

Veri Madenciliği Dersi Vize Ödevi

Kümeleme Algoritmaları

İsmail YILMAZ

02215076078

Makine Öğreniminde Algoritmalar

- Sahip olduğunuz verilere bağlı olarak makine öğrenimine üç farklı yaklaşım vardır.
- 1. Denetimli öğrenme:** Denetimli öğrenme algoritmaları, giriş verileri ve hedef çıktılar arasındaki ilişkiyi modellemek için kullanılır. Bu tür algoritmalar, eğitim veri setlerinde ki örneklerden öğrenme yapar ve ardından yeni girişler için tahminler yapabilir.
 - 2. Yarı denetimli öğrenme:** Yarı denetimli öğrenme, hem etiketli hem de etiketsiz veri kullanarak bir modelin eğitildiği bir öğrenme paradigmadır. Genellikle, etiketli veri sınırlıdır veya etiketleme süreci maliyetli olduğu için sınırlı bir şekilde gerçekleştirilebilmektedir.

Makine Öğreniminde Algoritmalar

- 3. Denetimsiz öğrenme:** Denetimsiz makine öğrenimi, herhangi bir etiketlenmiş çıktı verisi olmadan algoritma giriş verilerini verdiğiniz durumu ifade eder. Algoritma daha sonra kendi başına, veriler içindeki ve arasındaki kalıpları ve ilişkileri tanımlar.

Denetimsiz öğrenme Algoritmaları:

- Kümeleme
- Birliktelik kuralını öğrenme
- Olasılık yoğunluğu
- Boyutsallık azaltma

Kümeleme Algoritmaları

- **Yoğunluğa dayalı:**

- Veri setindeki yoğun bölgeleri tanımlayarak ve bu yoğun bölgeler arasındaki düşük yoğunluklu alanları belirleyerek kümeleme yaparlar

- **Dağıtım tabanlı:**

- Veri noktalarının olasılık dağılımına dayalı olarak kümeleme yaparlar. Bu tür algoritmalar, veri setindeki benzer özelliklere sahip veri noktalarını belirli bir dağılım ölçütüne göre gruplar.

Kümeleme Algoritmaları

- **Centroid tabanlı:**

- Kümeleme algoritmaları, veri noktalarını belirli bir uzaklık ölçüsü kullanarak küme merkezine atayan algoritmalarlardır. Bu algoritmalar, küme merkezlerini güncelleyerek ve veri noktalarını belirli bir kritere göre bir araya getirerek kümeleme işlemini gerçekleştirirler.

- **Hiyerarşik tabanlı:**

- Hiyerarşik tabanlı kümeleme algoritmaları, veri noktalarını hiyerarşik bir ağaç yapısı içinde kümelere ayırma veya birleştirme prensibi üzerine kurulmuş algoritmalarlardır. Bu algoritmalar, kümeleme sonuçlarını bir ağaç şeklinde temsil ederler.

K-Means Algoritması

- K-means kümelemesi en sık kullanılan kümeleme algoritmasıdır.
- Merkez tabanlı bir algoritmadır. Bu algoritma, bir küme içindeki veri noktalarının varyansını en aza indirmeye çalışır.
- K-means veri noktalarının tümünü yenilediği için, küçük veri kümelerinde daha kullanışlıdır.

DBSCAN Algoritması

- DBSCAN, gürültü içeren uygulamaların yoğunluğa dayalı bölgesel olarak kümelenmesi anlamına gelir.
- Farklı bölgelerdeki veri noktalarının yoğunluğuna göre keyfi biçimde şekillendirilmiş kümeler bulur. Yüksek yoğunluklu kümeler arasındaki aykırı değerleri tespit edebilmek için bölgeleri düşük yoğunluklu alanlara göre ayırır.
- Karmaşık verilerde daha iyidir.

Gaussian Mixture Algoritması

- Bir veri setini bir veya daha fazla Gauss dağılımı (normal dağılım) ile temsil etmeye çalışan bir olasılık tabanlı bir modeldir.
- GMM, veri setindeki karmaşık olasılık dağılımlarını modellere uygun bir şekilde yakalayabilir ve bu dağılımları kullanarak veri setini kümeleyebilir. Her küme bir Gauss dağılımına karşılık gelir.

BIRCH Algoritması

- BIRCH büyük veri setlerini ele almak üzere tasarlanmış bir kümeleme algoritmasıdır.
- BIRCH, özellikle bellek sınırlamalarına ve çok sayıda veri noktasına sahip veri setlerine uyum sağlamak üzere optimize edilmiştir.
- Algoritma, veri setini bir ağaç yapısı içinde temsil eder ve bu yapıyı kullanarak veriyi hiyerarşik bir şekilde kümelendirir.

Affinity Propagation Algoritması

- Bu algoritma, veri noktaları arasındaki "affinity" (benzerlik) ölçüsüne dayanarak özerk olarak öğrenme yeteneğine sahiptir ve küme merkezlerini ve kümeleri belirlemeye çalışır.
- Affinity Propagation, özellikle küme sayısını önceden belirlemek zorunda kalmadan çalışabilmesi ve farklı büyüklükte ve şekillerde kümeleri tanıyabilmesi açısından avantajlıdır.

Mean-Shift Algoritması

- Bir veri setindeki yoğunluk maksimumlarını bulmak ve bu maksimumlar etrafında kümeler oluşturmak için kullanılan bir kümeleme algoritmasıdır.
- Bu algoritma, veri noktalarını yoğunluk grafiği üzerinde hareket ettirerek kümeleri belirler.
- Mean-Shift, özellikle belirgin küme yapılarına sahip olan veri setlerinde ve küme sayısının önceden bilinmediği durumlarda kullanışlıdır.

OPTICS Algoritması

- Veri kümeleme ve küme yapısını anlama amacıyla kullanılan bir kümeleme algoritmasıdır.
- OPTICS, özellikle büyük ve gürültülü veri setleri üzerinde etkili çalışabilir.
- Bu algoritma, veri setindeki küme yapılarını hiyerarşik bir şekilde tanımlar ve kümeleme sonuçlarını bir sıralama listesi şeklinde sunar.

Agglomerative Hierarchy Algoritması

- Bir veri setindeki nesneleri birleştirerek veya bölerek hiyerarşik bir ağaç yapısı oluşturan bir kümeleme algoritmasıdır.
- Bu algoritma, agglomerative (birleştirici) bir yaklaşım kullanır ve genellikle alt düzeyde küçük kümelerle başlar, ardından bu kümelere birleştirme işlemi uygulayarak hiyerarşik bir yapı oluşturur.

Kaynaklar

- <https://aws.amazon.com/tr/compare/the-difference-between-machine-learning-supervised-and-unsupervised/#:~:text=Denetimsiz%20makine%20%C3%B6%C4%9Fre nimi%2C%20herhangi%20bir,denetimli%20%C3%B6%C4%9Frenme%20teknikleri%20t%C3%BCrleri%20vard%C4%B1r.>
- <https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/>
- <https://www.linkedin.com/pulse/denetimli-ve-denetimsiz-%C3%B6%C4%9Frenme-algoritmalar%C4%B1-abdulsamet-k%C4%B1l%C4%B1%C3%A7/?originalSubdomain=tr>