ÖRÜNTÜ TANIMA - BAHAR VİZE ÖDEVİ

Sevdenur YILMAZ – 180508036

NAIVE BAYES SINIFLANDIRICISI

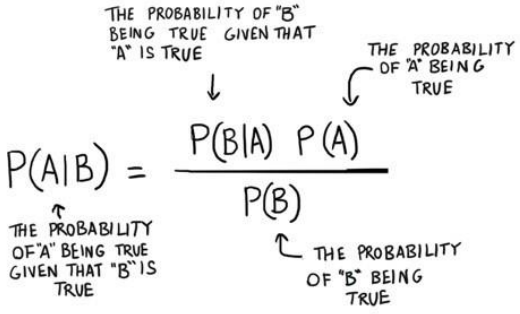
Sınıflandırıcı, belirli özellikleri temel alan farklı nesneleri ayırt etmek için kullanılan bir makine öğrenmesi modelidir. Veri setimizdeki verilerin belirli özelliklerine bakarak hedefimizi kategorilere ayırmamızı sağlar.

Naive Bayes 18.yy'da Thomas Bayes'in Bayes Teoremi temel alarak geliştirilmiştir.

Naive Bayes sınıflandırıcısı, örüntü tanıma problemine ilk bakışta oldukça kısıtlayıcı görülen bir önerme ile kullanılabilen olasılıksal bir yaklaşımdır. Bu önerme, örüntü tanımada kullanılacak her bir tanımlayıcı öznitelik ya da parametrenin istatistik açıdan bağımsız olması gerekliliğidir.

Naive Bayes sınıflandırıcısı, sınıflandırma görevi için kullanılan olasılıklı bir makine öğrenme modelidir. Sınıflandırıcının noktası, Bayes Teoreminin bağımsızlık önermesiyle basitleştirilmiş halidir. Bayes Teoremi, olasılık kuramı içinde incelenen önemli bir konudur. Bu teorem bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Bu şekli ile Bayes Teoremi kabul edilir bir ilişkiyi açıklar.

Olasılık teorisi içinde incelenen bir 'olay olarak B olayına koşullu bir A olayı (yani B olayının bilindiği halde A olayı) için olasılık değeri, A olayına koşullu olarak B olayı (yani A olayı bilindiği haldeki B olayı) için olasılık değerinden farklıdır. Ancak bu iki birbirine ters koşulluluk arasında çok belirli bir ilişki vardır ve bu ilişkiye Bayes Teoremi denilmektedir.



- ✓ P(*A*|*B*) ; B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır (bakınız koşullu olasılık)
- ✓ P(*B*|*A*) ; A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır
- ✓ P(A) ve P(B) ; A ve B olaylarının gerçekleşme olasılıklarıdır.
- ✓ Örneğin P(A) henüz elde veri toplanmadan A olayı hakkında sahip olunan bilgidir. Diğer taraftan P(*B*|*A*) ardıl olasılıktır çünkü veri toplandıktan sonra, A olayının gerçekleşmiş olduğu durumlarda B olayının gerçekleşme ihtimali hakkında bilgi verir.

SINIFLANDIRMA PROBLEMİ

Naive Bayes Sınıflandırması Makine öğreniminde öğreticili öğrenme alt sınıfındadır. Daha açık bir ifadeyle sınıflandırılması gereken sınıflar(kümeler) ve örnek verilerin hangi sınıflara ait olduğu bellidir.

Sınıflandırma işleminde genel olarak elde bir örüntü (pattern) vardır. Buradaki işlem de bu örüntüyü daha önceden tanımlanmış sınıflara sınıflandırmaktır. Her örüntü nicelik (feature ya da parametre) kümesi tarafından temsil edilir.

NİCELİK KÜMESİ

Yine yukarıda bahsedilen spam e-posta örneğinden devam edilecek olunursa; Posta kutumuzda bulunan spam e-postaları spam olmayan e-postalardan ayıran parametrelerden oluşan bir küme, mesela *ikramiye,ödül* gibi sözcüklerden oluşan, nicelik kümesine örnektir. Matematiksel bir ifadeyle nicelik kümesi;

$x(i), i = 1, 2, \ldots, L$, ise $x = [x(1), x(2), \ldots, x(L)]^T$ $\in R^L$ L-boyutlu nicelik vektörünü oluşturur.

$x \in R^L$ verildiğine göre ve *S* ayrıştırılacak sınıflar kümesiyse, Bayes teoremine göre aşağıdaki ifade yazılır.

$x(i), i = 1, 2, \ldots, L$, ve $p(x) = \sum_{i=1}^{L} p(x|S_i)P(S_i)$

- $P(S_i)$; S_i'nin öncel olasılığı i = 1, 2, . . . , L,
- $P(S_i|x)$; S_i'nin ardıl olasılığı
- p(x); x in Olasılık yoğunluk fonksiyonu (oyf)
- $p(x|S_i)$; i =1 = 2, . . . , L, x'in koşullu oyf'si

BAYES KARAR TEOREMİ

Elimizde sınıfı belli olmayan bir örüntü olsun. Bu durumda

$x = [x(1), x(2), \ldots, x(L)]^T \in R^L$

sınıfı belli olmayan örüntünün L-boyutlu nicelik vektörüdür. Spam e-posta örneğinden gidecek olursak spam olup olmadığını bilmediğimiz yeni bir e-posta sınıfı belli olmayan örüntüdür.
Yine S_i x'in atanacağı sınıf ise;
Bayes karar teorisine göre x sınıf S_i'ya aittir eğer

$$P(S_i|x) > P(S_j|x), \forall j \neq, i$$

diğer bir ifadeyle eğer

$$P(x|S_i)P(S_i) > P(x|S_j)P(S_j), \forall j \neq, i$$

NAİVE BAYES SINIFLANDIRMASI

Verilen bir x'in ($x = [x(1), x(2), \ldots, x(L)]^T \in R^L$) sınıf S_i'ye ait olup olmadığına karar vermek için kullanılan yukarıda formüle edilen Bayes karar teoreminde istatistik olarak bağımsızlık önermesinden yararlanılırsa bu tip sınıflandırmaya Naive bayes sınıflandırılması denir. Matematiksel bir ifadeyle

$$P(x|S_i)P(S_i) > P(x|S_j)P(S_j), \forall j \neq, i$$

İfadesindeki

$P(x|S_i)$ terimi yeniden aşağıdaki gibi yazılır

$$P(x|S_i) \approx \prod_{k=1}^{L} P(x_k|S_i)$$

böylece Bayes karar teoremi aşagıdaki şekli alır. Bayes karar teorisine göre x sınıf S_i'ya aittir eğer

$$P(S_i) \prod_{k=1}^{L} P(x_k|S_i) > P(S_j) \prod_{k=1}^{L} P(x_k|S_j)$$

$P(S_i)$ ve $P(S_j)$ i ve j sınıflarının öncel olasılıklarıdır. Elde olan veri kümesinden değerleri kolayca hesaplanabilir.
Naive bayes sınıflandırıcının kullanım alanı her ne kadar kısıtlı gözükse de yüksek boyutlu uzayda ve yeterli sayıda veriyle x'in (nicelik kümesi) bileşenlerinin istatistik olarak bağımsız olması koşulu esnetilerek başarılı sonuçlar elde edilebilinir.

Hava	Oynama Durumu
Güneşli	Hayır
Güneşli	Hayır
Bulutlu	Evet
Yağmurlu	Evet
Yağmurlu	Evet
Yağmurlu	Hayır
Bulutlu	Evet
Güneşli	Hayır
Güneşli	Evet
Yağmurlu	Evet
Güneşli	Evet
Bulutlu	Evet
Bulutlu	Evet
Yağmurlu	Hayır

ÖRNEK: Aşağıda hava durumu(X) ve ona karşılık gelen kategorik oyunu oynama durumları(y) yer alıyor. Hava durumuna göre futbol oynayıp oynamama durumlarını Naive Bayes algoritması ile tahminlemeye çalışılsın. Eğitim verisi yandaki gibidir:
Bir soru sorup ve olasılık hesaplama:

- Örnek soru: Hava yağmurlu olduğunda oyun oynar mıyım?
- P(Evet|Yağmurlu)=P(Yağmurlu|Evet)*P(Evet)/P(Yağmurlu)
- İlk olarak evet olduğu bilindiğinde yağmurlu olma olasılığı: Toplam 9 evet içerisinden 3 tanesi yağmurludur. P(Yağmurlu|Evet) = 3/9 = 0.33 olasılığı elde edilir.
- "Evet" ve "Yağmurlu" olma olasılıkları:
- Toplam 14 gözlem içerisinden 9 tanesi "Evet" şeklindedir. P(Evet) = 9/14 = 0.64
- Toplam 14 gözlem içerisinden 5 tanesi "Yağmurlu" şeklindedir. P(Yağmurlu) = 5/14 = 0.36
  - Değerleri yerine yazdıktan sonra ve hava yağmurlu olduğunda oyun oynama olasılığı:
  - P(Evet|Yağmurlu) = P(Yağmurlu|Evet)*P(Evet)/P(Yağmurlu) = 0.33*0.64/0.36 = **0.59**

Yani hava yağmurlu olduğunda futbol oynama olasılığı 0.59 olarak belirlendi. Tabi burada hesaplanan bir olasılık değeridir. Ancak bizden "oynarım" ya da "oynamam" gibi net bir tahmin beklenmektedir. Naive Bayes algoritması bu tahmini yaparken ihtimalin yüksek olduğu duruma göre yuvarlar. Burada "%59 ihtimalle oynarım" sonucu çıkmaktadır. Naive Bayes algoritmasına göre bu durum "Evet" şeklinde belirlenir.

NAIVE BAYES TÜRLERİ

- ✓ Gaussian Naive Bayes: Tahmin ediciler sürekli bir değer aldıklarında ve ayrık olmadıklarında, bu değerlerin bir gauss dağılımından örneklendiğini varsayıyoruz.
- ✓ Multinomial Naive Bayes: Bu çoğunlukla belge sınıflandırma problemi için kullanılır, yani bir belgenin spor, politika, teknoloji vb. kategorisine ait olup olmadığı. Sınıflandırıcının kullandığı özellikler / öngörücüler belgede bulunan kelimelerin sıklığıdır.
- ✓ Bernoulli Naive Bayes: Multinomial Naive Bayes'e benzer, ancak tahmin ediciler boole değişkenleridir. Sınıf değişkenini tahmin etmek için kullandığımız parametreler, örneğin metinde bir kelime olduğunda veya olmasa da, sadece evet veya hayır değerlerini alır.

NAIVE BAYES SINIFLANDIRICISININ AVANTAJLARI

- ✓ Her özellik birbirinden bağımsız kabul edildiği için Lojistk Regresyon gibi modellerden daha iyi performans gösterebilir.
- ✓ Az veriyle iyi işler başarabilir. Hızlıdır ve gerçek zamanlı sistemlerde kullanılabilir.
- ✓ Sürekli ve kesikli veriler(continuous and discrete) ile kullanılabilir.
- ✓ Yüksek boyutlu verilerde(High Dimension Data) çalışabilir.

NAIVE BAYES SINIFLANDIRICISININ DEZAVANTAJLARI

- ✓ Özellikler birbirinden bağımsız varsayılarak işlem yapıldığı için değişkenler arası ilişkiler modellenemez.
- ✓ Sıfır olasılık(Zero Probability) problemi; Sıfır olasılık istediğimiz örneğin veri setinde hiç bulunmaması durumudur. Yani herhangi bir işleme alındığında sonucu sıfır yapacaktır. Bunun için en basit yöntem tüm verilere minimum değer ekleyerek (genellikle 1) bu olasılık ortadan kaldırılabilir. Bu duruma Laplace kullanılarak tahminleme de denmektedir.

NAIVE BAYES UYGULAMA ALANLARI

Genel olarak veri madenciliğinde, biyomedikal mühendisliği alanında, hastalıkların tıbbi tanımlanmasında, EKG grafiğinin sınıflandırılmasında, EEG grafiklerinin ayrıştırılmasında, genetik araştırmalarında, yığın mesaj tanımlanmasında, metin ayrıştırılmasında (spam filtreleme, duygu analizi), gerçek zamanlı sistemlerde, öneri sistemlerinde, çok sınıflandırma problemlerinde vs kullanılır.