 2-4 arasındaki adımları tekrarlayın ve kayıp fonksiyonunu minimize eden karşı-olgusalların listesini döndürün.

Watcher et al. (2017)

Yöntemin bazı dezavantajları bulunmaktadır:

- Sadece ilk ve ikinci kriterleri dikkate alır, son iki kriteri ("sadece birkaç değişken değişikliği içeren ve olası değişken değerleri üreten karşı-olgusallar üretme") dikkate almaz.
- Gerçekçi olmayan değişkenler kombinasyonlarına ceza verilmez.
- Çok sayıda farklı seviyeye sahip kategorik değişkenler ile iyi çalışmaz.

Yöntemin yazarları, kategorik değişkenlerin değerlerinin her kombinasyonu için yöntemi ayrı ayrı çalıştırmayı önerdiler, ancak bu, çok sayıda kategorik değişken olması ve çok sayıda kategori olması durumunda kombinatoryel bir patlamaya yol açar. Örneğin: on farklı kategorisi olan bir kategorik değişken için bir milyon hesaplama adımına karşılık gelir.

Dandl et al. (2023)

Dandl et al., Watcher et al. yönteminden farklı olarak λ parametresini kullanmaz. Ayrıca Gower metriğini kullanarak kambinatoryal patlamanın önüne geçer.

Uygulama

Uygulama

German Credit Risk veri seti üzerinde bir SVM modelini ele alalım. Model, bir müşterinin kredi güven skorunu tahmin eder. Bunun için 522 gözlem ve 9 değişken kullanır.

Kredi güven skoru 0.242 olan bir müşteriyi ele alalım:

age	sex	job	housing	savings	amount	duration	purpose
58	f	unskilled	free	little	6143	48	car

Müşterinin kredi güven skorunun 0.50 üzerinde olması için hangi değişkenlerin nasıl değişmesi gerektiğini açıklayan karşı-olgusal açıklamaları oluşturmaya odaklanalım.

Uygulama

Aşağıdaki en iyi 10 karşı-olgusal açıklama elde edilmiştir:

age	sex	job	amount	duration	o_2	o_3	o_4	$\hat{f}(x')$
		skilled		-20	0.108	2	0.036	0.501
		skilled		-24	0.114	2	0.029	0.525
		skilled		-22	0.111	2	0.033	0.513
-6		skilled		-24	0.126	3	0.018	0.505
-3		skilled		-24	0.120	3	0.024	0.515
-1		skilled		-24	0.116	3	0.027	0.522
-3	m			-24	0.195	3	0.012	0.501
-6	m			-25	0.202	3	0.011	0.501
-30	m	skilled		-24	0.285	4	0.005	0.590
-4	m		-1254	-24	0.204	4	0.002	0.506

- Tüm açıklamalar, sürenin 48 aydan en az 23 aya indirilmesini öneriyor.
- Bazıları müşterinin vasıflı bir işte çalışıyor olmasını öneriyor.
- Hatta bazı açıklamalar, cinsiyetin kadından erkeğe değiştirilmesini bile öneriyor.
- Cinsiyet ile birlikte yaşta da bir değişim önerisi getiriliyor.
- Ancak bu iki öneri mantıklı (*plausible*) değillerdir.

Karşı-olgusal açıklamaların artı ve eksileri

Artıları

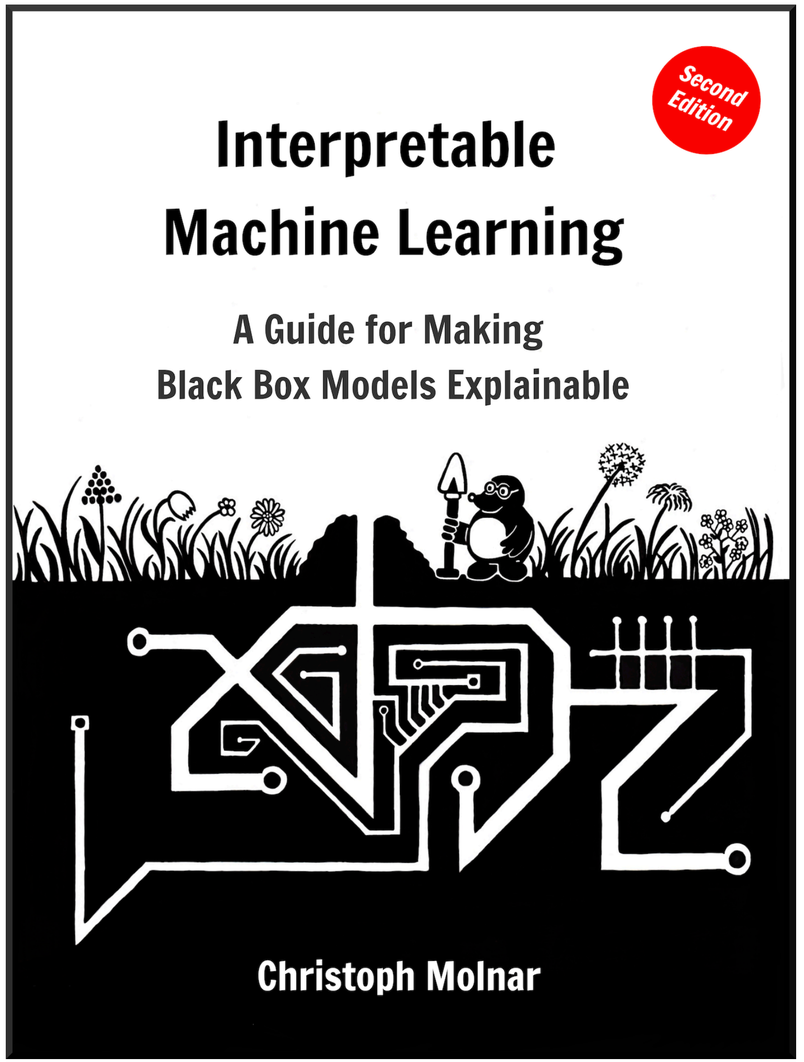
1. Karşı-olgusal açıklamaları **yorumlamak oldukça kolaydır**.
2. Hangi değişken değerinin değişmesi gerektiği üzerinde de yorumlanabildiği için **iki yönlü raporlama esnekliği sunar**.
3. **Veri veya modele erişim gerektirmez**. Bu durum ticari sırların saklanması ve verilerin korunmasına hizmet eder.
4. Yalnızca makine öğrenmesi sistemlerine değil **girdi-çıktı içeren herhangi bir sistemde kullanılabilir**.
5. Optimize edilebilen bir kayıp fonksiyonu üzerine kurulu bir yöntem ile oluşturulabildiği için **kolay bir şekilde uygulanabilir**.

Eksileri

1. Rashomon etkisi, çok sayıda karşı-olgusal açıklama oluşmasına neden olabilir. (Bu durum diğer taraftan seçim esnekliği sağladığı için bir artıdır.)

Kaynaklar

Bu materyalin hazırlanmasında **Interpretable Machine Learning (Molnar, 2023)** kitabından yararlanılmıştır. Kitabın ücretsiz online versiyonuna bağlantı üzerinden erişilebilir: **https://christophm.github.io/interpretable-ml-book**



Ders notlarına dersin **GitHub** reposu üzerinden ulaşabilirsiniz.

Ders ile ilgili sorularınız için mustafacavus@eskisehir.edu.tr adresi üzerinden benimle iletişime geçebilirsiniz.

Mustafa Cavus, Ph.D.
🏢 Eskişehir Teknik Üniversitesi - İstatistik Bölümü
📧 mustafacavus@eskisehir.edu.tr
🌐 linktr.ee/mustafacavus