T.C

SAKARYA ÜNİVERSİTESİ

BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BSM 310 – YAPAY ZEKA

Grup üyeleri:

G171210551 RECEP İLYASOĞLU
G171210041 MUHAMMET FATİH HAN UZUN
G171210013 YUNUS EMRE TOSUN
G171210049 AHMET BUDAK

Sakarya 2020

NAİVE BAYES SINIFLANDIRICI

Adını İngiliz Matematikçi Thomas Bayes 'ten alır. Naïve Bayes Sınıflandırıcı Örüntü tanıma problemine ilk bakışta oldukça kısıtlayıcı görülen bir önerme ile kullanılabilen olasılıkcı bir yaklaşımdır. Bu önerme örüntü tanımada kullanılacak her bir tanımlayıcı öznitelik ya da parametrenin istatiksel açıdan bağımsız olması gerekliliğidir. Her ne kadar bu önerme Naive Bayes sınıflandırıcının kullanım alanını kısıtlasa da , genelde istatistik bağımsızlık koşulu esnetilerek kullanıldığında da daha karmaşık yapay sinir ağları gibi metotlarla karşılaştırabilir sonuçlar vermektedir. Bir Naive Bayes sınıflandırıcı, her özniteliğin birbirinden koşulsal bağımsız olduğu ve öğrenilmek istenen kavramın tüm bu özniteliklere koşulsal bağlı olduğu bir Bayes ağı olarak da düşünülebilir.

Örnek

Örneğimiz, aşağıda verilen tablodaki verilerden bir sınıflandırıcı eğitimi olsun

Kısım	Maaş	Yaş	İş Tecrübesi
Yazılım	3000	26	4
Muhasebe	1500	22	2
Yazılım	5000	30	9
Muhasebe	2000	30	7
Muhasebe	500	18	3
Yazılım	2000	20	2
Yazılım	7000	29	5
Muhasebe	6000	45	15

Yukarıda, sadece muhasebe ve yazılım kısımlarında çalışan kişilerin maaş, yaş ve iş tecrübelerini içeren temsili bir tablo verilmiştir. Buna göre aşağıdaki şekilde bize bir bilgi verilse:
Maaş : 3000, Yaş : 30, Tecrübe : 5yıl
Bu kişinin hangi kısımda çalıştığını acaba bulabilir miyiz?
Öncelikle eğitim ile işe başlayalım sonra da testimizi yaparız. Basitçe veri kümemizideki (data set) Yazılım ve Muhasebe kısımlarının ortalama ve varyans değerlerini hesaplıyoruz. Bu değerler aşağıdaki şekildedir:

Muhasebe	2500	28,75	6,75
Yazılım	4250	26,25	5
Muhasebe	5833333,333	142,25	34,91666667
Yazılım	4916666,667	20,25	8,666666667

Yukarıdaki ilk iki satır ortalama ve ikinci iki satır ise varyans değerleridir. Bu değerleri basitçe Excel ile hesapladık.
Şimdi beklenen değeri hesaplayacağız. Yani gelen test verimizin Muhasebe kısmında birisine ait olması veya Yazılım kısmından birisine ait olması için beklenen durum hesabı yapacağız.
Hesaplamaya geçmeden önce yazının konusu olan naive bayes kavramını hızlıca açıklayalım. Aslında naive bayes sınıflandırıcısı basiçte bütün koşullu olasılıkların çarpımıdır. Aşağıdaki şekilde gösterilebilir:

$$sınıflandırma(s_1, s_2, \dots, s_n) = azami_c\, p(K=k) \prod_{i=1}^{n} p(S_i = s_i | K = k)$$

Bu formülde görüleceği üzere s1'den sn'e kadar olan sınıflar arasından bir seçim yapılırken aslında bu sınıfın olasılık değeri ve bu sınıfları yerine getiren k koşulları için çarpımından bir farkı yoktur.
Yani diğer bir deyişle her sınıfın bir koşullu olasılık değeri vardır ve biz sınıflardan hangisine ait olduğunu bulmak için bu koşullu olasılık değerlerini çarparız.
Bu durumda bizim formülümüz aşağıdaki şekildedir denilebilir:

$$beklenti(Yazılım) = \frac{P(Yazılım)\, p(maaş|yazılım)\, p(Yaş|yazılım)\, p(iş\ tecrübesi|yazılım)}{normalleştirme}$$

Yani bir kişinin yazılım kısmında olduğunu anlamak için öncelikle yazılım kısmında olan kişilerin oranını buluyoruz, P(Yazılım), ardından yazılım kısmındaki kişiler için verilen maaş, yaş ve iş tecrübesine göre koşullu olasılıklarını bulup bunu normalleştirme değerine bölüyoruz.
Benzer şekilde muhasebe kısmındaki kişilere ait beklenti de aşağıdaki gibi hesaplanabilir

$$beklenti(Muhasebe) = \frac{P(Muhasebe)\, p(maaş|Muhasebe)\, p(Yaş|Muhasebe)\, p(iş\ tecrübesi|Muhasebe)}{normalleştirme}$$

Buradaki fark, verilen bilgilerin muhasebe kısmındaki kişilere göre koşullu olasılığının alınmasıdır.
Normalleştirme değeri ise sistemimizde bulunan bütün ihtimalleri içeren ve beklenti değerlerini normalleştiren değerdir. Aşağıdaki şekilde hesaplanabilir:
normalleştirme = P(Muhasebe) p(maaş | Muhasebe) p (Yaş | Muhasebe) p (iş tecrübesi | Muhasebe) + {P(Yazılım) p(maaş | yazılım) p (Yaş | yazılım) p (iş tecrübesi | yazılım)
Şimdi iki sınıfımız olduğu ve aslında ikisi arasında seçim yapacağımız için, iki sınıfı da bölen ve pozitif çıkmasını beklediğimiz normalleştirme değerini göz ardı edebiliriz. Sırasıyla olasılıkları hesaplayalım:
P(Yazılım) = 8 kişiden 4'ü yazılım kısmında
P(Yazıılm) = 0.5
Benzer şekilde,
P(Muhasebe) = 0.5
olarak buluruz. Ardından koşullu olasılık değerlerini hesaplayacağız. Bu aşamada gauss dağılımını kullanmak isteyelim (farklı dağılımlar kullanılabilir ve başarı bu dağılıma göre değişebilir) ve dağılım fonksiyonunu hatırlayalım:

$$p(maaş|yazılım)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

varyans, maaş ve ortalama değerlerini yazacak olursak:

$$\frac{1}{\sqrt{2\pi 4.91E6^2}}\exp\left(\frac{-(3000-4250)^2}{24.91E6^2}\right)=\frac{1}{1.46E7}\exp\frac{(1250^2)}{4.82E13}=6.84E\text{-}8$$

Yani sonuç olarak 0.0000000687'e yakın bir değer bulunuyor. Bunun anlamı, verilen 3000 lira maaş ile çalışan kişinin Gauss dağılımında (dağılımın toplam alanının 1 olduğunu hatırlayınız), yazılım kısmında çalışan birisi olması ihtimalinin 6.87E-7 olduğudur. Daha doğru bir ifadeyle, bir kişinin yazılımcı olduğunu kabul edersek (verilen koşul) bu kişinin 3000 lira maaş alma durumunu hesaplamış olduk. Şimdi aynı maaş değerine sahip kişinin muhsabe kısmında çalışma olasılığını hesaplayalım.

$$\frac{1}{\sqrt{2\pi 5.83E6^2}}\exp\left(\frac{-(3000-2500)^2}{25.83E6^2}\right)=\frac{1}{4{,}91E7}\exp\frac{(-500^2)}{6{,}79E13}=8{,}11E\text{-}8$$

Yukarıdaki ikinci hesaplamanın sonucuna bakarak aslında bu maaşın, yazılım kısmındaki kişilere daha uygun olduğunu söyleyebiliriz.
Benzer hesaplamaları diğer koşullu olasılık durumları için de yaparsak aşağıdaki gibi bir tablo elde edebiliriz:

	Maaş	Yaş	Tecrübe
Muhasebe	6,84074E-08	0,002805118	0,011414151
Yazılım	8,11614E-08	0,019705849	0,046043474

Sonuçta naive bayes yöntemine göre bu verilen olasılıkların çarpımlarını alacağız:

$$beklenti\,(Yazılım) = \frac{P(Yazılım)\,p(maaş|yazılım)\,p(Yaş|yazılım)\,p(iş\ tecrübesi|yazılım)}{normalleştirme}$$

olduğunu hatırlayalım. Bu durumda beklenti(yazılım) aşağıdaki şekilde yazılabilir:

$$beklenti\,(Yazılım) = \frac{0.5\,x\,6{,}84\text{E-}8\,x\,0.0028\,x\,0{,}0114}{normalleştirme} = 1{,}9\text{E-}12$$

Burada normalleştirme değeri daha önce de belirtildiği üzere hesaba katılmamıştır.
Benzer şekilde muhasebe kısmına ait beklenti de aşağıdaki şekilde hesaplanabilir.

$$beklenti\,(Muhasebe) = \frac{0.5\,x\,8{,}11\text{E-}8\,x\,0.0019\,x\,0{,}046}{normalleştirme} = 3{,}68\text{E-}11$$

Görüldüğü üzere muhasebe beklentisi, yazılım beklentisinin yaklaşık 20 misli daha yüksek çıkmıştır. Demek ki naive bayes sınıflandırmasına göre bu kişinin muhasebe kısmında çalıştığını söyleyebiliriz, en azından beklentimiz bu yönde olur.

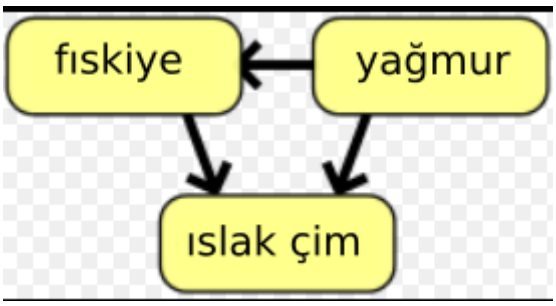
Bayes Ağı

Bir Bayes ağı, Bayes modeli ya da olasılıksal yönlü dönüşsüz çizge modeli bir olasılık çizgisel modelidir birbirleriyle koşulsal bağlılıklara sahip bir rassal değişkenler kümesini yönlü dönüşsüz çizge(YDÇ) şeklinde ifade eder.[1] Örneğin, bir Bayes ağı kullanılarak hastalıklar ve semptomları arasındaki olasılıksal koşul ilişkileri modellenebilir. Bu model kullanılarak, bir kişide görülen semptomlar verildiğinde bu kişinin bazı hastalıklara sahip olma olasılıkları hesaplanabilir.Bayes ağları, her düğümü bir rassal değişkeni ifade eden YDÇ'lerdir.[2] Gözlemlenebilir nicelikler, gizli değişkenler, bilinmeyen parametreler ya da hipotezler birer Bayes rassal değişkeni olabilirler. Birbirine herhangi bir şekilde bağlı olmayan düğümler birbirlerinden koşulsal bağımzsızdırlar. Her düğüm, girdi olarak ebeveyn düğümlerinin değerlerini alan ve çıktı olarak o düğümün ifade ettiği değişkenin alabileceği değerlerin olasılıklarını (duruma göre olasılık dağılımını) veren bir olasılık fonksiyonu ile ilişkilendirilmiştir.

Örneğin, eğer ebeveyn düğüm Bool değişkenini ifade ediyorsa olasılık fonksiyonu hücreli bir tablo ile gösterilebilir; ebeveyn değişkenlerinin alabileceği doğru ya da yanlış değerlerinin her biri için bir hücre.

Benzer fikirler yönsüz ve duruma göre dönüşlü çizgeler üzerinde uygulanabilir; böyleleri Markov Ağları olarak adlandırılır.

Bayes ağları üzerinde çıkarsama ve öğrenme yapan verimli algoritmalar vardır. Bir değişkenler dizisini modelleyen Bayes ağlarına dinamik bayes ağları denir.



Basit bir Bayes ağı. Yağmur yağma durumu fıskiyenin çalışma olasılığını etkiler. Çimlerin ıslak olma olasılığı ise hem fıskiyeye hem de yağmura bağlıdır.

Bayes Teoremi

Bayes Teoremi bir binom dağılımının parameteresinin olasılık dağılımının hesaplanmasını incelemekte olan, İngiliz Rahip Thomas Bayes tarafından bulunmuştur. Bu çalışma Bayes yaşamakta iken yayınlanmamış; ancak Bayes'in ölümünden sonra 1763'te yakın arkadaşı olan "Richard Price" tarafından yayına hazırlanıp bastırılmıştır.

Bayes'in çalışmalarından haberdar olmayan Fransız matematikci Pierre Simon Laplace aynı sonuçları aynen sırf kendi gayretiyle yeniden çıkartıp genişleterek 1774'te yazdığı bir makalede yayınlamıştır.Bir Amerikan istatistik profesörü (Stigler 1983), yaptığı bir araştırma sonucunda, Bayes Teoremi'nin, Bayes'ten bir süre önce Nicholas Saunderson tarafından bulunduğunu öne sürmüştür.

Naive Bayes sınıflandırıcısı Bayes teoreminin bağımsızlık önermesiyle basitleştirilmiş halidir. Bayes teoremi aşağıdaki denklemle ifade edilir;

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P(A|B) ; B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır.

P(B|A) ; A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır

P(A) ve P(B) ; A ve B olaylarının önsel olasılıklarıdır

Burada önsel olasılık Bayes teoreminine öznellik katar. Diğer bir ifadeyle 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.

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.

E-posta kutusuna gelen e-postaların spam olarak ayrıştırılması işlemi buna örnek verilebilir. Bu örnekte spam e-posta ve spam olmayan e-posta ayrıştırılacak iki sınıfı temsil eder. Elimizdeki spam ve spam olmayan e-postalardan yaralanarak gelecekte elimize ulaşacak e-postaların spam olup olmadığına karar verecek bir Algoritma da öğreticili makine öğrenmesine örnektir.

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.

Nicelik Kümesi

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. $\in R^L$ L-boyutlu nicelik vektörünü oluşturur.

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

$$P(S_i|x) \times p(x) = p(x|S_i) \times P(S_i)$$ 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

Örnek1

İki tabak dolusu bisküvi düşünülsün; tabak **#1** içinde 10 tane çikolatalı bisküvi ve 30 tane sade bisküvi bulunduğu kabul edilsin. Tabak **#2** içinde ise her iki tip bisküviden 20şer tane olduğu bilinsin. Evin küçük çocuğu bir tabağı rastgele seçip bu tabaktan rastgele bir bisküvi seçip alsın. Çocuğun bir tabağı diğerine ve bir tip bisküviyi diğerine tercih etmekte olduğuna dair elimizde hiçbir gösterge bulunmamaktadır. Çocuğun seçtiği bisküvinin sade olduğu görülsün. Çocuğun bu sade bisküviyi tabak **#1'**den seçmiş olmasının olasılığının ne olacağı problemi burada incelenmektedir.

Sezgi ile, tabak **#1** de sade bisküvi sayısının çikolatalı bisküvi sayısına göre daha fazla olduğunu göz önüne alınırsak incelenen olasılığın %50den daha fazla olacağı hemen algılanır. Bu soruya cevap Bayes teoremi kullanarak kesin olarak verilebilir.

Önce soruyu değiştirip Bayes teoremi uygulanabilecek şekle sokmak gerekmektedir: Çocuğun bir sade bisküvi seçmiş olduğu bilinmektedir; o halde bu koşulla birlikte tabak **#1** den seçim yapması olasılığı ne olacaktır?

Böylece Bayes teoremi formülüne uymak için *A* olayı çocuğun tabak **#1'**den seçim yapması; *B* olayı ise çocuğun bir sade bisküvi seçmesi olsun. İstenilen olasılık böylece Pr(*A*|*B*) olacaktır ve bunu hesaplamak için şu olasılıkların bulunması gerekir:

Pr(*A*) veya hiçbir diğer bilgi olmadan çocuğun tabak **#1**'den seçim yapması olasılığı;

İki tabak arasında tercih olmayıp seçimin eşit olasılığı olduğu kabul edilmektedir.Pr(*B*) veya hiçbir diğer bilgi olmadan çocuğun bir sade bisküvi seçmesi olasılığı: Diğer bir ifade ile, bu çocuğun her bir tabaktan bir sade bisküvi seçme olasılığıdır.

Bu olasılık, önce her iki tabaktan ayrı ayrı olarak seçilen bir tabaktan bir sade bisküvi seçme olasılığı ile bu tabağı seçme olasılığının birbirine çarpılması ve sonra bu iki çarpımın toplanması suretiyle elde edilir. Tabaklarda olan sade bisküvinin sayısının toplama orantısından bilinmektedir ki tabak **#1**'den bir sade bisküvi seçme olasılığı (30/40=) 0,75; tabak **#2**'den sade bisküvi seçme olasılığı (20/40=) 0,5 olur. Her iki tabaktan seçme olasılığı ise her tabak aynı şekilde uygulama gördüğü için 0,50 olur. Böylece bu problemin tümü için bir sade bisküvi seçme olasılığı $0.75\times0.5 + 0.5\times0.5 = 0.625$ olarak bulunur.

Pr(*B*|*A*), veya çocuğun tabak **#1**'den seçim yaptığı bilirken bir sade bisküvi seçmesi.: Bu 0,75 olarak bilinmektedir çünkü tabak **#1**'deki toplam 40 bisküviden 30'u sade bisküvidir.

Şimdi bu açıklanan tüm olasılık değerleri Bayes teoremi formülüne konulabilir:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.75 \times 0.5}{0.625} = 0.6$$

Böylece çocuğun sade bisküvi seçimi bilindiğine göre tabak **#1**'den alma olasılığı %60'dır ve sezgimize göre seçtiğimiz %50'den daha büyüktür.

Koşullu olasılıkları hesaplarken her bir bağımsız değişken için her mümkün sonucun ortaya çıkma sayısını veya her sonucun relatif çokluğunu gösteren basit bir tablo hazırlamak konuyu daha iyi anlamaya yardımcı olabilir. Bisküvi örneği için bu yöntemin kullanışını gösteren tablolar şöyle verilmiştir:

	Tabak #1	Tabak #2	Toplamlar
Çikolatalı	10	20	30
Sade	30	20	50
Toplam	40	40	80

	Tabak #1	Tabak #2	Toplamlar
Çikolatalı	0.125	0.250	0.375
Sade	0.375	0.250	0.625
Toplam	0.500	0.500	1.000

Örnek2

Yeni bir uyuşturucu madde testinin değerlendirilmesinde de Bayes teoremi yardımcı olabilir. Bu testin bir uyuşturucu madde sınamasında %99 kesin sonuç verdiğini kabul edelim; yani bu test bir uyuşturucu madde kullanan için %99 defa doğru olarak pozitif sonuç verecek ve uyuşturucu madde kullanmayan için negatif sonucu da %99 defa verecektir. Bu olasılıkların yüksek oluşu bu testin nispeten hatasız olduğu sonuç çıkarabilir; ancak Bayes teoremini kullanırsak bu düşüncemizin pek doğru olmadığı ortaya çıkacaktır.

Bir iş yeri çalışanlarını heroin kullanıp kullanmadıklarını sınamak istediğini ve çalışanlardan yüzde yarımının (%0,5) eroin kulladığını kabul edelim. Aradığımız netice, bir çalışan için bir test yapıldıktan ve pozitif sonuç alındıktan sonra bu çalışanın gerçekte eroin kullanıcısı olma olasılığının ne olduğunu bulmaktır.

"E" bir eroin kullanıcı olma olayı ve "N" eroin kullanıcı olmama olayı olarak gösterilsin. "+" test, pozitif sonuç göstermesi olayını belirtsin. Bu halde şunları bilmemiz gerekmektedir:

- Pr(*E*) veya başka bir bilgi bulunmadan çalışanın gerçekte eroin kullanıcı olması olasılığı

olsun; çalışanlardan yüzde yarımı (%0.5) eroin kullanıcısı olduğuna göre bu olasılık 0.005 olarak ifade edilinebilir.

- Pr(*N*) çalışanın gerçekte eroin kullanıcı **olmaması** olasılığı olsun. Bu 1-Pr(E) veya 0.995 olur.
- Pr(+|*E*) : Çalışanın bir eroin kullanıcısı olması bilindiği halde testin pozitif sonuç vermesinin olasılığı. Test %99 kesin sonuç verdiğine göre bu olasılık 0.99 olur.
- Pr(+) : Diğer bilgi olmadan testin pozitif sonuç vermesi olasılığıdır. Bu sonuç bulmak için önce eroin kullanması halinde testin pozitif verme olasılığı (ki bu %99 x %0,5 = %0,495) artı eroin kullanıcısı olmama halinde testin hatalı pozitif sonuç vermesi olasılığı (0,01 x 0,995= %0.995); yani 0.0149 %1.49 olur.

Bu bilgileri sıraladıktan sonra sınamada pozitif sonuç vermiş bir çalışanın gerçekte eroin kullanıcısı olması olasılığı şöyle bulunur:

$$
\begin{aligned}
P(E|+) &= \frac{P(+|E)P(E)}{P(+)} \\
&= \frac{P(+|E)P(E)}{P(+|E)P(E) + P(+|N)P(N)} \\
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\
&= 0.3322
\end{aligned}
$$

Test yüksek oranda "+" sonuç vermesine rağmen, genel olarak eroin kullanımı çok düşük olduğundan, pozitif testli bir çalışanın gerçekte eroin kullanıcısı olması olasılığı %33 olur. Test için konu alınan olay ne kadar nadir görülmekte ise testteki pozitif sonuçların hatalı pozitif olmaları olasılığı o kadar artacaktır. Bu uyuşturucu madde sınamasında testin tekrarlamanın neden gerektirdiğini de açıklamaktadır. Aynı örnek AIDS testi ve diğer testler için de geçerlidir.

Örnek3(Monty Hall Problemi)

Bir TV oyun programında üç tane (kırmızı, yeşil ve mavi boyalı) kapalı kapı gösterilmekte ve bu kapılardan birisinin arkasında bir armağan bulunmaktadır. Kırmızı kapıyı seçtiğimizi düşünelim; ama bu kapı program sunucunun bir faaliyet göstermesini bitirmeden açılmamaktadır. Program sunucusu hangi kapı arkasında armağan bulunduğunu bilmektedir; ama ona verilen direktife göre ne arkasında armağan bulunan kapıyı ne de seçtiğimiz kapıyı açabilir. Yeşil kapıyı açar ve arkasında bir armağan bulunmadığını gösterir ve şu soruyu yarışmacıya sorar: "İlk tercihiniz olan kırmızı kapı hakkında fikrinizi değiştirmek ister misiniz?"

İncelenecek sorun şudur: "Armağanın mavi veya kırmızı kapılar arkasında bulunma olasılıkları nedir?"

Yarışmanın ana sonuçları olan değişik renkli kapılar arkasında armağan bulunmasını şöyle ifade edelim: :**A_k**, **A_y** ve **A_m**. İlk olarak her bir kapı arkasında armağan bulunması birbirine eşit olasılık olduğu kabul edilir yani olur. Yine düşünelim kırmızı kapıyı yarışmacı seçmiş durumdadır. Sunucunun yeşil kapıyı açması olayına **B** olayı adını verelim. Arkasında armağan bulunan kapıyı bilmeseydi bu olay için olasılık %50 olacaktır.

- Eğer gerçekte armağan kırmızı kapı arkasinda ise, sunucu ya yeşil ya da mavi kapıyı acmakta serbest olacaktır. Bu halde $P(B|A_k) = 1/2$
- Eğer gerçekte armağan yeşil kapı arkasında ise, sunucu mavi kapıyı acacaktır. Yani $P(B|A_y) = 0$.
- Eğer gerçekte armağan mavi kapı arkasında ise, sunucu yeşil kapıyı acacaktır. Yani $P(B|A_m) = 1$.

Böylece

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(A_y|B) = \frac{P(B|A_y)P(A_y)}{P(B)} = \frac{0\frac{1}{3}}{\frac{1}{2}} = 0$$

$$P(A_m|B) = \frac{P(B|A_m)P(A_m)}{P(B)} = \frac{1\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Dikkatle incelenirse bunun P(B) değerine bağlı olduğu görülecektir. Bir an armağanın kırmızı kapı arkasında olmadığını farzedelim; o halde sunucunun yeşil kapıyı açma olasılığı çok *yüksek* olacaktır - diyelim %90. Bundan dolayı, eğer sunucu başka kapı açmaya zorlanmadıkça, yeşil kapıyı açmayı tercih edecektir.

Böylece, **B** olayı olasılığı 1/3 * 1 + 1/3 * 0 + 1/3 * 9/10 = 19/30 olur.

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} = \frac{\frac{9}{10}\frac{1}{3}}{\frac{19}{30}} = \frac{9}{19}$$

$$P(A_y|B) = \frac{P(B|A_y)P(A_y)}{P(B)} = \frac{0\frac{1}{3}}{\frac{19}{30}} = 0$$

$$P(A_m|B) = \frac{P(B|A_m)P(A_m)}{P(B)} = \frac{1\frac{1}{3}}{\frac{19}{30}} = \frac{10}{19}$$

Bu nedenle sunucunun yeşil kapıyı açması bize çok az bilgi vermektedir - zaten bu seçimi yapmak zorundadır. **Pr(A_m)** olasılığı 1/2'nin çok az üstündedir.

Buna karşılık, armağanın kırmızı kapı arkasında olduğunu farzedersek; o halde sunucunun yeşil kapı açma olasılığı çok *küçük* olacaktır - diyelim %10. Bu demektir ki özellikle zorlanmadıkça sunucu nerede ise hiçbir halde yeşil kapıyı açmayacaktır.

O halde **B** olasigi 1/3 * 1 + 1/3 * 0 + 1/3 * 1/10 = 11/30 olur

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} = \frac{\frac{1}{10}\frac{1}{3}}{\frac{11}{30}} = \frac{1}{11}$$

$$P(A_y|B) = \frac{P(B|A_y)P(A_y)}{P(B)} = \frac{0\frac{1}{3}}{\frac{11}{30}} = 0$$

$$P(A_m|B) = \frac{P(B|A_m)P(A_m)}{P(B)} = \frac{1\frac{1}{3}}{\frac{11}{30}} = \frac{10}{11}$$

Bu halde, gerçekte sunucunun yeşil kapıyı açması bize çok önemli bilgi vermektedir. Armağan nerede ise hiç şüphesiz olarak mavi kapı arkasında bulunmaktadır. Eğer mavi kapı arkasında değilse, sunucu çok muhtemelen mavi kapıyı açacaktı.

Bayes Karar Teoremi

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

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

diğer bir ifadeyle eğer

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

Naive Bayes Sınıflandırma

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 edilinebilir.

Naive Bayes Uygulama Alanları

Naive Bayes sınıflandırıcısı genel olarak

veri madenciliğinde, biyomedikal mühendisliği alanında, hastalıkların ya da anormalliklerin tıbbi tanımlanmasında (otomatik olarak mühendislik ürünü tıbbi cihazlar tarafından tanı konulması), elektrokardiyografi (EKG) grafiğinin sınıflandırılmasında, elektroensefelografi (EEG) grafiklerinin ayrıştırılmasında, genetik araştırmalarında, yığın mesaj tanımlanmasında, metin ayrıştırılmasında, ürün sınıflandırma ve diğer bazı alanlarda kullanılır.