4.3 Karar Ağaçları (ağaç algoritmaları)

Ağaç tabanlı öğrenme algoritmaları, en çok kullanılan ve denetimli öğrenme yöntemlerinden biri olarak düşünülmektedir. Ağaç tabanlı yöntemler, yüksek doğruluk, kararlılık ve yorumlanma kolaylığına sahiptir. Doğrusal modellerin aksine doğrusal olmayan ilişkileri de oldukça iyi eşleyebilirler. Sınıflandırma veya regresyon, elde edilen her türlü sorunun çözümünde uyarlanabilirler. Karar ağaçları, rastgele orman, gradyan güçlendirme gibi yöntemler, her türlü veri bilimi probleminde yaygın şekilde kullanılmaktadır

Karar ağacı öğrenmesi, endüktif(inductive) çıkarım için en yaygın kullanılan pratik yöntemlerden birisidir. Karar ağacı öğrenmesi, öğrenilen fonksiyonun bir karar ağacı tarafından temsil edildiği, kesikli değerli hedef fonksiyonlarını yaklaştırmak için kullanılan bir yöntemdir. Karar ağacı, sınıflandırma problemlerinde çoğunlukla kullanılan bir denetimli öğrenme algoritmasıdır (önceden tanımlanmış bir hedef değişkene sahiptir). Hem kategorik hem de sürekli giriş ve çıkış değişkenleri için çalışır. Öğrenilen ağaçlar insan okunabilirliğini artırmak için if-then kural setleri olarak temsil edilebilirler. Karar ağacı, bir ağaç yapısı biçiminde sınıflandırma veya regresyon modelleri oluşturur. Bir veri kümesini daha küçük ve daha küçük alt kümelere bölerken, aynı zamanda ilişkili bir karar ağacı aşamalı olarak geliştirilir. Nihai sonucu, karar düğümleri ve yaprak düğümleri olan bir ağaçtır. Bir karar düğümü, iki veya daha fazla dallara sahiptir. Yaprak düğüm bir sınıflandırma veya kararı temsil eder. Bir ağaçtaki en üstteki karar düğümü, kök düğüm olarak adlandırılan en iyi belirleyiciye karşılık gelir. Karar ağaçları hem kategorik hem de sayısal verileri işleyebilir. Karar ağaçlarını şöyle ikiye ayırabiliriz:

- **Kategorik Değişken Karar Ağacı:** Kategorik hedef değişkeni olan Karar Ağacı, kategorik değişken karar ağacı olarak adlandırılır. Sınıflandırma Karar Ağaçları da denilebilir.
- **Sürekli Değişken Karar Ağacı:** Karar Ağacı sürekli hedef değişkenine sahipse, Sürekli Değişken Karar Ağacı olarak adlandırılır. Regresyon Karar Ağaçları da denilebilir.

Karar ağaçları da diğer ağaç veri yapıları gibidir ve aynı terminolojiyi kullanmaktadır. Kısaca karar ağaç terminolojine bakacak olursak:

- **Kök Düğüm:** Tüm örneği temsil eder ve bu düğüm daha sonra iki veya daha fazla kümeye ayrılır.
- **Parçalama**: Bir düğümün iki veya daha fazla alt düğümlere bölünmesi işlemidir.
- **Karar Düğümü:** Bir alt düğüm başka alt düğümlere bölünürse, karar düğümü olarak adlandırılır.
- **Yaprak Düğümü:** Bölünmeyen düğümlere Yaprak veya Terminal düğümü denir.
- **Budama**: Karar düğümünün alt düğümlerini kaldırdığımızda, bu işleme budama denir. Yani parçalama işleminin tersi diyebiliriz.
- **Alt Ağaç:** Tüm ağacın bir alt kısmı şube veya alt-ağaç olarak adlandırılır.
- **Ana ve Çocuk Düğümü:** Alt düğümlere ayrılmış olan bir düğüme, alt

düğümlerin ana düğümü ve alt düğümlerine de çocuk düğümü adı verilir.

Karar Ağaç algoritmalarından bazılarını şöyle sıralayabiliriz, ID3, C4.5, C5.0 ve CART.

Tek karar ağacından daha iyi tahmin edici performans elde etmek için çeşitli karar ağaçlarını birleştiren topluluk yöntemleri vardır. Bunlara ağaç topluluk algoritmaları denir. Ağaç topluluk algoritmalarının ana amacı bir grup zayıf öğrenicinin bir araya gelerek güçlü bir öğrenici topluluk oluşturmaktır.

Topluluk karar ağaçlarını gerçekleştirmek için birkaç teknik vardır, bunlar:

1. Torba (bagging)

2. Arttırma (boosting)

Bagging (Bootstrap Aggregation), bir karar ağacının varyansını azaltmak istediğimiz zaman kullanılır. Buna örnek olarak Rastgele Orman algoritması verilebilir.

Boosting, bir öngörü koleksiyonu oluşturmak için kullanılan diğer bir topluluk tekniğidir. Bu teknikte öğreniciler, erken öğrenenlerin basit modellerini verilerle uyuşturduktan sonra hatalar için verileri analiz ederek sırayla öğrenmeye dayanır. Bu tekniğe örnek olarak Gradient Boosting örnek verilebilir.

4.3.1 Gini Dizini

Gini indeksi(dizini) veya Gini katsayısı, İtalyan istatistikçi Corrado Gini tarafından 1912'de geliştirilen istatistiksel bir ölçüdür. Katsayı, 0 (%0) ile 1 (%100) aralığındadır; 0, mükemmel eşitliği temsil eder ve 1, mükemmel eşitsizliği temsil eder. Gini indeksi, rastgele seçilen bir öğenin ne sıklıkta yanlış tespit edildiğini ölçmek için kullanılan bir metriktir. Düşük gini indeksi olan bir özellik tercih edilmelidir. Gini indeksi kategorik hedef değişkeni için başarılı veya başarısız olarak çalışır. Gini indeks, yalnızca ikili bölmeleri (binary: 1 veya 0) gerçekleştirir ve yüksek gini indeksi homojenliği artırır. CART (Sınıflama ve Regresyon Ağacı) ikili bölmeler oluşturmak için gini yöntemini kullanır.

$$\text{Gini Index} = 1 - \sum_{j} P_j^2$$

4.3.2 Entropi

Entropi, rasgele bir değişkenin belirsizliğinin ölçüsüdür, örneklerin keyfi bir koleksiyonunun saf olmayanlığını karakterize eder. Entropi ne kadar yüksek olursa elde edilen bilgi de o kadar fazla olur. Sezgisel olarak, belirli bir olayın öngörülebilirliğinden bahseder.

if a random variable x can take N different value ,the i^{th} value x_i with probability $p(x_i)$,we can assocoate the folloeing entropy with x:

$$H(x) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i)$$

4.3.3 Bilgi Kazanımı

Entropi, tipik olarak, eğitim örneklerini daha küçük alt gruplara bölmek için bir karar ağacında bir düğümü kullandığımızda değişir. Bilgi kazancı, entropideki bu değişimin bir ölçüsüdür.

Definition: Suppose S is a set of instances, A is an attribute, S_v is the subset of s with A = v and Values(A) is the set of all possible of A,then

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$|S|$ denotes the size of set S