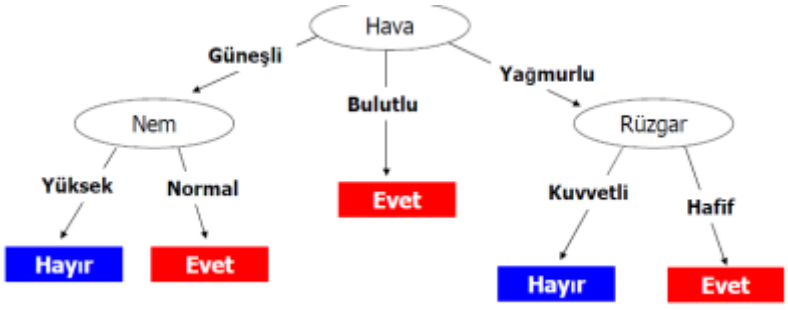
Karar Ağacı:

Karar ağacı algoritması, denetimli öğrenme kategorisine girer. Hem regresyon hem de sınıflandırma problemlerini çözmek için kullanılılar. Karar ağacı, her yaprak düğümün bir sınıf etiketine karşılık geldiği ve özniteliklerin ağacın iç düğümünde temsil edildiği sorunu çözmek için ağaç temsilini kullanır. Karar ağacını kullanarak herhangi bir boole fonksiyonunu ayrık öznitelikler üzerinde temsil edebiliriz.

Karar ağacı öğrenmesi, bir karar ağacını, bir öğeyle ilgili (dallarda temsil edilen) gözlemlerden öğenin hedef değeri (yapraklarda temsil edilen) ile ilgili sonuçlara gitmek için bir tahmin modeli olarak kullanır. İstatistik, veri madenciliği ve makine öğrenmesinde kullanılan öngörülü modelleme yaklaşımlarından biridir. Hedef değişkenin ayrı bir değer kümesi alabileceği ağaç modellerine sınıflandırma ağaçları denir; bu ağaç yapılarında yapraklar sınıf etiketlerini ve dallar bu sınıf etiketlerine yol açan özelliklerin birleşimlerini temsil eder. Hedef değişkenin sürekli değerler alabileceği karar ağaçlarına (tipik olarak gerçek sayılar) regresyon ağaçları denir. Karar analizinde, bir karar ağacı, kararları ve karar almayı görsel ve açık bir şekilde temsil etmek için kullanılabilir. Veri madenciliğinde, bir karar ağacı verileri tanımlar, ancak sonuçta ortaya çıkan sınıflandırma ağacı karar verme için bir girdi olabilir.

Karar ağaçları metotu, giriş verisinin bir algoritma yardımıyla gruplara bölünerek tüm elemanlarının aynı sınıf etiketine sahip olması için yapılan sınıflama işlemidir. Giriş verisinin bir kümeleme algoritması yardımıyla tekrar tekrar gruplara bölünmesine dayanır. Grubun tüm elemanları aynı sınıf etiketine sahip olana kadar kümeleme işlemi derinlemesine devam eder.



Karar ağacı kullanılırken yapılan bazı varsayımlar aşağıdadır:

- Başlangıçta, tüm eğitim seti kök olarak kabul edilir.

• Özellik değerlerinin kategorik olması tercih edilir. Değerler sürekli ise, model oluşturmadan önce ayrıklaştırılırlar.

• Öznitelik değerleri temelinde, kayıtlar özyinelemeli olarak dağıtılır.

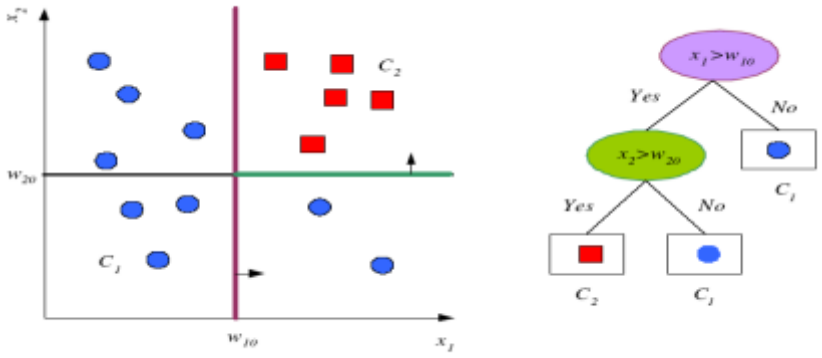
• Öznitelikleri kök veya dahili düğüm olarak sıralamak için istatistiksel yöntemler kullanılır. Karar ağacı tipleri ikiye ayrılır:

• Entropiye dayalı sınıflandırma ağaçları (ID3, C4.5)

• Regresyon ağaçları (CART). Karar ağaçları çok boyutlu (özellikli) veriyi belirlenmiş şartlara bağlı olarak parçalara böler. Her adımda verinin hangi özelliği üzerinde işlem yapılacağına karar vermek çok büyük bir kombinasyonun çözümüyle mümkündür. Örneğin, 5 özellik ve 20 örneğe sahip bir veride 10^6 dan fazla sayıda farklı karar ağacı oluşturulabilir. Bu sebeple her parçalanmanın metodolojik olması gerekir. Quinlan'e göre veri bir özelliğine göre bölündüğünde elde edilen her bir veri kümesinin belirsizliği minimum ve dolayısıyla bilgi kazancı maksimum ise en iyi seçim yapılmıs demektir. Buna göre önerdiği ilk algoritma ID3'te tek tek özellik vektörleri incelenir ve en yüksek bilgi kazancına sahip özellik, ağaçta dallanma yapmak için tercih edilir. Karar Ağacı Algoritması: Karar ağaçları eğiticili öğrenme için çok yaygın bir yöntemdir.

Algoritmanın adımları:

1. T öğrenme kümesini oluşturulur.
2. T kümesindeki örnekleri en iyi ayıran nitelikler belirlenir.
3. Seçilen nitelik ile ağacın düğümleri oluşturulur ve herbir düğümde alt düğümler veya ağacın yapraklarını oluşturulur. Alt düğümlere ait alt veri kümesinin örneklerini belirlenir
4. 3. adımda oluşturulan her alt veri kümesi için
    a. Örneklerin hepsi aynı sınıfa aitse
    b. Örnekleri bölecek nitelik kalmamışsa
    c. Kalan niteliklerin değerini taşıyan örnek yoksa işlemi sonlandır. Diğer durumda alt veri kümesini ayırmak için 2. adımdan devam edilir.



Ezber (Overfitting: Aşırı Uyum):

• Tüm makine öğrenmesi yöntemlerinde verinin ana hatlarının modellenmesi esas alındığı için öğrenme modelinde ezberden (overfitting) kaçınılmalıdır.

• Tüm karar ağaçları önlem alınmazsa ezber yapar. Bu yüzden ağaç olusturulurken veya olusturulduktan sonra budama yapılmalıdır. Ağaç Budama: Budama, sınıflandırmaya katkısı olmayan bölümlerin karar ağacından çıkarılması islemidir. Bu sayede karar ağacı hem sade hem de anlaşılabilir hale gelir. İki çesit budama yöntemi vardır;

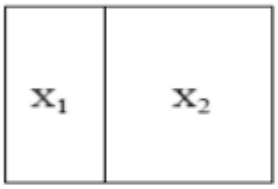
- Ön budama
- Sonradan budama

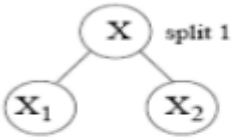
Ön budama islemi ağaç oluşturulurken yapılır. Bölünen nitelikler, değerleri belli bir esik değerinin (hata toleransının) üstünde değilse o noktada ağaç bölümleme islemi durdurulur ve o an elde bulunan kümedeki baskın sınıf etiketi, yaprak olarak olusturulur. Sonradan Budama: Sonradan budama islemi ağaç olusturulduktan sonra devreye girer. Alt ağaçları silerek yaprak olusturma, alt ağaçları yükseltme, dal kesme seklinde yapılabilir. Aşırı uyumu önlemek için ağacı büyütmeyi durdurabiliriz, ancak durdurma kriteri miyop olma eğilimindedir. Bu nedenle standart yaklaşım, "dolu" bir ağaç yetiştirmek ve ardından budama yapmaktır. Düğümdeki noktalar için yanlış bir sınıflandırma yapma olasılığı olduğu da unutulmamalıdır. Tüm ağaç için yanlış sınıflandırma olasılığını elde etmek için, toplam olasılık formülüne göre yaprak düğüm içi hata oranının ağırlıklı toplamı hesaplanır. Spesifik olarak, ana düğüm için ağırlıklı yanlış sınıflandırma oranının, sol ve sağ alt düğümlerin ağırlıklı yanlış sınıflandırma oranlarının toplamından daha büyük veya buna eşit olacaktır. Yeniden ikame hata oranını en aza indirirsek, her zaman daha büyük bir ağacı tercih edeceğimiz anlamına gelir. Aşırı uyuma karşı hiçbir savunma yoktur. Aşırı büyümüş ağaç er ya da geç budanacaktır. Ne zaman duracağınıza karar vermenin birkaç yolu vardır:

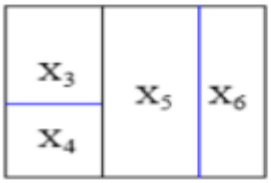
- ✓ Tüm terminal düğümleri saf olana kadar devam edilir.
- ✓ Her bir terminal düğümündeki veri sayısı belirli bir eşikten, örneğin 5'ten, hatta 1'den büyük olmayana kadar devam edilir.
- ✓ Ağaç yeterince büyük olduğu sürece, ilk ağacın boyutu kritik değildir.

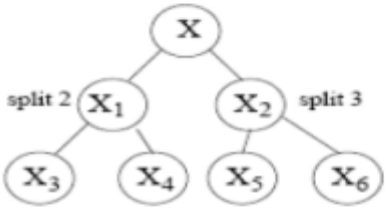
Buradaki anahtar, ilk ağacı yeniden budamadan önce yeterince büyük yapmaktır!

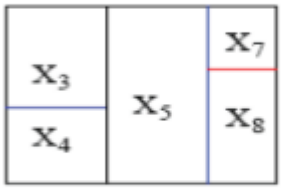
Sınıflandırma Ağaçları:

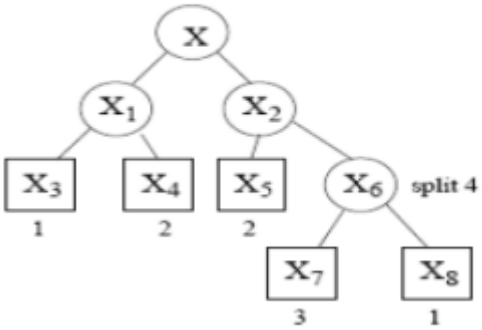


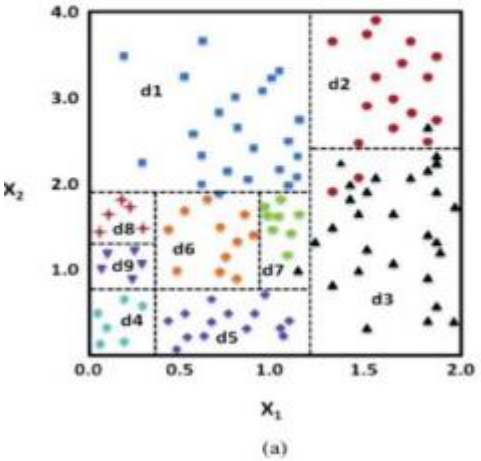


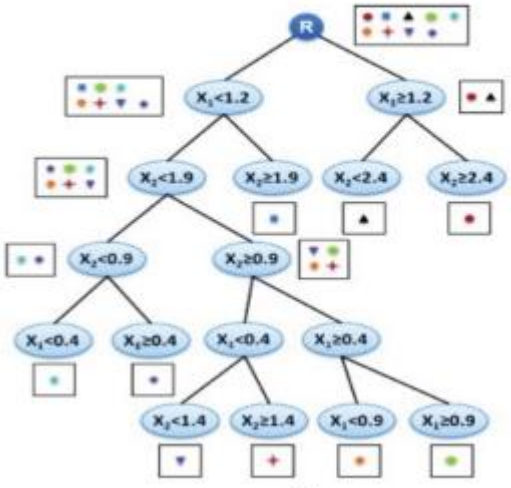












Karar Ağacında en büyük zorluk, her seviyede kök düğüm için özniteliğin tanımlanmasıdır. Bu işlem öznitelik seçimi olarak bilinir. İki popüler öznitelik seçim ölçüsü bulunmaktadır:

1. Bilgi Kazancı
2. Gini İndeksi

1) Bilgi Kazancı Eğitim örneklerini daha küçük alt kümelere bölmek için karar ağacında bir düğüm kullandığımızda entropi değişir. Bilgi kazancı, entropideki bu değişimin bir ölçüsüdür. Tanım: Diyelim ki S bir örnekler kümesi, A bir nitelik, Sv, S'nin A = v ile alt kümesi ve Değerler (A), A'nın tüm olası değerlerinin kümesidir, o zaman:

$$Gain(S, A) = Entropy(S) - \sum_{veValues(A)} \frac{|S_v|}{|S|}.Entropy(S_v)$$

Örnek: X = {a,a,a,b,b,b,b,b} kümesi için Toplam örnek: 8 b: 5 örnekleri a: 3'ün örnekleri

$$EntropyH(X) = -\left[\left(\frac{3}{8}\right) log_2 \frac{3}{8} + \left(\frac{5}{8}\right) log_2 \frac{5}{8}\right]$$

```
= -[0.375 * (-1.415) + 0.625 * (-0.678)]
=-(-0.53-0.424)
= 0.954
```

Bilgi Kazanımını Kullanarak Karar Ağacı Oluşturma Gereklilikler:

- Kök düğümle ilişkili tüm eğitim örnekleriyle başlanır
- Her bir düğümün hangi öznitelikle etiketleneceğini seçmek için bilgi kazancı kullanılır.
- Not: Hiçbir kökten yaprağa yol, aynı ayrık özniteliği iki kez içermemelidir
- Her alt ağacı, ağaçta o yolda sınıflandırılacak eğitim örneklerinin alt kümesinde yinelemeli olarak oluşturulur. Sınır vakaları:
- Tüm pozitif veya tüm negatif eğitim örnekleri kalırsa, o düğümü buna göre “evet” veya “hayır” olarak etiketlenr.
- Hiçbir öznitelik kalmazsa, o düğümde kalan eğitim örneklerinin çoğunluk oyu ile etiketlenir.
- Örnek kalmadıysa, ebeveynin eğitim örnekleri çoğunluk oyu ile etiketlenir.