Çalışma Notları: Sınıflandırma: Temel Kavramlar (Bölüm 8.1 - 8.2.2)

18 Kasım 2025

1 Bölüm 8: Sınıflandırma: Temel Kavramlar

Sınıflandırma (Classification), önemli veri sınıflarını tanımlayan modelleri (**sınıflandırıcıları**) çıkaran bir veri analizi türüdür. Bu modeller, **kategorik** (ayrık, sıralanmamış) sınıf etiketlerini tahmin eder.

Uygulama alanları arasında dolandırıcılık tespiti, hedefli pazarlama ve tıbbi teşhis yer alır.

1.1 Temel Kavramlar

1.1.1 Sınıflandırma Nedir?

Sınıflandırmadaki temel amaç, kategorik etiketleri tahmin etmek için bir model oluşturmaktır. Bu kategoriler arasındaki sıralamanın bir anlam ifade etmediği ayrık değerler (örneğin tedaviler A, B, C) ile temsil edilebilir.

1.1.2 Sınıflandırmaya Genel Yaklaşım

Sınıflandırma süreci iki temel adımdan oluşur:

1. **Model Oluşturma (Öğrenme):**
    - Önceden tanımlanmış veri sınıflarını veya kavramları tanımlayan bir sınıflandırıcı (model) oluşturulur.
    - Model oluşturmak için kullanılan verilere **eğitim seti** (training set) denir.

- Eğitim setindeki her veri örneği (tuple), bir uzmandan gelen veya veriden türetilen **ilişkili bir sınıf etiketine** sahiptir.
- Sınıf etiketlerinin bilinmesi nedeniyle bu adım, **denetimli öğrenme** (supervised learning) olarak da bilinir. Bu, etiketlerin bilinmediği **denetimsiz öğrenme** (kümeleme) yönteminin tersidir.
- Türetilmiş model; sınıflandırma kuralları (EĞER-O ZAMAN), karar ağaçları, matematiksel formüller veya sinir ağları gibi çeşitli şekillerde temsil edilebilir.

2. **Modelin Kullanımı:**

   - Oluşturulan sınıflandırıcı, yeni veya daha önce görülmemiş veri örneklerini sınıflandırmak için kullanılır.
   - Kullanım öncesinde, sınıflandırıcının **tahmini doğruluğu** (predictive accuracy) tahmin edilmelidir.
   - Doğruluğu ölçmek için eğitim örneklerinden bağımsız olan bir **test seti** (test set) kullanılır.
   - Eğer modelin doğruluğu eğitim seti kullanılarak ölçülürse, bu tahmin genellikle **iyimser** olur (aşırı uyum/overfit nedeniyle).
   - Modelin doğruluğu kabul edilebilir düzeydeyse, yeni veri örneklerini sınıflandırmak için kullanılabilir.

1.2 Karar Ağacı Türetme

Karar ağacı türetimi, sınıflandırma için popüler bir yöntemdir. Karar ağacı, eğitim setindeki verilerin tutarlı sınıflara ayrılmasına izin veren bir akış şeması benzeri yapıdır.

- **İç düğümler** (internal nodes) bir öznitelik üzerindeki testleri gösterir.
- **Dallar** (branches) testin sonuçlarını temsil eder.
- **Yaprak düğümleri** (leaf nodes) ise sınıf etiketlerini veya sınıf dağılımlarını temsil eder.

Karar ağaçları oluşturulurken, en iyi ayrımı sağlayan özniteliği seçmek için **öznitelik seçim ölçütleri** (attribute selection measures) kullanılır.

1.2.1 Karar Ağacı Türetme (Temel Algoritma)

Karar ağacı algoritmaları genellikle **açgözlü** (greedy) bir yaklaşımla çalışır: Ağacı, veriyi **yukarıdan aşağıya özyinelemeli bölme** (top-down recursive partitioning) yoluyla oluştururlar.

Önemli Algoritmalar:

- **ID3** (Iterative Dichotomiser).
- **C4.5** (ID3'ün ardılı).
- **CART** (Classification and Regression Trees), ikili karar ağaçlarının üretimini tanımlar.

Durdurma Koşulları (Temel Algoritma İçin):

1. Bir düğümdeki tüm örnekler aynı sınıfa aitse.
2. Test edilecek başka öznitelik kalmamışsa.
3. Dalın bölmesinde hiç örnek kalmamışsa.

1.2.2 Öznitelik Seçim Ölçütleri

Öznitelik seçim ölçütleri, eğitim verilerindeki örnekleri en iyi şekilde ayrı sınıflara bölen **bölme kriterini** (splitting criterion) seçmek için kullanılan **sezgisel** (heuristic) yöntemlerdir. En iyi puana sahip öznitelik, **bölme özniteliği** olarak seçilir.

1. Bilgi Kazancı (Information Gain) Bilgi Kazancı, bilgi teorisine dayanır ve en az bilgiyi gerektiren özniteliği seçer. D veri kümesini sınıflandırmak için gereken beklenen bilgi (entropi) miktarı ($Info(D)$) şu şekilde hesaplanır:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

A özniteliği kullanılarak bölme yapıldıktan sonra kalan beklenen bilgi miktarı ($Info_A(D)$) hesaplanır.

Bilgi Kazancı ($Gain(A)$**)** şu şekilde hesaplanır:

$$Gain(A) = Info(D) - Info_A(D)$$

En yüksek bilgi kazancını sağlayan öznitelik, bölme özniteliği olarak seçilir. Bilgi Kazancı, çok sayıda farklı değere sahip öznitelikleri kayırma eğilimindedir.

2. Kazanç Oranı (Gain Ratio) C4.5 algoritmasında kullanılan Kazanç Oranı, Bilgi Kazancı'nın önyargısını düzeltmek için tasarlanmıştır. Bir normalleştirme faktörü olan **Bölme Bilgisi** ($SplitInfo_A(D)$) kullanılır:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2(\frac{|D_j|}{|D|})$$

Kazanç Oranı ($GainRatio(A)$) formülü:

$$\text{GainRatio}(A) = Gain(A)/SplitInfo_A(D)$$

Maksimum kazanç oranına sahip öznitelik seçilir.

3. Gini İndeksi (Gini Index) CART sistemi tarafından kullanılan Gini İndeksi, bir D veri bölümünün **saflığını** (impurity) ölçer.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

(p_i, D'deki bir örneğin C_i sınıfına ait olma olasılığıdır).

Öznitelik A üzerindeki ikili bölme durumunda (D'yi D_1 ve D_2'ye ayırır), ağırlıklı Gini İndeksi hesaplanır:

$$Gini_A(D) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$$

En düşük Gini İndeksine (yani saflıkta en büyük azalmaya) sahip olan öznitelik seçilir.

Çok Değişkenli Bölmeler (Multivariate Splits) Bazı öznitelik seçim ölçütleri, birden fazla özniteliğin kombinasyonuna dayanan **çok değişkenli bölmeleri** dikkate alır. Bu, **öznitelik yapılandırması** (feature construction) olarak da bilinir.