- Olasılık ve Koşullu Olasılık
- Bayes Teoremi
- Naïve Bayes Sınıflayıcı

1

Olasılık

Olasılık ifadesinin birçok kullanım şekli vardır. Rasgele bir A olayının herhangi bir olaydan bağımsız olarak gerçekleşme ihtimalini ifade etmek için P(A) notasyonu kullanılır. A olayının olasılığı olarak bilinen bu ifade "önsel" (prior), "koşulsuz" (unconditional) veya "marjinal" (marginal) olasılık isimleriyle kullanılabilir.

2

Örnek

Bir para atma olayının iki kez tekrarlanması durumunda ardarda iki defa da tura gelme ihtimalini bulalım.

$$P(A \cap B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

3

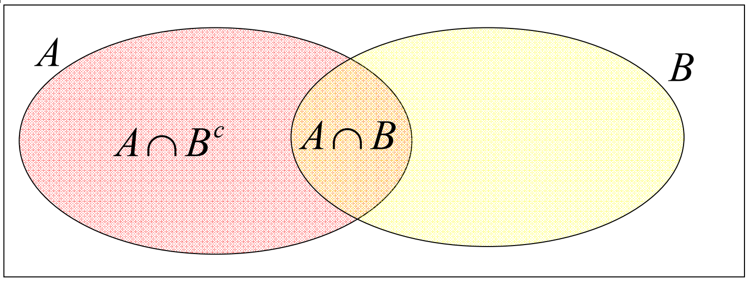
Koşullu Olasılık

Rasgele bir A olayının, farklı bir rasgele B olayına bağlı gerçekleşmesi ihtimalini ifade etmek için önsel olasılıklar yeterli olmaz. Bu yüzden "koşullu" (conditional) veya "sonsal" (posterior) olasılık olarak isimlendirilen P(A|B) notasyonu kullanılır. Bilinen bir B olayına göre A olayının koşullu olasılığı şöyle gösterilir.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

4

Koşullu Olasılık



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

5

Örnek

Elimizde bulunan iki kavanozdan birincisinde 3 mavi ve 4 sarı, ikincisinde 5 mavi 2 sarı top olsun. Bu durumda rasgele seçilen bir kavanozdan mavi top çekme olasılığı nedir?

A: Birinci kavanozu seçme durumu

B: İkinci kavanozu seçme durumu

C: Mavi top seçme durumu

$$P(C) = P(A)P(C \mid A) + P(B)P(C \mid B) = \frac{1}{2}\frac{3}{7} + \frac{1}{2}\frac{5}{7} = \frac{4}{7}$$

6

Bayes Teoremi

Birbirinden bağımsız ve rasgele iki olayın (A ve B) birbiri ardı sıra gerçekleştiği durumlarda bu iki olaydan birinin gerçekleşmesi durumunda ikinci olayın gerçekleşme olasılığı P(A,B) veya P(B,A) yada $P(A \cap B)$ ifadesi ile gösterilebilir. Değişme özelliği sayesinde aşağıdaki çarpım kuralı iki farklı ifade ile yazılabilir.

$$P(A \cap B) = P(A \mid B)P(B)$$
$$P(A \cap B) = P(B \mid A)P(A)$$

7

Bayes Teoremi

Bayes teoremi (kuralı veya kanunu) stokastik (rassal) bir sürece bağlı olarak ortaya çıkan rasgele bir A olayı ile diğer bir rasgele B olayı için koşullu olasılıklar ve marjinal olasılıklar arasındaki ilişkiyi tanımlar. Bu ilişkiyi ilk kez Thomas Bayes (1702-1761) ortaya atmış ve aşağıdaki eşitliği önermiştir.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

8

Örnek

İki sınıflı (kanser ve kanser değil) bir tıbbi teşhis problemini ele alalım. Tüm popülasyonun %0.8'nin kanser olduğunu varsayalım. Ayrıca kişilere + (pozitif) ve - (negatif) olmak üzere iki sonucu olan bir laboratuvar testi uygulanmış olsun. Test, hastalığın var olduğu durumların %98'inde +, olmadığı durumların ise %97'sinde - sonuçlar üretiyorsa;

$$P(kanser) = 0.008 \qquad P(\neg kanser) = 0.992$$
$$P(+ \mid kanser) = 0.98 \qquad P(+ \mid \neg kanser) = 0.03$$
$$P(- \mid kanser) = 0.02 \qquad P(- \mid \neg kanser) = 0.97$$

9

Örnek

Bu olasılıklara göre labortuvar sonucu + olan ve kanser şüphesiyle gelen bir kişinin kanser olup olmadığı Bayes Teoremine göre hesaplanırsa aşağıdaki sonuçlar elde edilir;

$$P(kanser \mid +) = \frac{0.98 \times 0.008}{P(+) = 0.0376} = 0.2085$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$$P(\neg kanser \mid +) = \frac{0.03 \times 0.992}{P(+) = 0.0376} = 0.7915$$

$P(kanser \mid +) < P(\neg kanser \mid +)$ olduğu için <u>kanser değildir</u> !

10

Naïve Bayes Sınıflayıcı

Bayes Teoremi sınıflandırma amaçlı kullanılırken en yüksek olasılıklı durum hedef sınıf olarak seçilir.

$$S_{hedef} = \arg\max_{s_j \in S} P(s_j \mid v)$$

Fakat girdi vektörlerinin (*v*) birden çok olduğu durumlarda Bayes formülü farklı bir forma dönüşür. Birçok özelliğin kesişimi görünümündeki veri örneği için hedef sınıf tahmininde tüm özellikler için koşullu olasılıkların çarpımı hesaplanmalıdır.

$$P(v_1, v_2, \ldots, v_n \mid s_j) = \prod_{i=1}^{n} P(v_i \mid s_j)$$

11

Naïve Bayes Sınıflayıcı

Naïve Bayes Sınıflayıcı ile Bayes Teoremi hesaplarında dikkat edilmesi gereken en önemli fark sınıflayıcıların olasılık değerinden ziyade hedef sınıfı bulmaya odaklanmasıdır. Bu yüzden paydada bulunan değer, tüm hedef sınıflara ait olasılık hesaplarında ortak olduğundan ihmal edilebilir. O zaman hedef sınıf bulurken dikkat edeceğimiz formül;

$$S_{hedef} = \arg\max_{s_j \in S} \left(P(s_j) \prod_{i=1}^{n} P(v_i \mid s_j) \right)$$

12

Naïve Bayes Sınıflayıcı

Bu yüzden Bayes Teoreminde bulunan sınıf olasılıkları toplamı 1 olmak zorunda iken Naïve Bayes Sınıflayıcı ile bulunan değerlerin toplamı (ihmal edilen paydadan dolayı) 1 olamaz. Buna göre Bayes Teoremi ile bir durumun olası sonuçlarının olasılıkları bulunurken Naïve Bayes ile normalizasyonsuz olasılıklar üzerinden sınıflandırma yapılabilir.

Örnek

v_1	v_2	v_3	v_4	Sınıf
Yes	No	No	Yes	B
Yes	No	No	No	B
No	Yes	Yes	No	M
No	No	Yes	Yes	M
Yes	No	No	Yes	B
Yes	No	No	No	M
Yes	Yes	Yes	No	M
Yes	Yes	No	Yes	M
No	No	No	Yes	B
No	No	Yes	No	M

Yandaki veri kümesi için sırasıyla gelen <Yes, No, Yes, Yes> örneğinin sınıfını bulalım.

Örnek

$$S_{hedef} = \arg\max_{s_j \in S} \left(P(s_j) \prod_{i=1}^{n} P(v_i \mid s_j) \right)$$

V1	V2	V3	V4	Sınıf
Yes	No	Yes	Yes	B
Yes	No	No	No	B
Yes	No	No	Yes	B
No	No	No	Yes	B
No	Yes	Yes	No	M
No	No	Yes	Yes	M
Yes	No	No	No	M
Yes	Yes	Yes	No	M
Yes	Yes	No	Yes	M
No	No	Yes	No	M

P(S | Yes, No, Yes, Yes) = ?

$$P(S = b \mid v_1, v_2, v_3, v_4) = \frac{2}{5}\left(\frac{3}{4} 1 \frac{1}{4}\frac{3}{4}\right) = \frac{9}{160}$$

$$P(S = m \mid v_1, v_2, v_3, v_4) = \frac{3}{5}\left(\frac{1}{2}\frac{1}{2}\frac{2}{3}\frac{1}{3}\right) = \frac{1}{30}$$

Sonuçlara göre hedef sınıf "**b**" bulunur.

15

Nümerik Değerler

Bayes Teoremi yalnızca kategorik veri özelliklerinde kullanılabilmektedir. Nümerik değerlere sahip özelliklere uygulayabilmek için ilgili özellik uzayında örneklerin Gauss (normal) Dağılımına sahip olduğu varsayılır. Aranan olasılık değeri, özellik vektörünün ortalaması μ ve standart sapması σ değerine bağlı aşağıdaki dağılım formülü ile hesaplanır.

$$P(v) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2}$$

16

Örnek

V_1	V_2	V_3	V_4	Yaş	Cinsiyet
Evet	Hayır	Evet	Hayır	38	k
Evet	Evet	Evet	Hayır	40	k
Evet	Evet	Evet	Hayır	41	k
Hayır	Hayır	Hayır	Hayır	55	k
Hayır	Evet	Hayır	Hayır	27	e
Evet	Evet	Evet	Evet	30	e
Evet	Hayır	Evet	Evet	35	e
Hayır	Hayır	Hayır	Hayır	42	e
Evet	Hayır	Hayır	Hayır	43	e
Evet	Hayır	Hayır	Hayır	45	e

V_1= Evet, V_2 = Evet,
V_3= Hayır, V_4= Hayır,
Yaş= 45 için
Cinsiyet = ?

Yaş özelliği nümeriktir. Normal dağılıma sahip olduğu varsayılarak olasılıkları hesaplanır.

17

Örnek

Yaş	Cinsiyet
38	**k**
40	**k**
41	**k**
55	**k**
27	e
30	e
35	e
42	e
43	e
45	e

P(Yaş | Cinsiyet=k) için μ=43.5 ve σ=7.77
P(Yaş | Cinsiyet=e) için μ=37 ve σ=7.46

Normal dağılıma göre; $P(v) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2}$

P(Yaş=45 | Cinsiyet=k) = 0.0504 ~ 5/100
P(Yaş=45 | Cinsiyet=e) = 0.0301 ~ 3/100

bulunur.

18

Örnek

v_1	v_2	v_3	v_4	Yaş	c
Evet	Hayır	Evet	Hayır	38	k
Evet	Evet	Evet	Hayır	40	k
Evet	Evet	Evet	Hayır	41	k
Hayır	Hayır	Hayır	Hayır	55	k
Hayır	Evet	Hayır	Hayır	27	e
Evet	Evet	Evet	Evet	30	e
Evet	Hayır	Evet	Evet	35	e
Hayır	Hayır	Hayır	Hayır	42	e
Evet	Hayır	Hayır	Hayır	43	e
Evet	Hayır	Hayır	Hayır	45	e

$P(C\,|$ Evet,Evet,Hayır,Hayır,45) = ?

$$P(C=e\,|\,v_1,v_2,v_3,v_4,Yaş)=\frac{3}{5}\left(\frac{2}{3}\frac{1}{3}\frac{2}{3}\frac{2}{3}\frac{3}{100}\right)\cong 0.0018$$

$$P(C=k\,|\,v_1,v_2,v_3,v_4,Yaş)=\frac{2}{5}\left(\frac{3}{4}\frac{1}{2}\frac{1}{4}1\frac{5}{100}\right)\cong 0.0019$$

Sonuçlara göre hedef sınıf "**k**" bulunur.

19

Sıfır Olasılık Sorunu

Bayes formülü, özelliklere ait koşullu olasılıkların çarpımı şeklinde düzenlenirken bir özellik vektöründe çözümü aranan değere rastlanmaması durumunda genel sonucu etkileyecek bir sıfır olasılık üretilir. Diğer özellik olasılıklarının büyük olduğu durumlarda sıfır olasılıklı özellik genel sonucu etkiler. Bunu engellemek için olasılıklar sıfır olamayacak şekilde tekrar organize edilir.

20

MATLAB Uygulaması

```
>edit NBayes_ornek.m
```

Makine öğrenmesinde popüler IRIS dataseti üzerinde Naive Bayes algoritması deneyi yapılmaktadır. Matlab grafik komutları ve `NaiveBayes.fit` komutu üzerinde değişiklikler yapılarak kodlar irdelenmelidir.

21

ÖDEV

Aşağıdaki yöntemleri MATLAB'de hazırlayınız.

- Lojistik Regresyon
- EM (Expectation Maximization)

22