countbayesie.com/blog/2019/6/12/logistic-regression-from-bayes-theorem

## Logistic Regression from Bayes' Theorem

- Lojistik regresyonu bayes teoreminden nasıl elde edeceğimizi,
- Lojistik regresyonun Bayes teoremindeki prior ve likelihoodu verilerimizden öğrenebilmemizin bir yolu olduğunu göreceğiz.

Bayes teoremini hatırlayacak olursak;

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Basic lineer modeli hatırlayacak olursak;

$$y = \beta x + \beta_0$$

modeli hedef değişken y'nin x çarpı $\beta$ kat sayıları artı bir miktar sabit $\beta_0$'ın lineer bir kombinasyonu olarak anlaşılabileceğini söylüyordu.

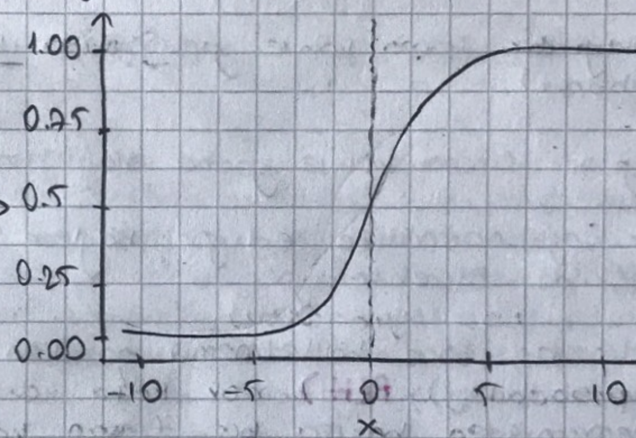
## Lojistik Regresyon Temelleri

Hızlı bir tazeleme olarak, lojistik regresyon, bazı hipotezlerin olasılığını tahmin etmek için verileri kullanmanın yaygın bir yöntemidir. Daha matematiksel olarak konuşursak, bir x girdimiz var, (bu birinin boyu gibi tek bir değer olabilir veya görüntüdeki pikseller gibi bir vektör olabilir) y hedef değişkeni bu bir kedi resmidir gibi bir çıktıyı temsil eder. Lojistik regresyondaki hedefimiz x verildiğinde y'nin olasılığını öğrenmektir. Model, y'nin binary bir çıktı, 1'in başarı ve 0'ın başarısızlık olduğu ve x'in y çıktısıyla sonuçlanan veriye karşılık gelen bir örnek olduğu örnekler üzerinde eğitilir. Modeli eğittiğimizde, bir y vektörüne ve satırları training örneklerini, sütunları özellikleri (features) temsil eden bir X matrisine sahibiz.

Buraya kadar her şey yolunda ama burada işler genellikle biraz kafa karıştırıcı hale geliyor. Bu konuyla ilgili x verildiğinde y'nin olasılığını nasıl hesapladığımızı gösteren aşağıdaki denklem hemen gösteriliyor.

$$p(y|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Lojistik fonksiyon




Bu ters lojit (inverse logit) veya lojistik fonksiyondur. $\beta_0$ ve $\beta$ eğitim sırasında öğreneceğimiz parametrelerdir.

Bunun arkasındaki sezgi, genellikle lojistik fonksiyonun çok büyük pozitif sayıları 1'e ve çok büyük negatif sayıları 0'a yakın olmaya zorlaması olarak açıklanır ki bu da nihayetinde olasılıkların nasıl görünmesini istediğimizdir. Bu iyi bir açıklama, ancak lojistik regresyon ve Bayes teoremi arasındaki gerçekten güzel (ve kullanışlı) bağlantıyı kaçırıyor.

## İyi bir fincan kahve yapmak

— Mükemmel bir fincan kahve yapmak için farklı miktarlarda kahve, farklı hacimlerde su vb. ile deneyler yapmaya başladım. Mükemmel fincanı olasılıksal olarak nasıl modelleyeceğimizi düşünelim.

İlk olarak verilerimizi temsil etmek istiyoruz. Aynı çekirdekleri kullansak bile, demleme kalitesine giren birçok faktör var: Çekirdeklerin yaşı, öğütmenin inceliği, telvenin ağırlığı, dökme süresi, suyun sıcaklığı, ve elbette, muhtemelen daha da fazla şey bulabiliriz. Bu yazıda gerçek veriler hakkında endişelenmeyeceğiz, sadece bu problemi olasılıksal olarak nasıl modelleyeceğimize bakacağız.

## Bayesian Brewing (Demleme)

Bir fincan hazırlarken açıklayabileceğimiz tüm olası şeyleri veri vektörümüz D olarak düşünelim. Ve "harika bir fincan kahve" olan H hipotezimizin olasılığını bilmek istiyoruz. Şimdi sadece "Demleme düzenim göz önüne alındığında harika bir fincan kahve alma şansım nedir?" sorusuna cevap vermek istiyoruz. Yani P(H|D)'i bilmek istiyoruz. Verilerimizle bir hipotezin olasılığını bilmek istediğimizde Bayes teoremine dönebiliriz.

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Bu denklemin herbir parçası için Bayes terimlerini ve ne anlama geldiklerine bakalım.

- P(H|D), düzenimiz göz önüne alındığında harika bir fincan kahve yapma olasılığımızdır (our posterior probability).

- P(D|H), iyi bir fincan kahve yaptığımızda bu düzene sahip olma olasılığımızdır (likelihood).

- P(H), iyi bir fincan kahve yapma olasılığımızdır (prior probability).

- P(D) her şeyi normalize ediyor, böylece olasılığımız 0 ile 1 arasında uygun bir şekilde ölçekleniyor.

→ Devam etmeden önce halletmemiz gereken birkaç şey var. Önsel olasılığımız (our prior probability), P(H) çok kafa karıştırıcı değil, sadece genel olarak demlememizden harika bir fincan kahve alacağımıza dair ön inanç. Belki hatırladığımız fincanların yarısı harikadır, belki sadece 1000'de 1'i, ama her iki şekilde de bunun temsil ettiği şey budur.

→ P(D|H)'i (likelihoodu) düşünmek biraz zor, çünkü "iyi bir fincan kahve istediğimde bu düzeni kurma ihtimalim ne kadar"ı bilmek istiyoruz. Bu, problemimiz hakkında düşünmenin tuhaf bir yolu. Ama tam da bu yüzden burada "makine öğrenmini" kullanmak istiyoruz. Bu likelihood için basit bir model bulacağız, böylece bunu verilerden öğreneceğiz.

→ Geriye sadece verilerimizin olasılığını ifade eden P(D) kaldı. "Kahve demlemek için bu düzenin kurulma olasılığı nedir?" Elbette kahve demleme düzeni için neredeyse sonsuz olasılık var, bu yüzden bu bir problem. Bunu çözemezsek posterior probabilityi (P(H|D)) çözemeyiz.

## Posterior oddsları Kullanarak Hipotezleri Karşılaştırma

P(D)'i gerçekten bilmediğimiz gerçeğiyle ilgilenmemiz gerekiyor, sorunumuzu yeniden çerçeveleyerek bunu çözmenin bir yolu var. Şu anda sadece bir hipotezi düşünüyoruz, o da bizim bir fincan kahvemizin harika olması durumu, ama belli ki buna alternatif bir durum var.

Sadece kahvemizin harika olmadığı inancı olan $\bar{H}$'yi düşünürsek, posteriorları karşılaştırabilir ve problemimize P(D)'e ihtiyaç duymadan bakabiliriz. Şimdi elimizde bir de $P(\bar{H}|D)$ var ve bunun $P(H|D)$ ile oranına bakacağız.

$$\frac{P(H|D)}{P(\bar{H}|D)} = \frac{P(D|H)\,P(H)\,.\,\frac{1}{P(D)}}{P(D|\bar{H})\,.\,P(\bar{H})\,.\,\frac{1}{P(D)}}$$

Tabi ki $1/P(D)$ hem payda hem de paydada görünüyor, bu yüzden ondan kurtulabiliriz. Bu, artık P(D) için endişelenmemize gerek olmadığı anlamına gelir.

$$\frac{P(H|D)}{P(\bar{H}|D)} = \frac{P(D|H)\,P(H)}{P(D|\bar{H})\,P(\bar{H})}$$

Burada elde ettiğimiz şey, H için posterior oddsu hesaplama formülüdür. Odds, $P(H|D)$'nin $P(\bar{H}|D)$'den kaç kat olası olduğunu ifade eder.
Ve $P(H|D)+P(\bar{H}|D)=1$ olduğu için en nihayetinde oddslarımızdan $P(H|D)$'e geri dönebiliriz. (Bu şekilde birbirinin tamamlayıcısı olmayan iki hipotezi karşılaştırmamız doğru olmaz) Modelimizi öğrenmenin püf noktası aslında $\bar{H}$ örneklerine sahip olmamızdır.

Artık olasılıklar yerine odds hakkında konuştuğumuz için prior odds'umuz olan $P(H)/P(\bar{H})$ ifadesini O(H) olarak ifade edelim. Daha temiz bir formül yazacak olursak;

$$O(H|D) = \frac{P(D|H)}{P(D|\bar{H})}\,.\,O(H)$$

Sorunumuzu çözmek için fena bir başlangıç değil. Ardından, bu denklemin sağ tarafını verilerimizden öğrenmenin bir yoluna ihtiyacımız var.

## Lineer Bir Model Bulma

Kullanılacak en basit ve çoğu zaman en kullanışlı model genellikle lineer bir modeldir.

$$y = \beta x + \beta_0$$

Bu sadece, y'nin bazı β sabit oranlarından bazı βeta0 kesişmeler ile arttığı veya azaldığı anlamına gelir. Bu, şeylere bakmanın basit bir yoludur, ancak elimizde veri varsa, bunun için en uygun parametreyi doğrusal regresyon ile öğrenebiliriz. Ne yazık ki, kahve sorununa şu andaki olasılıksal çözümümüz böyle bir şeye benzemiyor... henüz.

## Kurtarma için Log Dönüşümü

Olasılık problemimizin odds formunun içinde sadece çarpma ve bölme olduğunu farkedeceksiniz, bu da güzel bir lineer çözümden birkaç uzakta olduğumuzu gösteriyor. Ancak bunu doğru biçimde elde etmek için çok faydalı bir trick var: basitçe log dönüşümü yapabiliriz. 10 tabanındaki şeyler hakkında sezgiler yapmak daha kolay olduğu için şimdilik log10 kullanacağız.

$$O(H|D) = \frac{P(D|H)}{P(D|\bar{H})} \cdot O(H)$$

$$\log(O(H|D)) = \log_{10}\left(\frac{P(D|H)}{P(D|\bar{H})} \cdot O(H)\right)$$

$$\log(O(H|D)) = \log_{10}\left(\frac{P(D|H)}{P(D|\bar{H})}\right) + \log_{10}(O(H))$$

Dikkat edersek şimdi bir lineer denklemimiz var. $O(H)$'ın veri vektörümüze hiç bağlı olmadığına dikkat edin, tıpkı lineer modelde olduğu gibi $\beta_0$ sadece bir sabittir ve X'e bağlı değildir. Yani lineer modelimiz için şunu söyleyebiliriz:

$$\beta_0 = \log_{10}(O(H))$$

$D \to x$

$\log_{10}\left(\frac{P(D|H)}{P(D|\bar{H})}\right) \to \beta x$

Veya $\beta_0$, prior odds'un logudur.

Şimdi modelimizin kalbine geliyoruz. Prioru göz ardı ederek, log likelihood oranının basitçe D'nin lineer bir fonksiyonu olduğu şeklindeki basitleştirici varsayımı yapmaya devam edeceğiz. Örneğin, belki de suyumuzun sıcaklığındaki bir artma log-likelihood'un lineer olarak artmasına neden olur. Bu harika bir özellik, çünkü, bir şeyin olasılığını 0.01'den 0.1'e çıkarırsak olasılıkta 0.09'luk doğrusal bir artış değil, üstel bir artış olur! Dolayısıyla olasılıkları doğrusal bir şekilde modellemek istiyorsak, log dönüşümü yapılmış veriler açısından düşünmek isteyeceğiz.

$\log_{10}(0.01) = -2$
$\log_{10}(0.1) = -1$ $)\,{-1}$

$(10^{-1})$ artış

Bu varsayımı yaparsak, likelihood oranını $\beta D$ olarak modelleyebiliriz. Ve şimdi problemimize güzel bir lineer çözümümüz var. Burada $\log(O(H|D))$'yi $lo(H|D)$ olarak göstereceğiz.

$$lo(H|D) = \beta D + \beta_0$$

Bu lineer form ile log formunda verinin lineer bir fonksiyonu olarak likelihood oranını ve prior oddsu öğrenebiliriz. Lojistik regresyonu lineer bir model yapan şey budur, üstünde likelihoodun $P(D|H)$'in nihai olarak inputlarla lineer bir ilişkisi olduğunu varsayıyoruz. Ancak bu doğrusal ilişkiyi görmek için çıktımızı log oddsa dönüştürmemiz gerekiyordu.

## Where we are so far: probabilities, odds and log odds

Şimdiye kadar neler olduğunu bildiğimizden emin olmak için biraz özetleyelim. $P(H|D)$'i bilmek istedik ki bu, D demleme düzenini göz önüne aldığımızda harika bir kahve alma olasılığımızı ifade ediyordu. Bayes teoremiyle $P(D)$'yi hesaplamanın bir yolunu bulamasak da, bu sorunu neredeyse çözebildik. Bu, yalnızca $P(H|D)$'e bakmak yerine kahvenin harika olma olasılığıyla olmama olasılığı $\bar{H}$'yi karşılaştıran odds'a $O(H|D)$ bakmamız gerektiği anlamına geldi. Odds, bir hipotezin diğerine ne kadar muhtemel olduğunun oranları açısından bize sonuçlar verecektir:

- $O(H|D) = 10$ kahvenin iyi olma ihtimalinin iyi olmama ihtimalinden onlarca kat daha fazla olduğu anlamına gelir.

- $O(H|D) = 1/10$ kahvenin iyi olmama olasılığının 10 kat daha fazla olduğu anlamına gelir.

Odds formatının asimetrik olduğuna dikkat edin, çünkü hipotezimiz için kanıtlar büyüdükçe sonuç sonsuza doğru büyür ve hipotezimize karşı kanıtlar büyüdükçe odds 0'a düşer.

Oddsu $\log_{10} O(H|D)$'a çevirdiğimizde bu asimetriyi düzeltiriz:

- $\log_{10} O(H|D) = 1$ harika kahvenin 10 kat daha olası olduğu anlamına gelir
- $\log_{10} O(H|D) = 2$ harika kahvenin 100 kat daha olası olduğu anlamına gelir
- $\log_{10} O(H|D) = -1$ harika kahvenin 10 kat daha az olası olduğu anlamına gelir
- $\log_{10} O(H|D) = -2$ harika kahvenin 100 kat daha az olası olduğu anlamına gelir

Bu yüzden, problemimize bakmamız için bize güzel bir lineer yol vermesinin yanı sıra, problemimizi log odds ile çerçevelemek, sonuçları yorumlamaya çalıştığımızda aslında çok mantıklı geliyor.

## Modelimizi Öğrenme Sorunu

Problemimiz için temelde lineer regresyona benzeyen güzel bir lineer formatımız var.

$$y = \beta x + \beta_0$$

Şimdi bitirmemiz ve problemimizi diğer herhangi bir lineer regresyon probleminde yaptığımız gibi en küçük kareleri en aza indirerek (minimizing least squares) çözebilmemiz doğal görünüyor. Bu yaklaşımda, D'yi olduğu gibi kullandığımızdan ve verilerimizin değerlerine bağlı olarak log oddsun arttığını veya azaldığını varsaydığımız için verilerimizi dönüştürmemize gerek yok. Ancak hedef değişkenimizi nasıl dönüştüreceğimizi düşünmemiz gerekiyor. Üzerinde çalışmak istediğimiz hedef y için zaten $P(H|D)$ formunda verimiz var. Başarılı olan durumlar için $P(H|D)=1$ olduğunu ve başarısız olan durumlar için $P(H|D)=0$ olduğunu biliyoruz. Lineer modelimizi eğitmek için önce hedef verilerimizi log odds formuna dönüştürmemiz gerekiyor.

Ama burada can sıkıcı bir sorun var! Bir olasılığı oddsa dönüştürmek için şu basit kuralı takip edebiliriz:

$$O(H) = \frac{P(H)}{1-P(H)}$$

Ancak burada bir sorun olduğunu görebiliriz çünkü olasılıklar mutlak 1 veya 0'dır. Pozitif caseler için odds tanımsız olan 1/0'dır. Ve negatif durum için oddsun logunu almak istediğimizde bu sorunu çözsek bile yapamayız çünkü bunlar 0/1 olacak ve log0 da anlamsız.

## Log Oddsu Tekrar Olasılıklara Çevirmek: Ters Logit (Inverse logit)

Çözümümüze sınır bozucu bir şekilde yakınız ve henüz tam olarak oraya ulaşamasak da değerli bir şey öğrendik. Olasılık problemimizin merkezinde lineer bir model var. Hedef değişkenimizi dönüştüremiyoruz, ancak bu lineer modelin kendisini tekrar bir olasılığa dönüştürebilirsek, o zaman çözümümüzü bulmuş olacağız.

Bu şaşırtıcı derecede kolay. Yaptığımız her şeyi geri almamız gerekiyor, ama bu sefer bunu lineer modele yapıyoruz. Modelimiz şu anda log odds cinsinden yazılmıştır, bu yüzden yapmamız gereken ilk şey log dönüşümünü geri almaktır. Bunu lineer denklemimizin 10. kuvvetini alarak yapabiliriz.

$$\log(O(H|D)) = \beta D + \beta_0 \longrightarrow O(H|D) = 10^{(\beta D+\beta_0)}$$

Bu oldukça kolaydı! Şimdi sadece oddsu olasılıklara çevirmemiz gerekiyor, bu olasılıkları oddsa dönüştürmek kadar kolay. Bu kuralı kullanabiliriz!

$$P(X) = \frac{O(X)}{1+O(X)}$$

$$O(X) = \frac{P(X)}{1-P(X)}$$

$$O(X) - O(X).P(X) = P(X)$$

$$O(X) = P(X)\,(1+O(X))$$

$$\frac{O(X)}{1+O(X)} = P(X)$$

Bunu yaparsak şunu görebiliriz!

$$P(H|D) = \frac{10^{(\beta D+\beta)}}{1+10^{(\beta D+\beta)}}$$

Hedef değerlerimizi dönüştüremediğimiz için $\beta D + \beta_0$'ı dönüştürmek zorunda kaldık, ama sorun değil çünkü her iki şekilde de aynı etkiye sahip ve bu sefer sonuçlarımızın (outcomes) mutlak 1'ler ve 0'lar olduğu gerçeğini kaldırabiliriz.

Bunu daha güzel ve matematiksel olarak daha kabul edilebilir kılmak için yapabileceğimiz iki basitleştirme daha var. İlk olarak, hiçbir ciddi matematikçi $\log_{10}$'u kullanmaz, bunun yerine $\ln$'i tercih eder. Bu yüzden bu 10'u e ile değiştirmemiz gerekir. Sonuçlarımızı taban 10'da tutmakla ilgili özel bir durumumuz yok, bu yüzden bu değişiklikte hiçbir sorun yok, etkiler aynı olacak.

$$P(H|D) = \frac{e^{(\beta D + \beta_0)}}{1 + e^{(\beta D + \beta_0)}}$$

Ve aynı zamanda oldukça uygun bir şekilde ortaya çıkıyor:

$$\frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Bu, son denklemimizi başlangıçta gördüğümüz gizemli formüle dönüştürebileceğimiz anlamına gelir.

$$P(H|D) = \frac{1}{1 + e^{-(\beta D + \beta_0)}}$$

Burada, bu formülün log oddsu tekrar olasılığa dönüştürmenin bir yolu olduğunu görüyoruz. Bu, elbette, kelimenin tam anlamıyla "inverse logit"in ne anlama geldiğidir, "logit" "log odds" fonksiyonudur. Logit fonksiyonu olasılıkları alır ve bunları log oddsa dönüştürür. "Inverse logit" log oddsu alır ve onları olasılıklara dönüştürür.

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Bayes' Theorem

$$P(H|D) = \frac{1}{1 + e^{-(\beta D + \beta_0)}}$$

inverse logit model

$$O(H|D) = \frac{P(D|H)}{P(D|\bar{H})} \cdot \frac{P(H)}{P(\bar{H})}$$

posterior odds

$$lo(H|D) = \beta D + \beta_0$$

log odds linear model

Lojistik regresyonu türetmekten elde ettiğimiz temel anlayışlardan biri, lojistik regresyonun nasıl lineer bir model olduğunu çok açık bir şekilde görmektir. Başlangıçta problemimizi Bayes teoremi olarak modelliyoruz, ancak hipotezimiz için verilen verilerin likelihoodunu ve prior probabilityi bilmiyoruz. Bunları verilerden öğrenebilmek istiyoruz ve bunu yapmak için nihayetinde hipotezimizin log oddları ile verimiz D arasında lineer bir ilişki olduğu gibi basitleştirici bir varsayımda bulunuyoruz.

Tüm bunları gözden geçirdikten sonra, Bayes teoremi ve lineer regresyonun temellerinden başlayarak bunu istediğimiz zaman yeniden oluşturabiliriz. Artık inverse logit kullanarak $\beta$ ve $\beta_0$'ı öğrenmek için bir model eğitebiliriz.

## Sonuç

Bu, bir fincan kahve yapmak için oldukça karmaşık bir yol olsa da, Bayes teoremi ile lojistik regresyon arasındaki ilişkiyi öğrenmek, lojistik regresyonun nasıl çalıştığına dair oldukça güçlü içgörüler sağlar.

Ayrıca genelleştirilmiş lineer modelin (generalized linear model) gücünün de biraz tadına baktık. Burada yaptığımızın büyük bir kısmı, lineer bir modeli dönüştürebilecek bir fonksiyon, bu durumda ters logit bulmaktı. Biraz ayrıştırma ile bunun, lineer modelin diğer dönüşümlerine dayalı olarak çeşitli diğer regresyonları gerçekleştirmek için genişletilebileceğini görebiliriz.