Makine Öğrenmesi

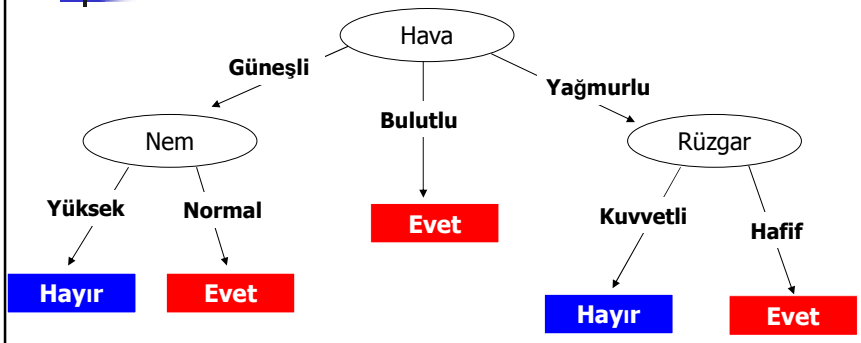
3. hafta

- Entropi
- Karar Ağaçları (Desicion Trees)
  - ID3
  - C4.5
- Sınıflandırma ve Regresyon Ağaçları (CART)

Karar Ağacı Nedir?

Temel fikir, 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 Nedir?



Karar Ağacı Tipleri

Entropiye dayalı sınıflandırma ağaçları (ID3, C4.5) ve Regresyon ağaçları (CART) olmak üzere iki kategoride birçok algoritma önerilmiştir.

Önce entropiye dayalı karar ağaçlarını inceleyeceğiz. Bu algoritmaları iyi anlayabilmek için önce entropiyi iyi bilmek gerekmektedir.

Entropi, Belirsizlik ve Enformasyon

Rassal bir değişkenin belirsizlik ölçütü olarak bilinen Entropi, bir süreç için tüm örnekler tarafından içerilen enformasyonun beklenen değeridir. Enformasyon ise rassal bir olayın gerçekleşmesine ilişkin bir bilgi ölçütüdür. Eşit olasılıklı durumlar yüksek belirsizliği temsil eder. Shannon'a göre bir sistemdeki durum değiştiğinde entropideki değişim kazanılan enformasyonu tanımlar. Buna göre maksimum belirsizlik durumundaki değişim muhtemelen maksimum enformasyonu sağlayacaktır.

Enformasyon

Aslında zıt şeyleri temsil etmelerine rağmen Shannon'a göre maksimum belirsizlik maksimum enformasyon sağladığı için Enformasyon ve Belirsizlik terimleri benzerdir. Enformasyon (self-information) formülü aşağıdaki gibidir. Shannon bilgiyi bitlerle temsil ettiği için logartimayı iki tabanında kullanımıştır.

$$I(x) = \log \frac{1}{P(x)} = -\log P(x)$$

Entropi

Shannon'a göre entropi, iletilen bir mesajın taşıdığı enformasyonun beklenen değeridir. Shannon Entropisi (*H*) adıyla anılan terim, tüm a_i durumlarına ait P_i olasılıklarına bağlı bir değerdir.

$$H(X) = E(I(X)) = \sum_{1 \le i \le n} P(x_i).I(x_i)$$

$$= \sum_{i=1}^{n} P(x_i) \log_2 \frac{1}{P(x_i)} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

Entropi

Bir paranın havaya atılması olayı, rassal X sürecini temsil etsin. Yazı ve tura gelme olasılıkları eşit olduğu için X sürecinin entropisi aşağıdaki gibidir.

$$H(X) = -\sum_{i=1}^{2} p_i \log_2 p_i$$

$$= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$$

Entropisi 1 olan para atma olayı (X) gerçekleştiğinde 1 bitlik bilgi kazanılacaktır.

Karar Ağacında Entropi

Karar ağaçları çok boyutlu (özellikli) veriyi belirlenmiş özellik üzerindeki bir şart ile parçalara böler. Her seferinde verinin hangi özelliği üzerinde hangi şarta göre işlem yapacağına karar vermek çok büyük bir kombinasyonun çözümüyle mümkündür. 5 özellik ve 20 örneğe sahip bir veride 10^6 dan fazla sayıda farklı karar ağacı oluşturulabilir. Bu sebeple her parçalanmanın metodolojik olması gerekir.

Karar Ağacında Entropi

Quinlan'e göre veri, bir özelliğe göre bölündüğünde elde edilen her bir veri kümesinin belirsizliği minimum ve dolayısıyla bilgi kazancı maksimum ise en iyi seçim yapılmış 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.

ID3 Algoritması

Sadece kategorik veri ile çalışan bir yöntemdir. Her iterasyonun ilk adımında veri örneklerine ait sınıf bilgilerini taşıyan vektörün entropisi belirlenir. Daha sonra özellik vektörlerinin sınıfa bağımlı entropileri hesaplanarak ilk adımda hesaplanan entropiden çıkartılır. Bu şekilde elde edilen değer ilgili özellik vektörüne ait kazanç değeridir. En büyük kazanca sahip özellik vektörü ağacın o iterasyonda belirlenen dallanmasını gerçekleştirir.

ID3 Örneği

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

2 özellik vektörü (V1 ve V2) ile S sınıf vektörüne sahip 4 örnekli veri kümesi verilmiştir. ID3 algoritması ile ilk dallanma hangi özellik üzerinde gerçekleşir ?

H(S) - H(V1,S)

H(S) - H(V2,S)

ID3 Örneği

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

Sınıf Entropisi

$$H(S) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

V1 Entropisi

$$H(V1) = \frac{1}{4}H(A) + \frac{3}{4}H(B)$$

$$= \frac{1}{4}0 - \frac{3}{4}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)$$

$$= 0 + \frac{3}{4}0{,}9183 = 0{,}6887$$

V2 Entropisi

$$H(V2) = \frac{1}{2}H(C) + \frac{1}{2}H(D) = \frac{1}{2} + \frac{1}{2} = 1$$

V1 seçilir...

C4.5 Algoritması

ID3 algoritmasının nümerik özellik içeren veriye uygulanabilen şeklidir. ID3'ten tek farkı nümerik özelliklerin kategorik hale getirilebilmesini sağlayan bir eşikleme yöntemini içermesidir. Temel mantık nümerik özellik vektöründeki tüm değerler ikili olarak ele alınarak ortalamaları eşik olarak denenir. Hangi eşik değeriyle bilgi kazanımı en iyi ise o değer seçilir. Seçilen eşiğe göre özellik vektörü kategorize edilir ve ID3 uygulanır.

Kayıp Veri

Eğer veride bazı örneklerin bazı özellikleri kayıpsa izlenecek iki yol vardır:

- Kayıp özelliklere sahip örnek veriden tamamen çıkartılır.
- Kayıp verilerle çalışabilecek şekilde algoritma düzenlenir.

Eğer kayıplı örneklerin sayısı birinci seçenek uygulanamayacak kadar çoksa ikinci seçenek uygulanmalıdır.

Kayıp Veri

Kayıp bilgiye sahip özellik vektörü için kazanç hesaplanırken kayıplı örnekler hariç tutularak bilgi kazancı normal şekilde hesaplanır ve daha sonra F katsayısıyla çarpılır. F, kayıpsız verinin tamamına oranıdır.

$$IG(X) = F.(H(X) - H(V, X))$$

Kayıp Veri

Kayıp bilgiye sahip özellik vektörü içinde en sık tekrarlanan değerin kayıp bilgi yerine yazılması da önerilen yöntemlerdendir.

Ezber (Overfitting)

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 oluşturulurken veya oluşturulduktan sonra budama yapılmalıdır.

Ağaç Budama

Budama, sınıflandırmaya katkısı olmayan bölümlerin karar ağacından çıkarılması işlemidir. Bu sayede karar ağacı hem sade hem de anlaşılabilir hale gelir. İki çeşit budama yöntemi vardır;

- ön budama
- sonradan budama

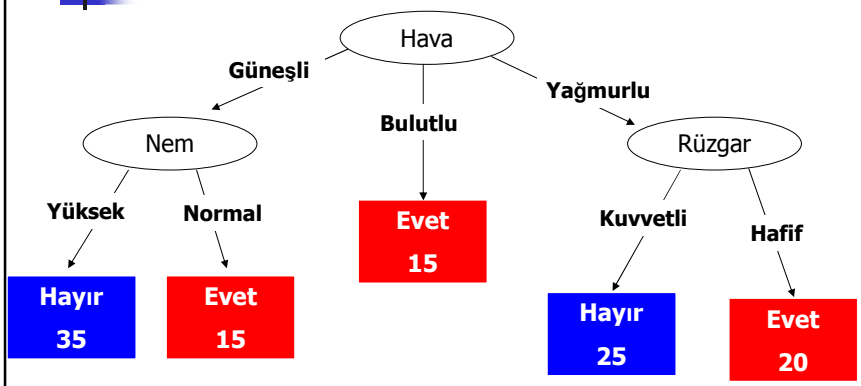
Ön Budama

Ön budama işlemi ağaç oluşturulurken yapılır. Bölünen nitelikler, değerleri belli bir eşik değerinin (hata toleransının) üstünde değilse o noktada ağaç bölümleme işlemi durdurulur ve o an elde bulunan kümedeki baskın sınıf etiketi, yaprak olarak oluşturulur.

Sonradan Budama

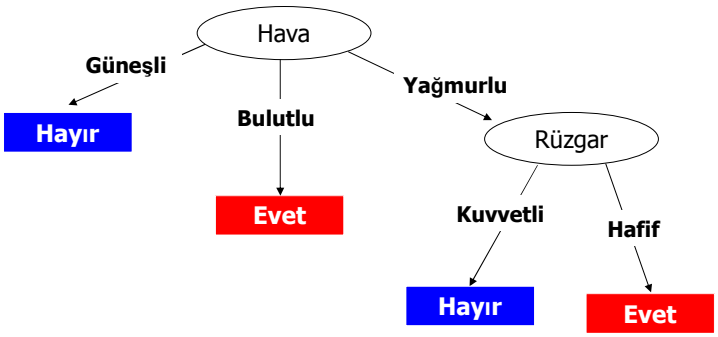
Sonradan budama işlemi ağaç oluşturulduktan sonra devreye girer. Alt ağaçları silerek yaprak oluşturma, alt ağaçları yükseltme, dal kesme şeklinde yapılabilir.

Ağaç Budama



Hata toleransı %33 seçilirse "Nem" düğümünün alt dallarındaki "Evet" oranı %30'dur. Bu yüzden "Nem" düğümü budanıp yerine "Hayır" yaprağı konur.

Ağaç Budama



Sınıflandırma ve Regresyon Ağaçları (CART)

CART karar ağaçlarının temel prensibi herbir düğümde ağacı iki dala ayırmasıdır. En çok bilinen iki algoritması:

- Twoing algoritması
- Gini algoritması

MATLAB Uygulaması

```
>edit C4_5_ornek.m
```

Data1 isimli dataset yüklenerek üzerinde C 4.5 algoritması deneyi yapılmaktadır. Haritalama için kullanılan Matlab grafik komutlarını ve `C4-5` komutu üzerinde farklı deneyler yapılarak kodlar irdelenmelidir.

ÖDEVLER

Aşağıdaki CART algoritmalarını MATLAB'de hazırlayınız.

- Twoing
- Gini