Kümeleme Performans Metrikleri

Atıl Samancıoğlu

Giriş

Kümeleme algoritmaları veriyi önceden belirlenmiş etiketler olmadan gruplara ayırır. Bu grupların ne kadar başarılı oluşturulduğunu değerlendirebilmek için çeşitli metriklere ihtiyaç duyarız. Denetimsiz öğrenmede "başarı"yı ölçmek zor olduğundan, bu metrikler genellikle **küme içi benzerlik** ve **kümeler arası ayrışma** kavramlarına dayanır.

Bu dökümanda en çok kullanılan üç metriği ele alacağız:

- Silhouette Skoru
- Davies-Bouldin Skoru
- Calinski-Harabasz Skoru

Her bir metriği sırayla matematiksel altyapısıyla birlikte açıklayacağız.

1 Silhouette Skoru

Silhouette skoru, her veri noktasının kendi kümesiyle ne kadar iyi eşleştiğini ve diğer kümelerden ne kadar ayrıldığını ölçer.

Tanım

Her veri noktası için iki değer hesaplanır:

- $a(i)$: Nokta i'nin kendi kümesindeki diğer noktalara olan ortalama mesafesi.
- $b(i)$: Nokta i'nin ait olmadığı kümelerden en yakın olanına olan ortalama mesafesi.

Bu iki değerle silhouette skoru şöyle hesaplanır:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Bu değer -1 ile 1 arasında değişir:

- $s(i) \approx 1$: Nokta doğru kümeye atanmıştır.
- $s(i) \approx 0$: Sınırda kalmış, iki kümeye de yakın olabilir.
- $s(i) \approx -1$: Muhtemelen yanlış kümeye atanmıştır.

Tüm veriler için ortalama $s(i)$ hesaplanarak genel skor elde edilir.

Avantajları

Silhouette skoru:

- Küme içi sıkılığı ve kümeler arası ayrılığı birlikte değerlendirir.
- Negatif skorlar yanlış kümelenmeye işaret eder.
- Küme sayısı seçiminde (örneğin K-Means) yol gösterici olabilir.

2 Davies-Bouldin Skoru

Davies-Bouldin (DB) skoru, kümeler arası benzerliği değerlendirerek bir kümeleme modelinin ne kadar ayrık olduğunu ölçer.

Tanım

- Her küme için, küme içi yayılım S_i:

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - \mu_i\|$$

  Burada C_i, i. küme; μ_i kümenin merkezi.

- Her iki küme için ayrım:

$$R_{ij} = \frac{S_i + S_j}{\|\mu_i - \mu_j\|}$$

- Her küme için en kötü durumda olan diğer kümeyle benzerliği:

$$D_i = \max_{j \neq i} R_{ij}$$

- Tüm kümeler için ortalama alınır:

$$DB = \frac{1}{k} \sum_{i=1}^{k} D_i$$

Not: Daha düşük DB skoru daha iyi kümelenme demektir (çünkü kümeler arası ayrım daha fazladır).

Avantajları

- Sayısal olarak yorumlaması kolaydır.
- Küme sayısı arttıkça genellikle azalır; bu yüzden aşırı kümelenmeye dikkat edilmelidir.

3 Calinski-Harabasz Skoru

Calinski-Harabasz (CH) skoru, kümeler arası varyansın kümeler içi varyansa oranını ölçer. Bu metrik aynı zamanda *Variance Ratio Criterion* olarak da bilinir.

Tanım

- Küme merkezlerinin genel merkeze olan uzaklıklarının toplamı (inter-cluster dispersion):

$$B = \sum_{i=1}^{k} |C_i| \cdot \|\mu_i - \mu\|^2$$

  Burada μ tüm verinin ortalaması.

- Küme içi uzaklıkların toplamı (intra-cluster dispersion):

$$W = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

- CH skoru:

$$CH = \frac{B/(k-1)}{W/(n-k)}$$

  Burada n toplam örnek sayısıdır.

Not: Daha yüksek CH skoru daha iyi kümelenmeyi ifade eder.

Avantajları

- Doğrudan varyans oranı üzerinden çalışır.
- Küme sayısını belirlemede oldukça etkilidir.
- Skorun boyutları ve örnek sayısı dikkate alınarak normalize edilmiştir.

Özet

Metrik	Amaç	Skor Aralığı	Tercih Edilen
Silhouette	Ayrım + sıkılık	-1 ila 1	Yüksek olması
Davies-Bouldin	Küme benzerliği	$[0, \infty)$	Düşük olması
Calinski-Harabasz	Varyans oranı	$[0, \infty)$	Yüksek olması

Table 1: Kümeleme değerlendirme metriklerinin karşılaştırması

Sonuç

Bu metrikler, kümeleme kalitesini ölçmek için vazgeçilmezdir. Denetimsiz öğrenmede doğru modeli bulmak, bu metriklerle daha sistematik hale gelir. Uygulamada genellikle birkaç metriğin birlikte yorumlanması tavsiye edilir.