ML Notes
Ayşe Bilge İnce
30 January 2017

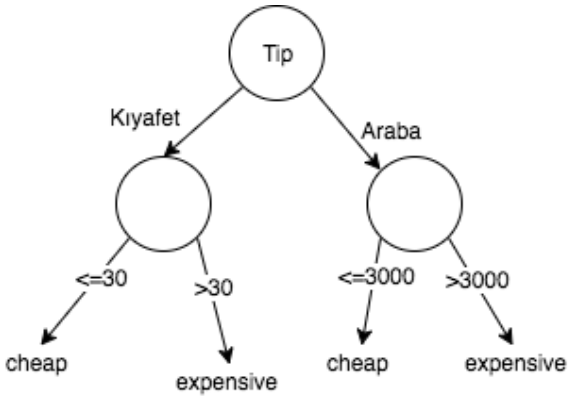
Machine Learning 101

Öğrenme Yöntemleri

1. Eğitmenli Öğrenme
2. Denetimsiz Öğrenme
3. Destekli Öğrenme

Karar Ağacı (ID3)

Parametrik olmayan, eğitimli bir yöntemdir. Sınıflandırma ve regresyon için kullanılırlar. Bir karar ağacı ara karar düğümü ve terminal yapraklardan oluşur. Yaprak düğümlerin bir output değeri vardır. Bu output sınıflandırmada sınıf kodu, regresyonda da nümerik bir değer olarak görünür. Karar ağaçlarında alt kümelere bölmenin amacı her alt kümeyi olabildiğince homojen hale getirmektir. Dezavantajı karar ağacı algoritmalarının greedy yöntemler olmasıdır.



Entropi

2-sınıflı Entropi(S) = $-(p_+ * \log_2 p_+ + p_- * \log_2 p_-)$

n-sınıflı Entropi(S) => $E(S) = \sum_{i=1}^{n} -p_i * log_2 p_i$

Örnek 1:

Örnek Sayısı	Örnek Sınıfı
9	+
5	-

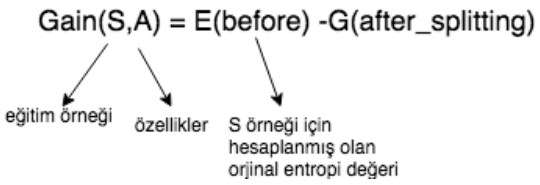
$p_+ = 9/(9+5) = 9/14$ (aynı eğitim sınıfında bulunan bir örneğin + sınıfta bulunma olasılığı)

$p- = 5/14$ (aynı eğitim sınıfında bulunan bir örneğin - sınıfta bulunma olasılığı)

$E= - 9/14*\log_2(9/14) - 5/14*\log_2(5/14) = 0.94$

Kazanım

Karar ağaçlarında kökler, düğümler ve yapraklar kazanım değerine göre oluşturulur.



Örnek 2:

Aşağıdaki tabloyu eğitim seti olarak kabul edip, tabloyu ID3 karar ağacı algoritmasına göre sınıflandırınız:

Haftasonu	Hava Durumu	Ebeveyn Durumu	Para Durumu	Karar(Sınıf)
H1	Güneşli	Evet	Zengin	Sinema
H2	Güneşli	Hayır	Zengin	Tenis
H3	Rüzgarlı	Evet	Zengin	Sinema
H4	Yağmurlu	Evet	Fakir	Sinema
H5	Yağmurlu	Hayır	Zengin	Ev
H6	Yağmurlu	Evet	Fakir	Sinema
H7	Rüzgarlı	Hayır	Fakir	Sinema
H8	Rüzgarlı	Hayır	Zengin	Alışveriş
H9	Rüzgarlı	Evet	Zengin	Sinema
H10	Güneşli	Hayır	Zengin	Tenis

Not: ID3 algoritması WEKA aracında C4.5 olarak gösterilmektedir.

Öncelikle tüm tabloyu en doğru şekilde ikiye bölecek olan özellik seçilmelidir. Bunun için de en yüksek kazanım veren özellik belirlenmelidir.

10 adet eğitim örneği için değerler şu şekilde bölünmektedir.

* 6 adet Sinema
* 2 adet Tenis
* 1 adet Ev
* 1 adet Alışveriş

Başlangıç için bu değerler üzerinden Entropi değeri hesaplanmalıdır.

$E(S) = -((6/10)*\log_2(6/10) + (2/10)*\log_2(2/10) + (1/10)*\log_2(1/10) + (1/10)*\log_2(1/10))$

$E(S) = 1.571$ (bu değer başlangıç entropi değeri olarak Information Gain'i hesaplamak için kenarda tutulacak.)

Tek tek tüm özelliklerin kazanım değerlerleri hesaplanarak en yüksek kazanım değerine sahip olan özellik kök düğümü olarak seçilir:

Gain(S,Hava Durumu) = ?

Güneşli = 3 (1 Sinema + 2 Tenis)

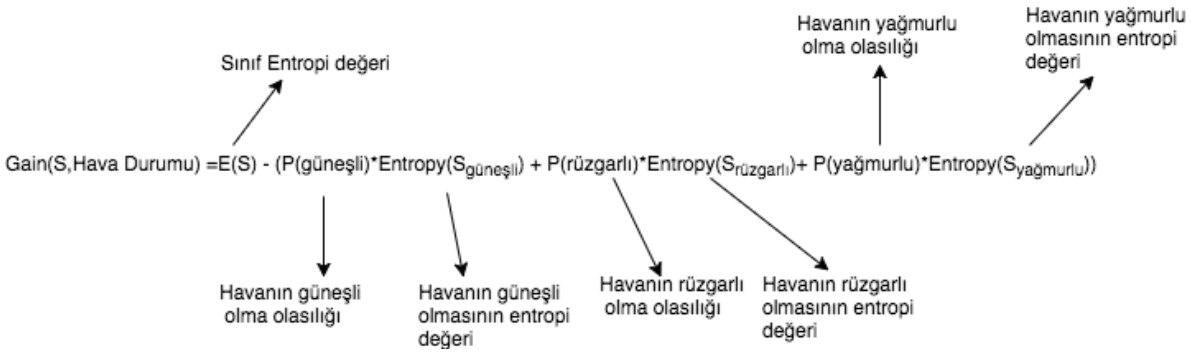
Rüzgarlı = 4 (3 Sinema + 1 Alışveriş)

Yağmurlu = 3 (2 Sinema + 1 Ev)

$Entropy(S_{güneşli}) = -(1/3)*\log_2(1/3) - (2/3)*\log_2(2/3) = 0.918$

$Entropy(S_{rüzgarlı}) = -(3/4)*\log_2(3/4) - (1/4)*\log_2(1/4) = 0.811$

$Entropy(S_{yağmurlu}) = -(2/3)*\log_2(2/3) - (1/3)*\log_2(1/3) = 0.918$



Gain(S, Hava Durumu) = 1.571 - (((1+2)/10)*0.918 + ((3+1)/10)*0.811 + ((2+1)/10)*0.918)
Gain(S, Hava Durumu) = 0.70

Gain(S, Ebeveyn) = ?

Evet = 5 (5 adet Sinema)

Hayır = 5 (2 adet Tenis + 1 adet Sinema + 1 adet Alışveriş + 1 adet Ev)

Entropy(S_{evet}) = - (5/5)*$\log_2$(5/5) = 0

Entropy($S_{hayır}$) = -(2/5)*$\log_2$(2/5) -3*(1/5)*$\log_2$(1/5) = 1.922

Gain(S, Ebeveyn) = Entropy(S) - (P(evet)*Entropy(S_{evet}) + P(hayır)*Entropy($S_{hayır}$))

Gain(S, Ebeveyn) = 1.571 - ((5/10)*Entropy(S_{evet}) + (5/10)*Entropy($S_{hayır}$))

Gain(S, Ebeveyn) = 0.61

Gain(S, Para)=?

Zengin = 7 (3 Sinema + 2 Tenis + 1 Alışveriş + 1 Ev)

Fakir = 3 (3 Sinema)

Entropy(S_{zengin}) = 1.842

Entropy(S_{fakir})=0

Gain(S, Para) = Entropy(S) - (P(zengin)*Entropy(S_{zengin}) + P(fakir)*Entropy(S_{fakir}))

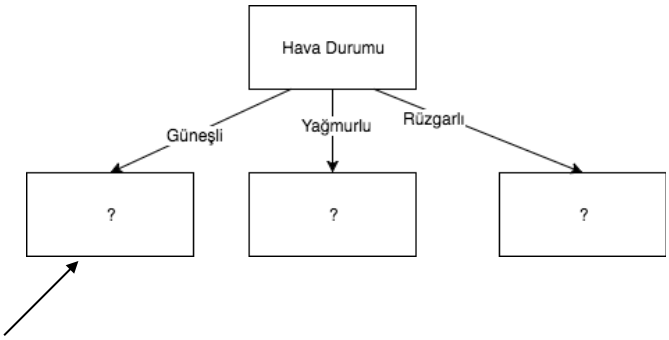
Gain(S, Para) = 0.2816

Son durumda tüm kazanım değerleri alt alta sıralanıp içlerinden en yüksek kazanım değerine sahip olan özellik kök düğüm olarak seçilir:

Gain(S, Hava Durumu) = 0.70

Gain(S, Ebeveyn) = 0.61

Gain(S, Para) = 0.2816



Hafta sonu	Hava	Ebeveyn	Para	Karar (Sınıf)
H1	Güneşli	Evet	Zengin	Sinema
H2	Güneşli	Hayır	Zengin	Tenis
H10	Güneşli	Hayır	Zengin	Tenis

Yukarıdaki örnekte görüldüğü gibi kök seçildikten sonra özelliğe ait tekil değerlerin her biri yaprak olarak belirlenmiş ve bu kez veri seti bu tekil değerler ile özelleştirilmiştir. Yani yukarıdaki tablo hava durumunun "Güneşli" olması durumunda diğer özelliklerin alabileceği

değerleri gösteren ayrı bir veri setine dönüştürülmüştür. Böylelikle hava durumu güneşli olduğunda bir alt yaprağın hangi özelliğe ait olacağı bu tabloya göre aşağıdaki gibi bulunacaktır:

* 1 adet Sinema
* 2 adet Tenis

$Entropy(S_{güneşli}) = -(1/3)*\log_2(1/3) -(2/3)*\log_2(2/3) = 0.918$

$Gain(S_{güneşli},Ebeveyn) = ?$

$Gain(S_{güneşli},Ebeveyn) = Entropy(S_{güneşli}) - (P(evet \mid S_{güneşli})*Entropy(S_{evet}) + P(hayır \mid S_{güneşli})*Entropy(S_{hayır}))$

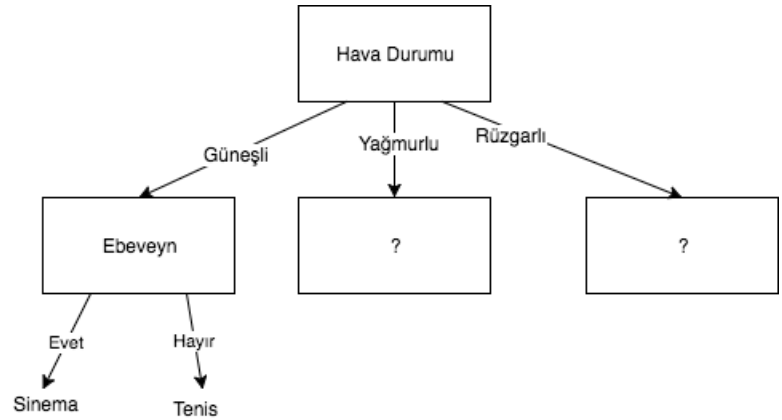
$Entropy(S_{evet}) = -(1/3)*\log_2(0) - 0 = 0$

$Entropy(S_{hayır}) = -(2/3)*\log_2(0) - 0 = 0$

$Gain(S_{güneşli},Ebeveyn) = 0.928 - ((1/3)*0 + (2/3)*0) = 0.928$

$Gain(S_{güneşli}, Para) = 0.918 - ((3/3)*0.918 + (0/3)*0) = 0$

Ebeveyn özelliğinin kazanım değeri daha fazla olduğu için Güneşli durumunun yaprağı Ebeveyn olacaktır. Güneşli durumu için oluşturulan veri setine göre de eğer Evet ise Sinema kararı, hayır ise de Tenis kararı çıkacaktır:



Rüzgarlı ve Yağmurlu durumları için de aynı işlemler yapılarak, son durumda karar ağacı tamamlanacaktır.

Overfitting: Öğrenmenin ezberlemeye dönüşmesidir. Başarım grafiğinde belli bir andan sonra başarımın yükselmesinin durup, düşmeye başlamasıdır.