

Bu bölümde, keşfetmek için takdire şayan bir algoritma sunuyoruz: Lojistik Regresyon. Gerçekten de Lojistik Regresyon, Sosyal Bilimler ve Doğa Bilimlerindeki en önemli analitik araçlardan biridir. Doğal dil işlemede, lojistik regresyon, sınıflandırma için temel denetimli makine öğrenme algoritmasıdır ve ayrıca sinir ağları ile çok yakın bir ilişkiye sahiptir.

Sınıflandırma Algoritması Nedir?

Sınıflandırma algoritmaları fikri oldukça basittir. Veri kümesini analiz ederek hedef sınıfı tahmin edersiniz. Bu, veri bilimi açısından en önemli konudur.

Sınıflandırma Algoritmalarında Temel Terminoloji

- **Sınıflandırıcı:** giriş verilerini belirli bir kategoriye ekleyen bir algoritma.
- **Sınıflandırma modeli:** bir sınıflandırma modeli, eğitim için verilen girdi değerlerinden bir miktar sonuç çıkarmaya çalışır. Yeni veriler için sınıf etiketlerini/kategorilerini tahmin edecektir.
- **Özellik:** gözlemlenen bir fenomenin bireysel ölçülebilir bir niteliğidir.
- **İkili sınıflandırma:** iki olası sonuç ile Sınıflandırma görevi.
Örneğin: Cinsiyet sınıflandırması (Erkek / Kadın)
- **Çok sınıflı sınıflandırma:** ikiden fazla sınıf ile sınıflandırma. Her örnek sadece bir hedef etikete atanır. Örneğin: bir hayvan bir kedi veya köpek olabilir, ikisi birden aynı anda olamaz.
- **Çoklu etiket sınıflandırması:** her bir numunenin bir dizi hedef etiketle (birden fazla sınıf) eşleştirildiği sınıflandırmadır. Örneğin: bir haber makalesi aynı anda spor, bir kişi ve konum hakkında olabilir.

Sınıflandırma algoritmalarının uygulamaları

- E-posta spam sınıflandırması
- Banka müşterilerinin kredi ödeme istekliliğini tahmin edilmesi.
- Kanseri tümör hücrelerinin tanımlanması.
- Duygu analizi
- İlaçların sınıflandırılması

Sınıflandırma algoritmalarının türleri

Sınıflandırma algoritmaları genel olarak aşağıdaki gibi sınıflandırılabilir:

- *Doğrusal Sınıflandırıcılar*
 - Lojistik regresyon
 - Naive Bayes Sınıflandırma Algoritması
 - Fisher's doğrusal diskriminantı
- *Destek vektör makinesi*
 - Least squares Support Vector Machines
- *Kuadratik sınıflandırıcılar*
- *Çekirdek tahmini*
 - K-NN Algoritması
- *Karar ağaçları*
 - Rassal orman
- *Yapay sinir ağları*
- *Öğrenmeli Vektör Kuantalama*

Lojistik Regresyon Nedir?

Lojistik regresyon, bir regresyon algoritması değil, bir sınıflandırmadır. Belirli bir bağımsız değişken(ler) kümesine dayanarak ayrık değerleri (0/1, Evet/Hayır, doğru/yanlış gibi ikili değerler) tahmin eder. Basitçe söylemek gerekirse, temel olarak, verileri bir logit işlevine uydurarak bir olayın oluşma olasılığını tahmin eder. Bu nedenle "logit regresyonu" olarak da bilinir. Elde edilen değerler, olasılığı öngördüğü için her zaman 0 ve 1 arasında olacaktır. Başka bir örnek üzerinden bu anlayalım:

Diyelim ki matematik testinde bir miktar var. Sadece 2 sonucu olabilir, değil mi? Ya bunu çözersiniz ya da çözemezsiniz (ve burada yöntem için puan almayalım). Şimdi, hangi bölümleri iyi anladığınızı anlamak için çok çeşitli toplamalar verildiğini hayal edin. Bu çalışmanın sonucu böyle bir şey olacaktır – eğer Trigonometri tabanlı bir problem verilirse, bunu çözme olasılığınız %70 olur. Öte yandan, eğer aritmetik bir problemse, o zaman bir cevap alma olasılığınız sadece %30'dur. Lojistik Regresyonun size sağladığı şey budur.

Lojistik Regresyon Modeli

$\ln[p/(1-p)] = \beta_0 + \beta_1 x$ p, Y olayının meydana gelme

olasılığıdır. $p(Y=1)$

$[range=0,1]$

$p/(1-p)$ "oran oranı" dır.

$[range=0,\infty]$

$\ln[p/(1-p)]$: log oran oranı "logit" $[range=-\infty, +\infty]$

$f(x) = 1 / (1 + e^{-w \cdot x})$ bunu bu şekilde yorumlayacağız $\rightarrow f(x) = P(y=1 | x)$ (kısaca p)

Ayrıca:

$P(y=1 | x) = f(x)$

$P(y=0 | x) = 1 - f(x)$

Bu iki kuralı birleştirerek daha kompakt bir şekilde yazılabiliriz:

$P(y | x) = (f(x))^y (1 - f(x))^{1-y}, y \in \{0, 1\}$