VERİ SIKIŞTIRMA

Bilgi Teorisi

Prof.Dr. Banu DİRİ

Bilgi Teorisi

Tabiata baktığımızda bir olayın gerçekleşme ihtimali ne kadar düşük ise o olayın taşıdığı bilgi o kadar fazladır. Tersini de aynı şekilde düşünebiliriz; Bir olayın gerçekleşme ihtimali ne kadar fazla ise o olay fazla bilgi içermez.

Örnek : Sıfır derecede karın yağması çok olasıdır ve fazla bilgi taşımaz. Ancak, 20 derecede kar yağarsa olağanüstü bir durum vardır ve bu olay fazla bilgi taşımaktadır.

$I(A) = log_x \frac{1}{P(A)}$

$I(A) = - \log_x P(A)$

P(A) : A olayının gerçekleşme ihtimali
I(A) : A olayının taşıdığı bilgi

Sıkıştırma tekniklerinde de performansı arttırmak için bu uygulanır. Sıkıştırılacak kaynakta bir sembol ne kadar fazla ise, sıkıştırma sırasında kodlama yapılırken bu sembol daha az bitle ifade edilmektedir. Sıkıştırılacak kaynakta bir sembol ne kadar nadir yer alıyorsa, kodlamada bu sembol daha fazla bit ile ifade edilmektedir.

Formülümüzde $\log_x$'i $\log_2$ olarak yani 2 tabanında kullanırız.

I(A) = $log_2 \frac{1}{P(A)}$ olur ve burada I(A)'yı kodlama sonucunda oluşan bit uzunluğu olarak ifade ederiz.

Örnek : Hatasız bir parayı attığımızda yazı veya tura gelme olasılığı her ikisi içinde $\frac{1}{2}$'dir.

Yani burada **P(Y)=P(T)=1/2** ve **I(Y)=I(T)=1** olarak çıkar.

Örnek : Para sahte ise ve P(Y)= 1/8 ve P(T)=7/8 olasılığı varsa kod uzunlukları I(Y)=3 bits ve I(T)=0,193 bits olacaktır.

Tek bir para atma olasılığı yerine bağımsız bir çok olayı ele alabiliriz. Bu bağımsız olaylardan bir set oluşturup, bu seti S ile ifade edersek: **S=U A_i** diye gösterebiliriz.

S örnek seti içindeki ortalama bilgiyi (self information) elde etmeye çalışırsak $\mathbf{H} = \sum \boldsymbol{P(Ai)I(Ai)}$ formülünden yararlanırız. Her bir olayın gerçekleşme ihtimalinin o olayın taşıdığı bilgiyle çarpımının genel toplamı bize bu olay setinin ortalama bilgisini verir yani **Entropy 'sini** verir.
Bu olayları kodlamaya geçtiğimizde entropy bize olması gereken ideal ortalama binary sembol uzunluğunu verecektir.

Shannon göstermiştir ki en iyi kayıpsız sıkıştırmada kodlanan sembollerin ortalama bit uzunluğu teoride bulduğumuz kaynağın entropy'sine eşit olmalıdır. Ancak, bu ideal sıkıştırma da mümkün olmamaktadır.

$$\mathbf{H} = \sum \boldsymbol{P(Ai)I(Ai)}$$

Kaynak alfabesi a_1, a_2, a_3 ve a_4'ten oluşan bir set olup sembollerin olasılıkları a_1 =0,49 a_2 =0,25 a_3 =0,25 ve a_4 =0,01 olsun.

$$\mathbf{H} = -\sum \boldsymbol{P(Ai)log_2P(Ai)}$$

Bu setin entropy H= - (0,49 $\log_2$0,49 + 0,25 $\log_2$0,25 + 0,25 $\log_2$0,25 + 0,01 $\log_2$0,01) = 1,57

Bu set ortalama 1,57 bits ile kodlanabilir.

Şayet

a_1 =00 a_2 =01 a_3 =10 ve a_4 =11 ile gösterecek olursak kodlanan sembollerin ortalama bit uzunluğu 2 olacaktır. **S=2**

Buradan artıklığı (redundancy) hesaplarsak R = |H – S|= |1,57 – 2|= 0,43 elde edilir.

Not: 2 tabanında logaritma nasıl alınır?

log_2 x = q → x=2^q

ln(x)=ln(2^q)

ln(x) = q ln(2)

q=$\frac{\ln(x)}{\ln(2)}$ → log_2 x = $\frac{log_{10}x}{log_{10}2}$

Örnek

1 2 3 2 3 4 5 4 5 6 7 8 9 8 9 10 şeklinde bir sayı dizisi verilmiş olsun. Bu sayı dizisini kadlayacak olursak ortalama bit uzunluğu ne olur?

$$\mathbf{H} = -\sum \boldsymbol{P(Ai) log_2 P(Ai)}$$

P(1) = P(6) = P(7) = P(10) = 1/16

P(2) = P(3) = P(4) = P(5) = P(8) = P(9) = 2/16

$\mathbf{H} = -\sum_{i=1}^{10} \boldsymbol{P_i log_2 P_i}$ = 3,25 bits

Örnek

1 2 3 2 3 4 5 4 5 6 7 8 9 8 9 10 şeklinde verilen sayı dizisinden bir «resiudal sequence» oluşturduğumuzda, bu sayı dizisini ortalama kaç bit uzunluğunda kodlayabiliriz ?

1 1 1 -1 1 1 1 -1 1 1 1 1 1 -1 1 1

P(1) = 13/16 P(-1) = 3/16

H= 0,70 bits

Alıcı bu diziden orijinal veriye dönüş yapabilir mi?

Sekansın modelinin çıkarılması gerekir ?

x_n : orijinal dizinin n.elemanı

r_n : residual dizinin n.elemanı

$x_n = x_{n-1} + r_n$

Bu model statik model olarak adlandırılır. Çünkü parametreler **n** değiştikçe, değişmez.

Verinin karakteristiğine bağlı olarak parametreler **n** ile değişiyorsa dinamik model adını alır.

Örnek

1 2 1 2 3 3 3 3 1 2 3 3 3 3 1 2 3 3 1 2 şeklinde verilen sayı dizisini ortalama kaç bit uzunluğunda kodlayabiliriz ?

P(1) = P(2) = $\frac{1}{4}$ P(3) = 1/2

H= 1,5 bits/ sembol

Sekans 20 sembolden oluştuğu için, kodlamak için gerekli bit sayısı 30

Aynı sekansta blokları ikişer sembolden oluşacak şekilde alırsak sonuç değişir mi?

P(12) = P(33) = $\frac{1}{4}$

H= 1 bit/ sembol

Sekans 10 sembolden oluştuğu için, kodlamak için gerekli bit sayısı 10

Modeller

1- Fiziksel Model
2- Olasılık Model
3- Markov Model

Verinin modelini iyi veya doğru bir şekilde oluşturmak, kaynağın entropy'sini tahmin etmek ve etkili bir sıkıştırma algoritması seçmek için önemlidir.

Fiziksel Model

İşlenecek verinin fiziki şartları hakkında yeterli bilgiye sahip olmak modelin doğru kurulması için önemlid

- Ses ile ilgili bir uygulamada sesin alındığı ortamın şartları
- Kullanılan elektriğin faturalandırılması

Olasılık Model

En basit istatistiksel model olup, kaynak tarafından üretilen her sembolün birbirinden bağımsız ve eşit olasılıkla üretildiği durumdur. Bu model **«ignorance model»** olarak adlandırılır ve kaynak hakkında hiçbir bilgi yoksa kullanılır.

Eşit olasılık yerine her sembolün meydana gelme sıklığı göz önüne alınarak bir olasılığın hesaplandığı modelde de **«probability model»** olarak adlandırılır. İstatistiksel veri sıkıştırma yöntemleri bu modeli kullanır.

Alfabe = {a1, a2, …am}

P = { P(a1), P(a2), …, P(m)}

Markov Model

Sembollerin olasılıkları birbirlerine bağımlı olarak hesaplanır (adını A.A. Markov Rus matematikçiden alır).

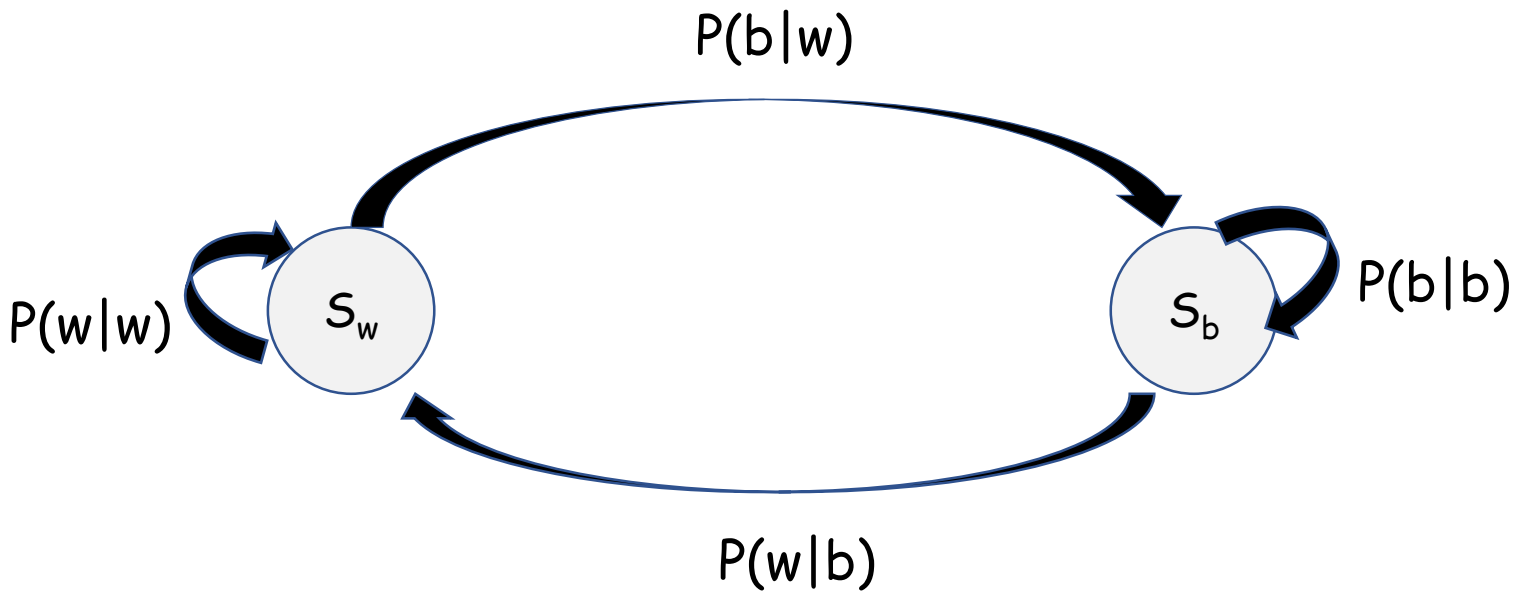
Kayıpsız sıkıştırma yöntemleri için kullanılır.
«Discrete Time Markov Chain» olarak da bilinir.

{xn} gözlemlenen bir sekans olsun. Bu sekans **«kth-order Markov»** model olarak

$P(x_n|x_{n-1}, \ldots, x_{n-k}) \approx P(x_n|x_{n-1},\ldots,x_{n-k}, \ldots)$ olarak gösterilir.

Geçmişteki **k** sembol bilgisi, işlemin geçmişteki tüm bilgisine yaklaşık olarak eşittir.

Elimizde binary (B&W) bir görüntü olduğunu düşünelim. Bu görüntüyü Markov Modele göre kodlamak isteyelim ve olasılıklarımızı belirleyelim.



- Siyah ve Beyaz 2 adet state
- Olasılıkları da P(Sw) ve P(Sb)

$$\mathbf{H} = \sum_{i=1}^{M} \boldsymbol{P}_{i(\boldsymbol{S}_i)} \boldsymbol{H}(\boldsymbol{Si})$$

$$H(S_b) = -P\left(\frac{w}{b}\right) log_2 P\left(\frac{w}{b}\right) - P\left(\frac{b}{b}\right) log_2 P\left(\frac{b}{b}\right)$$

$$H(S_w) = -P\left(\frac{b}{w}\right) log_2 P\left(\frac{b}{w}\right) - P\left(\frac{w}{w}\right) log_2 P\left(\frac{w}{w}\right)$$

P(w/w) = 1 – P(b/w)
P(b/b) = 1 - P(w/b)

Örnek

P(Sw) = 0,8 P(Sb) = 0,2 P(w/b) = 0,3 P(b/w) = 0,01 olarak verilmiş olsun. Hem Markov Model ile hem de Olasılıksal Model ile sistemin Entropy'sini hesaplayınız.

Olasılıksal Model

$H = -(0{,}8 \log_2 0{,}8 + 0{,}2 \log_2 0{,}2) =$ **0,722 bits**

Markov Model

$H(Sb) = -(0{,}3 \log_2 0{,}3 + 0{,}7 \log_2 0{,}7) = 0{,}881$ bits
$H(Sw) = -(0{,}01 \log_2 0{,}01 + 0{,}99 \log_2 0{,}99) = 0{,}081$ bits

$H = 0{,}2 * 0{,}881 + 0{,}8 * 0{,}081 =$ **0,241 bits**