Entropi

Entropi Nedir? Entropi, bilgi teorisinde, bir veri setinin belirsizlik veya düzensizlik miktarını ölçen bir kavramdır. Claude Shannon tarafından geliştirilen bu kavram, bir rastgele değişkenin belirsizliğini ve dolayısıyla bilgi içeriğini ifade eder.

Matematiksel olarak entropi, şu şekilde tanımlanır: H(X)=−∑i=1np(xi)log⁡2p(xi)H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)H(X)=−∑i=1np(xi)log2p(xi)

Burada:

- XXX: Rastgele değişken
- xix_ixi: Rastgele değişkenin olası değerleri
- p(xi)p(x_i)p(xi): xix_ixi değerinin olasılığı

Entropinin Özellikleri:

- **Belirsizlik Ölçüsü:** Yüksek entropi, daha fazla belirsizlik ve düzensizlik anlamına gelir. Düşük entropi ise daha az belirsizlik demektir.
- **Maksimum Entropi:** Olasılıklar eşit olduğunda entropi maksimum olur. Örneğin, adil bir zarın (1, 2, 3, 4, 5, 6) atılması durumunda entropi maksimumdur çünkü her sonucun olasılığı eşittir.
- **Minimum Entropi:** Tüm olasılıklar bir olayda yoğunlaştığında entropi minimumdur (sıfır). Örneğin, bir zarın her zaman "1" gelmesi durumunda entropi sıfırdır çünkü belirsizlik yoktur.

$$H = -\sum p(x) \log p(x)$$

Bilgi Kazancı

Bilgi Kazancı Nedir? Bilgi kazancı, bir veri setindeki bir özniteliğin, hedef değişkenin belirsizliğini ne kadar azalttığını ölçer. Başka bir deyişle, bilgi kazancı, bir öznitelik kullanılarak veri setinin ne kadar daha düzenli hale getirildiğini ifade eder. Karar ağaçlarında, bilgi kazancı, düğüm bölünmelerini değerlendirmek için kullanılır.

Matematiksel olarak bilgi kazancı, şu şekilde hesaplanır:
Bilgi Kazancı=H(D)−H(D∣A)\text{Bilgi Kazancı} = H(D) - H(D|A)Bilgi Kazancı=H(D)−H(D∣A)

Burada:

- H(D)H(D)H(D): Veri setinin entropisi
- H(D∣A)H(D|A)H(D∣A): Özniteliğe göre bölünmüş veri setinin entropisi

Örnek: Diyelim ki, bir veri setimizde "Yağmurlu" (R) ve "Güneşli" (S) olarak etiketlenmiş hava durumu bilgileri var. Bu veri setinde havanın güneşli veya yağmurlu olmasının olasılıkları şöyle olsun:

- P(R)=0.5P(R) = 0.5P(R)=0.5
- P(S)=0.5P(S) = 0.5P(S)=0.5

Veri setinin entropisi şu şekilde hesaplanır:
H(D)=−[P(R)log⁡2P(R)+P(S)log⁡2P(S)]=−[0.5log⁡20.5+0.5log⁡20.5]=1H(D) = -[P(R) \log_2 P(R) + P(S) \log_2 P(S)] = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1H(D)=−[P(R)log2 P(R)+P(S)log2P(S)]=−[0.5log20.5+0.5log20.5]=1

Bir öznitelik ekleyelim: "Rüzgarlı" (W) ve "Rüzgarsız" (NW). Diyelim ki, veri seti bu özniteliğe göre şu şekilde bölünmüş olsun:

- Rüzgarlı günlerde: %70 yağmurlu (R), %30 güneşli (S)
- Rüzgarsız günlerde: %20 yağmurlu (R), %80 güneşli (S)

Bu durumda, rüzgarlı ve rüzgarsız günlerin entropileri şu şekilde hesaplanır:
H(D|W)=−[0.7log⁡20.7+0.3log⁡20.3]≈0.88H(D|W) = -[0.7 \log_2 0.7 + 0.3 \log_2 0.3] \approx 0.88H(D|W)=−[0.7log20.7+0.3log20.3]≈0.88
H(D|NW)=−[0.2log⁡20.2+0.8log⁡20.8]≈0.72H(D|NW) = -[0.2 \log_2 0.2 + 0.8 \log_2 0.8] \approx 0.72H(D|NW)=−[0.2log20.2+0.8log20.8]≈0.72

Eğer veri setinde günlerin %40'ı rüzgarlı ve %60'ı rüzgarsız ise, şartlı entropi şöyle hesaplanır: H(D|A)=0.4⋅H(D|W)+0.6⋅H(D|NW)=0.4⋅0.88+0.6⋅0.72=0.792H(D|A) = 0.4 \cdot H(D|W) + 0.6 \cdot H(D|NW) = 0.4 \cdot 0.88 + 0.6 \cdot 0.72 = 0.792H(D|A)=0.4⋅H(D|W)+0.6⋅H(D|NW)=0.4⋅0.88+0.6⋅0.72=0.792

Son olarak, bilgi kazancı şöyle hesaplanır:
Bilgi Kazancı=H(D)−H(D|A)=1−0.792=0.208\text{Bilgi Kazancı} = H(D) - H(D|A) = 1 - 0.792 = 0.208Bilgi Kazancı=H(D)−H(D|A)=1−0.792=0.208

Bu bilgi kazancı, "Rüzgarlı" özniteliğinin hava durumunun belirsizliğini azaltma miktarını gösterir.

Karar Ağaçlarında Kullanımı

Karar ağaçlarında bilgi kazancı, ağacın dallarını oluştururken hangi özniteliklerin en iyi ayrımı yaptığını belirlemek için kullanılır. En yüksek bilgi kazancına sahip öznitelik, ağacın kök düğümü olarak seçilir ve bu süreç tekrarlanarak ağaç derinleştirilir.

Sonuç

Entropi ve bilgi kazancı, makine öğrenmesi ve bilgi teorisinde veri setlerindeki belirsizlikleri ölçmek ve azaltmak için kullanılan temel kavramlardır. Entropi, verilerin ne kadar belirsiz olduğunu ölçerken, bilgi kazancı bu belirsizliği azaltma kapasitesini gösterir. Bu kavramlar özellikle karar ağaçları gibi algoritmalarda kritik öneme sahiptir.