Shannon Bilgi Kuramı

Çağatay Yücel

Yaşar Üniversitesi
Bilgisayar Mühendisliği

08.03.2012 / Shannon Bilgi Kuramı

İçerik

Bilgi Kuramı ve Kavramı

- Bilgi kavramı ile uğraşan, bilgiyi ölçmeye çalışan ve kullanım alanlarını inceleyen bilim alanına *bilgi kuramı* denilebilir.
- Genel bakış açısıyla, bilgi kavramı 3'e ayrılır:
  - Sözdizimsel Bilgi: Mesajları oluşturan semboller ve bu sembollerin arasındaki ilişkileri ile alakalıdır.
    Örn: Gramer kurallarının tamamı, Büyük ünlü uyumu.
  - Anlamsal Bilgi: Mesajların ne anlama geldiği ile ilgilenir.
  - İşlevsel Bilgi: Mesajların nasıl kullanıldığı ve etkileri üzerinedir.

Shannon Bilgi Kuramı

- Bilginin sözdizimsel olarak incelendiği ve modellendiği kuramdır.
- İstatistiksel Bilgi Kuramı veya İletişim Kuramı olarak da anılır.
- Claude E. Shannon’ın 1948’de yayınladığı “A Mathematical Theory of Communication” makalesi ile temelleri atılmıştır.
- Bu kuramın oluşumundaki tarihsel sürece kısaca bir göz atalım.

Harry Nyquist, (1924)

- “Certain Factors Affecting Telegraph Speed (1924)” adlı makalesi ile telgraf üzerinden maksimum hızda ve bozulma olmadan mesaj göndermenin yolunu açıkladı.
- Kullanılan devreye göre iletişim kanalının limitlerinin var olduğunu açıkladı.

Ralph Hartley, (1928)

- “Transmission of Information (1928)” adlı makalesi ile ilk olarak bilgi ölçülmeye çalışılmıştır.
- s karakterlik bir alfabemiz olsun ve n uzunluğundaki bir mesajın içerisindeki bilgi miktarı

$$H_H(s^n) = log(s^n) = n \cdot log(s)$$

  olarak tanımlanmıştır.
- Bu durumda:

$$H_H(s^n) = n \cdot H_H(s^1)$$

Claude Elwood Shannon, (1948)

- “A Mathematical Theory of Communication (1948)”
- Nyquist ve Hartley’nin kuramlarını genişleterek günümüz bilgi kuramını oluşturmuştur.
- Bir mesajın içerisindeki belirsizliği olasılık kavramı ile ilişkilendirerek mesajın içerisindeki bilgi miktarını tanımlamıştır.
- Tüm karakterlerin oluşma olasılıklarının eşit olduğu varsayıldığında Hartley’nin ölçüsüne dönülmektedir. Bu sebeple Hartley’nin ölçüsü Shannon ölçüsünün özel bir durumudur denilebilir.

Temel Olasılık Bilgileri

- $X = x_1, x_2, \ldots, x_n$ - deneyini ele alalım. x_i olası sonuçlar olsun.
- $P = p_1, p_2, \ldots, p_n$ - örnek uzaydaki olayların ortaya çıkma olasılıkları olsun.

Tanım

Eğer aşağıdaki koşullar sağlanıyor ise $P = p_1, p_2, \ldots, p_n$ bir olasılık dağılımıdır :

1. $p_i \geq 0$
2. $\sum p_i = 1$

Ortak Olasılık

- $Y = y_1, y_2, \ldots, y_m$
- $Q = q_1, q_2, \ldots, q_m$
- $(X, Y) = (x_1, y_1), \ldots, (x_n, y_m)$ olayını tek bir örnek uzay olarak tanımlayabiliriz.

Tanım

(x_i, y_i) sonucunun ortaya çıkma olasılığına *ortak olasılık* denir ve $r(x_i, y_i)$ ile gösterilir.

Marjinal Olasılık

- Ortak olasılık biliniyor ise, marjinal olasılıklar hesaplanabilir.

Tanım

r_{ij} ortak olasılık dağılışı olsun. Bu durumda marjinal dağılışlar, p_i, q_j:

$$p_i = \sum_{j=1}^{m} r_{ij}$$

$$q_j = \sum_{i=1}^{n} r_{ij}$$

Koşullu - Marjinal- Ortak olasılık

- X olayının, bir diğer Y olayına koşullu olasılığı (veya Y biliniyorken X'in olasılığı), $P(X|Y)$ olarak tanımlanır;
- Koşullu olasılık, ortak olasılık ve marjinal olasılıklar arasında aşağıdaki gibi bir ilişki vardır.

$$r(x_i, y_j) = q(y_j) \cdot p(x_i|y_j) = p(x_i) \cdot q(y_j|x_i)$$

Birbirinden Bağımsız Olaylar

- X, Y olaylarının birbirinden bağımsız olması durumunda:

  $$p(x_i|y_j) = p(x_i)$$

  $$q(y_j|x_i) = q(y_j)$$

  $$r(x_i, y_j) = p(x_i) \cdot q(y_j)$$

  formülleri geçerlidir.

Mesajdaki Belirsizlik ve Bilgi Miktarı

- Bir mesajdaki belirsizliğin o mesajdaki bilgi miktarına eşit olduğunu aşağıdaki örnekle açıklayalım.
- Bir torbaya her renkten toptan bir adet attığımızda, herhangi bir rengin gelme olasılığı $1/n$ olur.
- Çekilen bir topun hangi renk olacağı belirsizdir ve çekilişin ardıdan sonuç bize bir miktar bilgi vermektedir.
- Sadece tek bir renkten n adet top atılsa idi, çekilen topun rengini zaten biliyor olduğumuzdan bu çekilişin ardından herhangi bir şey öğrenmemiş olurduk.
- Böylelikle belirsizliğin bilgi miktarına eşit olduğunu düşünebiliriz.

Bilgi Miktarı

Tanım

X bir olasılık deneyi olsun. $p(x_i)$ ise x_i sonucunun ortaya çıkması durumunun olasılığı olarak verildiğinde, bu deneydeki ortalama bilgi miktarı:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot log(p(x_i)) = -\sum_{i=1}^{n} p_i log(p_i)$$

Bilgi Miktarı

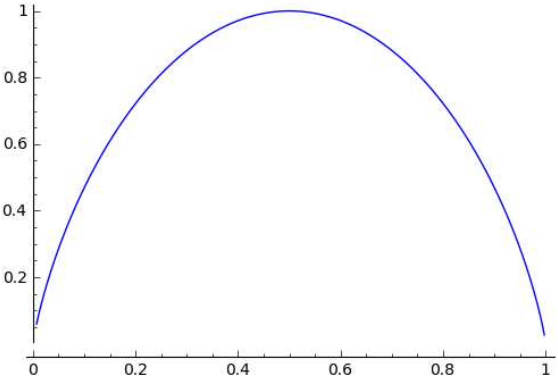
- Logaritmanın tabanı 2 olarak seçildiğinde, bilgi miktarının birimi bit olmaktadır.
- $log(1) = 0$ olduğundan kesin olarak gerçekleşeceğini bildiğimiz bir olay bize 0 bitlik bilgi vermektedir.
- Tüm olasılıklar eşit olduğunda ise bilgi miktarı maksimum değerini almaktadır.
- Denklemdeki - işareti bilgi miktarı için $H(X) \geq 0$ koşulunu yerine getirir.

Aksiyomlar

- $H(X)$ süreklidir.
- $H(X)$ simetriktir.
- $H(X)$ için toplama özelliği vardır.
- $H(X)$, X olayı uniform dağılış gösterdiğinde maksimum değerini alır.

Süreklilik

- $H(X)$, p'de süreklidir.
- $\mathbf{p} = (p, 1 - p)$



Simetri

- Olasılıkların sıralanışının bir önemi yoktur.
- Örneğin:
  $X =$ (Kırmızı top çekilir, Mavi top çekilir)
  $P_x = (0.8, 0.2)$ için
  $H(X) = -0.8 \cdot log(0.8) - 0.2 \cdot log(0.2) = 0.72$ bit.

  $Y =$ (Kar yağar, Kar yağmaz)
  $P_y = (0.2, 0.8)$ için
  $H(Y) = -0.2 \cdot log(0.2) - 0.8 \cdot log(0.8) = 0.72$ bit.

Toplama Özelliğinin Sağlanması

- Birbirinden bağımsız iki olay X, Y için:

$$H(X,Y) = H(X) + H(Y)$$

- Örneğin:
  İki zar atıldığında, birbirinden bağımsız olaylar söz konusu olduğu için, beraber atılmaları ile birinin diğerinden sonra atılmasının bir farkı yoktur.
  $H(X) = H(Y)$
  $H(p_1, q_1, \ldots, p_6, q_1, \ldots, p_6, q_6) = H(p_1, p_2, \ldots, p_6) + H(q_1, q_2, \ldots, q_6)$

Entropinin Maksimum ve Minimum Olma Durumları

Teorem

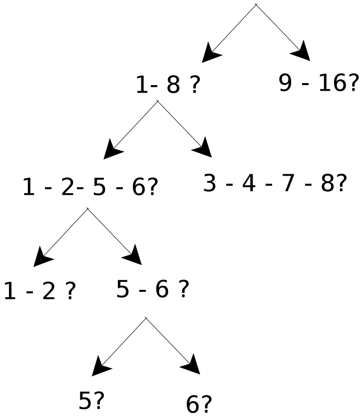
$X = (x_1, \ldots, x_n)$ örnek uzayımız olsun. $P = (p_1, \ldots, p_n)$ ise X örnek uzayına karşılık gelen olasılık dağılışımız olsun.

1. *$H(X) \leq log(n)$, eşitlik durumu $= p_i = 1/n$*
2. *$H(X) \geq 0$, eşitlik durumu $= p_k = 1$ değerini sağlayan bir k değeri mevcuttur.*

İspat için [1]

Örnek: Taralı alanı tahmin etmek

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16



- Tüm karelerin taralı olma olasılıklarının eşit olduğunu varsayarsak:
- Minimum 4 soru sorarak taralı alanı bulabiliriz.
- $H(X) = -\sum_{i=1}^{16} \frac{1}{16} \cdot log(\frac{1}{16}) = 4bit$

Ortak, Koşullu ve Marjinal Entropi

Tanım

(X,Y) raslantısal iki boyutlu vektörel değişkenimiz olsun. r_{ij}, x_i ve y_j sonuçlarının ortaya çıkma olasılığı ise, *Ortak (joint) Bilgi Miktarı*

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \cdot log(r_{ij})$$

Tanım

(X,Y) raslantısal iki boyutlu vektörel değişkenimiz olsun. r_{ij}, x_i ve y_j sonuçlarının ortaya çıkma olasılığı, q, Y'nin marjinal olasılığı ise $H(Y|X)$, Y'nin X olayına bağlı, yani *Koşullu Bilgi Miktarı*:

$$H(Y|X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \cdot log(q(y_j|x_i))$$

Ortak, Koşullu ve Marjinal Entropi

Teorem

$H(Y|X)$ koşullu bilgi miktarı olsun.

- $H(Y|X) \geq 0$
- *$H(Y|X) \leq H(Y)$ ve eşitlik durumu ancak X ve Y bağımsız ise gerçekleşir.*

Teorem

$H(Y|X)$ koşullu, $H(X)$,$H(Y)$ marjinal ve $H(X,Y)$ ortak bilgi miktarları olsun.

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

İspatları [1]'de bulabilirsiniz.

Karşılıklı Bilgi Miktarı

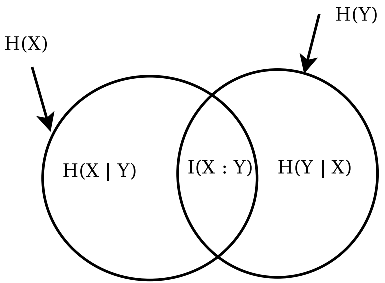
Tanım

X ve Y'ye bağlı *karşılıklı bilgi miktarı* aşağıdaki gibi tanımlanır.

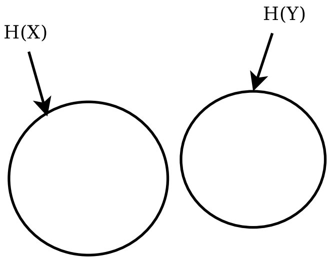
$$I(X:Y) = H(Y) - H(Y|X) = \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \cdot log(\frac{r_{ij}}{p_i \cdot q_j})$$

- Y ve X arasındaki bağımlılık miktarının ölçütüdür.
- X, Y bağımsız değişkenler ise $I(X:Y) = 0$ dır.
- Y, X'e tamamen bağımlı ise; $H(Y|X) = 0$ ve $I(X:Y) = H(Y)$.

Bilgi Miktarları Arasındaki İlişki

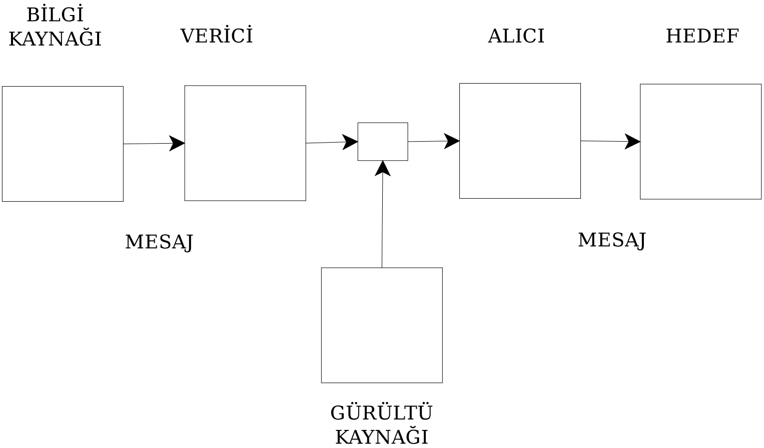


GENEL DURUM

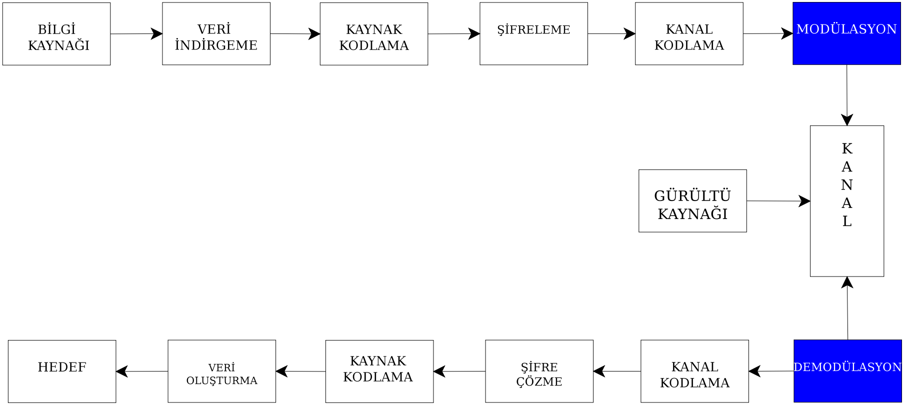


Bağımsız Olayların Entropisi

Shannon - Weaver İletişim Modeli



Detaylı İletişim Modeli



Belleksiz Kesikli Bilgi Kaynağı

- Bir bilgi kaynağının alfabesini, sonlu sayıda sembolden oluşan bir küme olarak düşünebiliriz.
- $U = (u_1, u_2, \dots, u_n)$
- Mesaj veya kelime ardarda gelen bir grup sembolden oluşmaktadır. Alfabede n sembol varsa ve mesajımızın uzunluğu l sembol ise n^l adet olası mesaj oluşturabiliriz.
- Bilgi kaynağımızın stokastik olması, alfabedeki her sembol ile ilintili bir oluşma olasılığının, $p(u_i) = p_i$, var olması anlamına gelmektedir.
- Bu durumda, bu bilgi kaynağı tarafından üretlien bilgi miktarı, aşağıdaki formül ile bulunabilir.

$$H(U) = -p_i \cdot log(p_i) bit/sembol$$

- Maksimum üretilebileceğimiz bilgi miktarı ise:

$$max[H(U)] = log(n) bit/sembol$$

Fazlalık (Redundancy) Kavramı

Tanım

Bir bilgi kaynağındaki fazlalık

$$Red = 1 - \frac{H(U)}{max[H(U)]} = 1 - \frac{H(U)}{log(n)}$$

ile gösterilir. Formüldeki n kaynak alfabedeki olası sembol sayısıdır.

Kaynak Kodlama

- $U = (u_1, u_2, \ldots, u_n)$ alfabesini kullanarak mesaj üreten bir kaynağımız olduğunu varsayalım ve mesajların bir iletişim kanalına yazıldığını düşünelim.
- Verimliliği sağlamak adına mesajların mümkün olduğunca fazlalıklardan kurtulması gerekir.
- Üretilen mesajın fazlalıklardan kurtulmasını sağlamak adına yeni bir alfabe ile kodlanmasına *kaynak kodlama* denir.

Kaynak Kodlama

- $S = (s_1, s_2, \ldots, s_n)$ kod alfabemiz olsun.
- U alfabesinde üretilen her bir mesajın S alfabesinde bir karşılığı mevcuttur, bu karşılığa kod kelimesi (code word) denir.
- Eğer olası tüm kod kelimeler birbirinden farklıysa, bu kodlama tekniğine (non - singular) tekil olmayan kodlama denir.
- Kodlamanın başarılı bir şekilde geriye çevrilebilmesi için, her bir kod kelimesinin tek bir karşılığı olması gerekmektedir.
- Aynı zamanda kod sembollerinin okunduğu anda orjinal mesajın alfabesine çevrilebilmesine ani kodlama (instantaneous code) denir.

Verimlilik

Tanım

Bir kodlama tekniğinin verimliliği, η, aşağıdaki şekilde tanımlanır.

$$\eta = \frac{H(U)}{L \cdot log(r)}$$

$H(U)$ mesajın kaynağının bilgi miktarı, L kod kelimelerinin ortalama uzunluğu ve r ise kod alfabesinin uzunluğudur.

Huffman Kodlaması - Örnek

- Kayıpsız (lossless) olarak veriyi sıkıştırıp tekrar açmak için kullanılır.
- Huffman Kodlamasının en büyük avantajlarından birisi kullanılan karakterlerin frekanslarına göre bir ikili ağaç (binary tree) yapısında tutulmasıdır.
- Sık kullanılan karakterler daha kısa kod kelimelerle ifade edilip ağacın üst düğümlerinde yer alır.
- Nadir kullanılanlar ise daha uzun kelimelerle ifade edilip daha aşağıdaki düğümlerde yer alır.
- Örnek: Türkçe dilinde yapılan Huffman Kodlaması çalışması. . .

Bellekli Kesikli Bilgi Kaynağı

- Belleksiz olma özelliği bir çok dil için geçerli değildir. Çoğu zaman önceki karakterlerden bir sonraki karakterin tahmini yapılabilmektedir.
- Örn: *bilg* harflerini gördüğümüzde bir sonraki harfin çok büyük olasılıkla "i" olacağını tahmin edebiliriz.
- Bilgi kaynağımızın bellekli olması demek, o ana kadar kaynaktan gelen sembollerin bir sonraki sembolün olasılıklarını etkilemesi anlamına gelmektedir.
- Bellekli bilgi kaynakların ürettiği diziler markov zincirleri ile gosterilebilir.

Markov Zincirleri

Tanım

$X_n, (n = 0, 1, 2, \dots,)$ sonlu sayıda değerler alan bir stokastik proses olsun. $X_n = i$ notasyonu, prosesin n anında i değerinde olduğunu göstermektedir. Proses herhangi bir n anında i durumunda olduğunda, bir sonraki durumun j olma olasılığı sabittir.

$$P\{X_{n+1} = j | X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} = P_{ij}$$

i, j ve $n \geq 0$.
Bu özelliği sağlayan proseslere *Markov Zinciri* denir.

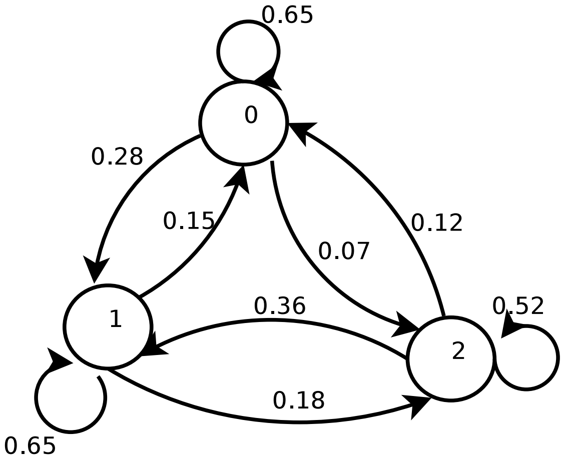
- Markov zincirleri durum geçiş matrisleri ile ifade edilir.

Örn: Markov Zincirleri

- Alfabemiz $U = 0, 1, 2$ olsun.
- Durum geçiş matrisimiz ise şu şekilde olsun.

		Bir sonraki durum		
		0	1	2
	0	0.65	0.28	0.07
Şu anki durum	1	0.15	0.65	0.18
	2	0.12	0.36	0.52

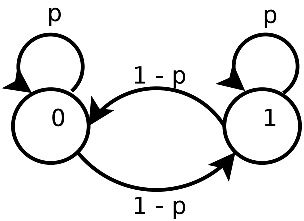
Örn: Markov Zincirleri



Örn: İletişim Sistemi

- Bir ağ üzerinden 1 ve 0 gönderen bir iletişim sistemi düşünelim. Her bir adımda bir sonraki sembolün değişme olasılığı p olsun. Bu durumda geçiş matrisimiz şu şekildedir.

$$\begin{Vmatrix} p & 1-p \\ 1-p & p \end{Vmatrix}$$



Chapman - Kolmogorov Denklemleri

- Bir prosesin n adımdan sonra durum değiştirme olasılıklarını Chapman - Kolmogorov denklemleri vermektedir.
- Bu denklemler

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^{n} \cdot P_{kj}^{m}$$

tüm $n, m \geq 0$ ve tüm i, j için sağlanır

N. dereceden Markov Zincirleri

İletişim sistemini tekrardan değerlendirirsek, $p = 0.7$ iken, 2. bitin ve 3. bitin olasılıklarını hesaplamak için:

$$P = \begin{vmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{vmatrix}$$

$$P^2 = P^{1+1} = P \cdot P$$

$$P^2 = \begin{vmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{vmatrix} \cdot \begin{vmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{vmatrix}$$

$$P^2 = \begin{vmatrix} 0.58 & 0.42 \\ 0.42 & 0.58 \end{vmatrix}$$

$$P^3 = P^2 \cdot P^1$$

$$P^3 = \begin{vmatrix} 0.58 & 0.42 \\ 0.42 & 0.58 \end{vmatrix} \cdot \begin{vmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{vmatrix}$$

$$P^3 = \begin{vmatrix} 0.532 & 0.468 \\ 0.468 & 0.532 \end{vmatrix}$$

1. Dereceden Markov Proseslerinin Bilgi Miktarı

- $u_i, i = 1, 2, \ldots, m$ alfabemizdeki semboller olsun.
- $S_i, i = 1, 2, \ldots, m$ prosesimizin durumları olsun.
- u_{1i} durumundan u_{2j} durumuna geçiş durumunu düşünelim. Bu olasılığı $p(u_{2j}|u_{1i})$ ile gösterirsek, bu geçişteki bilgi miktarı:

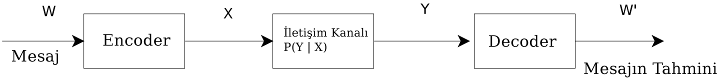
$$H(U_2|U_1) = -\sum_{i=1}^{m}\sum_{j=1}^{m} p(u_{2j}|u_{1i}) \cdot log(p(u_{2j}|u_{1i})) \qquad (1)$$

N. Dereceden Markov Proseslerinin Bilgi Miktarı

- $n-1$ önceki sembollerin bilindiği durumda, N. sembolün koşullu entropisi aşağıdaki gibi hesaplanır.

$$F_n(U) = H\left(U_n \middle| U_{n-1}, U_{n-2}, \ldots, U_1\right)$$

İletişim Kanalı Kapasitesi



- Güvenilir bir şekilde bir iletişim kanalından gönderebileceğiniz maksimum bilgi miktarına *Operasyonel iletişim kanalı kapasitesi* denir.
- *Bilgi Kanalı İletişim Kapasitesi* ise şu şekilde tanımlanır.

$$C = max_{\{p(x)\}}(I(X : Y))$$

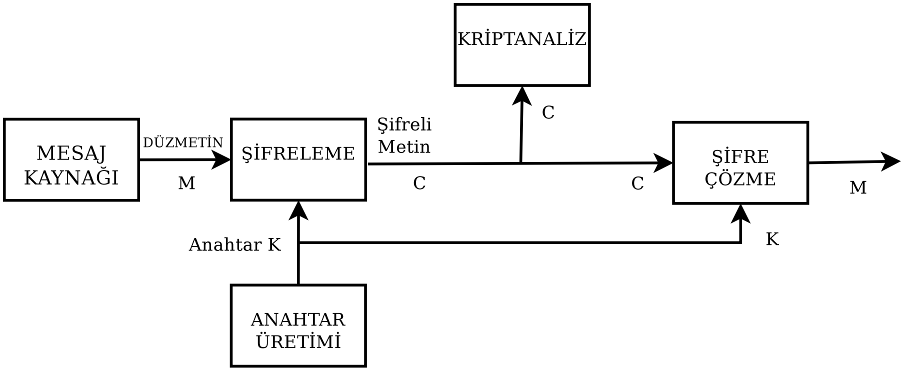
- Shannon'ın İletişim Kanalı Kodlama Teorisi'ne göre:

bilgi iletişim kapasitesi = operasyonel iletişim kapasitesi

Veri Sıkıştırma

- Kayıpsız sıkıştırma algoritmaları
  - Huffman
  - Fano - Shannon
  - Arithmetic Coding
- Kayıplı sıkıştırma algoritmaları
  - JPEG, DJVU
  - MPEG 1-2-4,
  - MP2, MP3, WMA

Simetrik Kriptosistemler



Simetrik Kriptosistemler

- M_i, $H(M_i) = -\sum_{i=1}^{n} p(M_i) \cdot log(p(M_i))$
- C_i, $H(M_i) = -\sum_{i=1}^{n} p(C_i) \cdot log(p(C_i))$
- K_i, $H(K_i) = -\sum_{i=1}^{n} p(K_i) \cdot log(p(K_i))$
- $H(K|C)$, *key equivocation*

$$H(K|C) = -\sum_{h=1}^{l}\sum_{j=1}^{m} p(K_h, C_j) \cdot log(p(K_h|C_j))$$

- $H(M|C)$, *message equivocation* aynı şekilde hesaplanmaktadır.
- $H(M|C,K) = 0$ olmasını bekleriz.
- Kusursuz gizlilik için ideal olarak $I(M:C) = 0$ olması gerekir.

Kolmogorov Karmaşası

- 1111111111111111
- 1011010110101101

Hangisini daha karmaşık olarak nitelendirebiliriz?

Kolmogorov Karmaşası

- 1111111111111111
- Yukarıdaki bit dizisi 16 adet 1 yazan Turing Makinesi ile ifade edilebilir.
- 1011010110101101
- Rasgele bit dizisi olarak nitelendirildiğinden tanımlayıcı Turing Makinesi bit dizisini sıkıştırmadan içermelidir.
- Kolmogorov Karmaşası ile bu fikri formal hale getirebiliriz.

Kolmogorov Karmaşası

Tanım

Bir x bit dizisinin Kolmogorov karmaşası, x bit dizisini üreten en küçük Turing Makinesi programının uzunluğuna eşittir.

$$K_f(x) = min_p\{|p| : f(p) = x\}$$

"Algorithmic information theory is the result of putting Shannon's information theory and Turing's computability theory into a cocktail shaker and shaking vigorously." —G. J. Chaitin

Shannon Entropisi ve Kolmogorov Karmaşası

- Kolmogorov karmaşasının beklenen değeri bize Shannon Entropisini vermektedir.
- Kodlama teorisinie göre mesajımız fazlalıkları azaltacak şekilde kodlanmaktadır. Bu kodlamayı Kolmogorov Karmaşasının tanımındaki f olarak değerlendirebiliriz.
- Bu durumda kodlama sisteminin beklenen değeri ile Shannon Entropisi arasındaki bağlantı kurulmuş olur.
- Formal teoremler ve ispatlar için [4]

Teşekkürler. . .

References

Jan C.A. van der Lubbe.
Information Theory.
Cambridge Press, 1997.

Kaan Kurtel.
Information theory lecture notes.

Neil Conway.
Kolmogorov complexity lecture notes.

Grunwald p. Vitanyi, P.
Shannon information and komogorov complexity.
2010.

C. E. Shannon.
A mathematical theory of communication.
Bell system technical journal, 27, 1948.

C. Shannon.
Communication theory of secrecy systems.
Bell System Technical Journal, 28, 1949.

Thomas M. Cover and Joy A. Thomas.
Elements of information theory.
Wiley-Interscience, 1991.

Sheldon M. Ross.
Introduction to Probability Models.
Academic Press, New York, 10th edition, 2003.