13. Ders - Makine Öğrenmesi

18 Eylül 2025 Perşembe 09:19

Entropi (Entropy)

Tanım: Bir veri kümesinin düzensizlik veya belirsizlik ölçüsüdür.

Decision Tree'deki Rolü: Bir düğümdeki verilerin ne kadar "karışık" (impure) olduğunu ölçer. Hedef değişkenin (sınıf etiketlerinin) ne kadar rastgele dağıldığını gösterir.

Değer Aralığı: 0 ile 1 arasındadır.

Entropi = 0: Düğüm tamamen saftır. Tüm veri noktaları aynı sınıftandır (Örn: Hepsi "Evet").

Entropi = 1 (veya max): Düğüm tamamen karışıktır. Veri noktaları sınıflar arasında eşit olarak dağılmıştır (Örn: %50 "Evet", %50 "Hayır").

Formül :

$$H(P) = -\sum_{x \in C} P(x) \log P(x)$$

Bilgi Kazanımı (Information Gain)

Tanım: Bir özelliği (feature) kullanarak veriyi bölmenin sonucunda entropide sağlanan azalmadır. Bir karar kuralının ne kadar "faydalı" olduğunun ölçüsüdür.

Decision Tree'deki Rolü: Ağacı oluşturmak için en iyi özelliği ve bölünme noktasını seçmek için kullanılan ana ölçüttür. Algoritma, Bilgi Kazanımı en yüksek olan özelliği seçerek ilerler.

"Bu soruyu sormak, belirsizliği (entropiyi) ne kadar çok azaltıyor?"

Formül:

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

Özet:

Entropi: Problemin ne kadar zor olduğunu (belirsizlik).

Bilgi Kazanımı: Bir özelliğin bu zorluğu çözmede ne kadar iyi olduğunu (belirsizlikteki azalma).
Decision Tree algoritması, her adımda Bilgi Kazanımını maksimize ederek en saf alt kümeleri (yaprakları) oluşturmaya çalışır.

Gini Index

Farklı sınıfa düşme ihtimalidir.

Tıpkı entropi gibi, bir düğümün saflığını (purity) veya karışıklığını (impurity) ölçen bir metriktir.

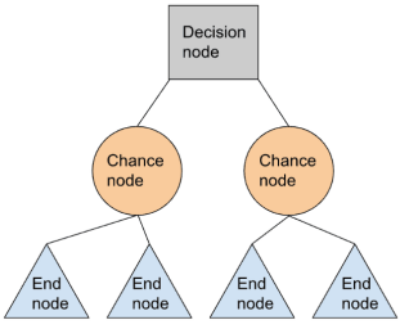
Decision Tree'deki Rolü: Bilgi Kazanımı (Information Gain) yerine de kullanılabilen ve genellikle daha hızlı hesaplanan bir "düzensizlik ölçütü'dür. Temel amaç, bu indeksi *minimize etmektir*.

Değer Aralığı: 0 ile 1 arasındadır (0 ile 0.5 arasında da görülebilir).

Gini = 0: Düğüm tamamen saftır. Tüm veri noktaları aynı sınıftandır (Mükemmel bölünme).

Gini ~ 0.5 (veya max): Düğüm tamamen karışıktır. Veri noktaları sınıflar arasında eşit olarak dağılmıştır (Örn: 2 sınıf için 0.5).

$$Gini = 1 - \sum_j p_j^2$$



Decision Tree (Karar Ağacı)

Gözetimli öğrenme için kullanılan, bir dizi kurala dayalı kararı temsil eden ağaç benzeri bir yapıdır.
Veriyi, bir dizi basit "evet/hayır" sorusuna (kararlara) dayanarak alt gruplara (karar düğümlerine) böler. Tıpkı "20 Soru" oyunu gibi çalışır.

Bileşenleri:
Kök Düğüm (Root Node): Ağacın en tepesindeki, tüm veriyi içeren ilk düğüm.
İç Düğümler (Internal Nodes): Bir özelliği (feature) test eden karar noktaları.
Yapraklar/Dallar (Leaves/Branches): Kararın sonucunu (sınıf etiketini veya tahmin edilen değeri) gösteren son düğümler.

Ağaç algoritma Türleri :

1. CART (Classification and Regression Trees)

Hem sınıflandırma (Classification) hem de regresyon (Regression) problemleri için kullanılabilen, çok amaçlı bir karar ağacı algoritmasıdır. İkili (binary) bölünme yapar (her düğüm evet/hayır sorusu gibi iki dala ayrılır). Sınıflandırmada Gini Indeksi, regresyonda Varyans (MSE) kullanır.

2. MARS (Multivariate Adaptive Regression Splines)

Esnek bir regresyon modelidir. Doğrusal regresyon ile karar ağacının bir melezidir (hybrid). Veriyi parçalara (spline'lar) böler ve her parça için basit doğrusal modeller oluşturur. Doğrusal olmayan ilişkileri modellemekte çok başarılıdır.

3. Conditional Inference Trees

İstatistiksel hipotez testlerine (p-değeri) dayalı bir karar ağacı algoritmasıdır. Aşırı öğrenmeyi (overfitting) azaltmak için değişken seçimini ve bölünme noktalarını istatistiksel anlamlılık testiyle belirler. CART'tan daha az yanlı (less biased) kabul edilir.

4. CHAID (Chi-Square Automatic Interaction Detection)

Özellikle kategorik değişkenlerle çalışmak için tasarlanmış bir sınıflandırma algoritmasıdır. Çoklu dallanmaya izin verir (ikiden fazla dal). Bölünmeleri belirlemek için Ki-Kare (Chi-Square) istatistik testini kullanır.

5. ID3 (Iterative Dichotomiser 3)

Sadece sınıflandırma için kullanılan, karar ağacı algoritmalarının temel ve eski bir türüdür. Çoklu dallanma yapabilir. Bölünme kriteri olarak Bilgi Kazanımı (Information Gain - Entropi) kullanır. Sürekli sayısal özellikleri ve ağaç budamayı desteklemez.

SPLIT MANTIĞI

- **Gini :** Daha hızlı, genellikle CART algoritmalarda production ortamında kullanılır.
- **Entropy :** Daha hassas, ama algoritma içeriğinden maliyetli.
- **Max_depth :** Veri büyüklüğü, özellik sayısı, sınıf dağılımları, hedef performansı, Grid search /Cross Validation teknikleriyle belirlenmesi daha gerçekçi olur.
- **Min_samples_split :** Bir düğüm bölünebilmesi için gereken örnek sayısı.
- **Min_samples_leat =** Yaprakta kalması gereken minimum örnek sayısı.
- Kategorik verilerde yüksek kardinaliteli özellikler bir kolona ağaç takılır ve anlamsız bölmeler gerçekleşir. Bu tip kolonlarda önceden gruplamak (binning) ya da one-hot-encoding yerine target encoding kullanılır.

- Overfitting ihtimalinde çok derin ağaçlarda pruning (budama) ile fazla dalları sonradan kesmek gerekir. Random forest veya Gradient Boosting algoritmalarında tek ağaç yerine esnemble yöntemleri kullanmak daha mantıklı olur. (Ağaç tabanlı topluluk yönetimi XGBoost, CatBoost, LightGBM)