ELM472 Makine Öğrenmesinin Temelleri

Ödev5

Ayşe Serra Şimşek
asimsek2019@gtu.edu.tr
[2]Elektronik Mühendisliği Bölümü, GTÜ, Kocaeli, Türkiye

I. GİRİŞ

Bu çalışmanın konusu, Ödev3'te kullanılan Salinas haritasındaki her bir vektörün boyutunun PCA kullanarak boyutlarının küçültülmesi işlemidir. Başarımın değerlendirilmesindeki metriklerin kullanımı açıklanmıştır.

II. TEORİK BİLGİ

PCA (Principal Component Analysis - Temel Bileşen Analizi), çok değişkenli verilerin boyutunu azaltan bir yöntemdir. Temel amacı, veri setindeki değişkenlik miktarını koruyarak, veriyi daha az sayıda, ancak orijinal verinin büyük bir kısmını temsil eden değişken setine dönüştürmektir.

PCA'nın çalışma prensibi şu adımları içerir:

- Veri Merkezleme: Veri setindeki her bir özellikten, o özelliğin ortalaması çıkarılarak veri merkezlenir. Bu, veri setinin merkezini (0,0) noktasına taşır.

- Kovaryans Matrisi Hesaplama: Merkezlenmiş veri setinin kovaryans matrisi hesaplanır. Kovaryans matrisi, değişkenler arasındaki ilişkiyi gösterir.

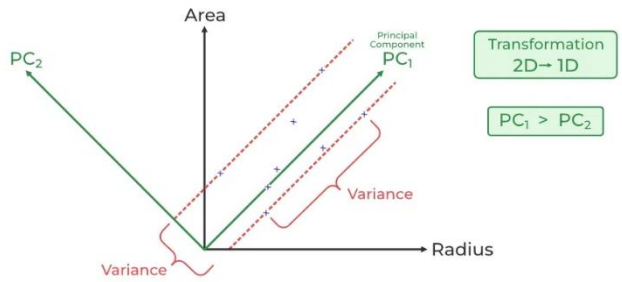
- Eigenvalues ve Eigenvectors Hesaplama: Kovaryans matrisinin eigenvalues (özdeğerler) ve eigenvectors (özvektörler) hesaplanır. Eigenvectors, veri setindeki değişkenliği temsil eden yeni bir bazı tanımlar.

- Eigenvalues'e Göre Sıralama: Eigenvalues, veri setindeki değişkenliğin miktarını temsil eder. Büyük eigenvalues'e sahip eigenvectors, daha fazla değişkenliği temsil eder. Eigenvalues'e göre sıralama yapılır.

- İstenen Boyuta Göre Seçim: İstenen boyuta kadar olan en büyük eigenvalues'e sahip eigenvectors seçilir. Bu, boyutun azaltılacağı boyut sayısını ifade eder.

- Yeni Veri Matrisi Oluşturma: Seçilen eigenvectors kullanılarak yeni bir veri matrisi oluşturulur. Bu matris, orijinal veri setinin daha düşük boyutlu bir temsilini içerir.

Yani PCA'nın temel hedefi, veri setindeki değişkenliği korurken, bu değişkenliği daha az sayıda özellikle temsil edebilen bir alt uzayı bulmaktır. Bu, özellikle büyük boyutlu veri setlerinde, veriyi daha anlaşılır ve işlenebilir bir formata getirmek için yaygın olarak kullanılır.



Şekil 1. Principal Component Analysis (PCA)

PCA uygulamanın bir veri setine birkaç önemli etkisi vardır:

- Boyut Azaltma: PCA, veri setindeki boyutu azaltarak, orijinal verinin büyük bir kısmını temsil eden daha az sayıda özellik elde etmenizi sağlar. Bu, veri setinizin karmaşıklığını azaltır ve analizini kolaylaştırır.

- Değişkenliği Koruma: PCA, veri setindeki değişkenliği korur. Eigenvalues ve eigenvectors aracılığıyla, veri setindeki değişkenlik miktarını temsil eden ana bileşenleri belirler. En önemli bilgiyi koruyarak boyut azaltma işlemini gerçekleştirir.

- Korelasyonu Azaltma: PCA, orijinal değişkenler arasındaki korelasyonu azaltır. Bu, veri setindeki birbirine bağlı değişkenleri daha bağımsız hale getirir ve analiz sırasında aşırı öğrenme (overfitting) gibi sorunları önler.

- Gürültüyü Azaltma: PCA, veri setindeki gürültüyü azaltabilir. En düşük eigenvalues'e sahip olan bileşenler, genellikle veri setindeki gürültüyü temsil eder. Bu bileşenlerin ihmal edilmesi, daha temiz bir temsile yol açabilir.

- Model Performansını Artırma: Boyut azaltma, özellikle makine öğrenimi modelleri gibi algoritmaların çalışma süresini azaltabilir. Aynı zamanda, gereksiz özellikleri çıkardığı için model performansını artırabilir.

Ancak, PCA uygulamanın bazı olumsuz etkileri de olabilir:

- Bilgi Kaybı: Boyut azaltma sırasında, veri setindeki bazı detaylar ve özellikler kaybolabilir. Özellikle küçük eigenvalues'e sahip bileşenlerin ihmal edilmesi, bu tür bilgi kayıplarına yol açabilir.

- Overfitting Sorunu: PCA, modeldeki özellik sayısını azaltarak overfitting sorununu azaltabilir. Ancak, gereksiz boyut azaltma, veri setindeki önemli özellikleri kaçırabilir ve modelin genelleme yeteneğini etkileyebilir.

III. ÇALIŞMA

Bu çalışmada, Ödev3'te yapılan kümelendirme işlemleri, Salinas haritasına PCA işlemi uygulandıktan sonra yapılmıştır. Önceki ödevde sadece K-means kümelendirmesi uygulandığı için burada da yapılan işlemler sonrasında K-means ile kıyaslanmıştır.

- Öncelikle salinas.mat dosyası "loadmat" komutu ile yüklenmiş ve 105 bandında görselleştirilmiştir.
- Daha sonra salinas.mat dosyasına PCA işlemi uygulanmıştır. Veri setini PCA kullanarak boyut küçültme işlemi, önce veriyi düzenlemeyi, ardından kovaryans matrisini hesaplamayı, bu matrisin özdeğerlerini ve özvektörlerini bulmayı ve en sonunda istenilen boyutu seçmeyi içerir.
- Bu noktada, N, boyut değerinin belirlenmesi üzerine iki yaklaşım vardır:

1. **Toplam varyansın belirli bir yüzdesini koruma**: Örneğin, %95 varyansı korumak için yeterli boyutu seçebilirsiniz.
2. **Eşik değeri belirleme**: İlk eigenvalues'ların toplamını üzerinden bir eşik değeri belirleyerek bu eşik değeri geçene kadar boyutları seçebilirsiniz.

Yapılan çalışmada birinci yaklaşım ele alınmıştır. %95'lik varyans değerini korumak için özdeğerler kullanılarak bir algoritma kurulmuş ve N değeri belirlenmiştir. Algoritma Şekil.2'de gösterilmiştir.

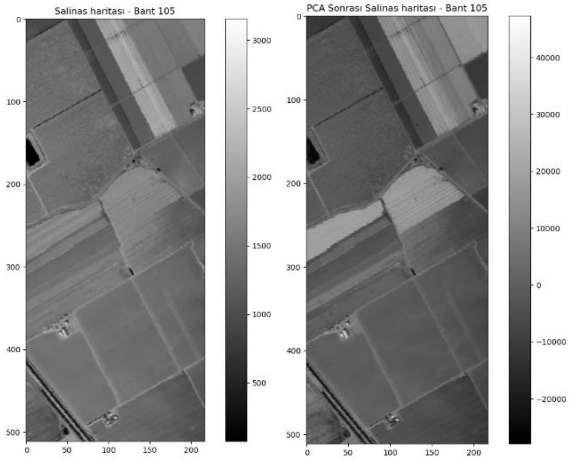
```
total_variance = np.sum(eigenvalues)
explained_variance = np.cumsum(eigenvalues) /
total_variance
N = np.argmax(explained_variance >= 0.95) + 1
```

Şekil 2. Boyut indirgenme algoritması

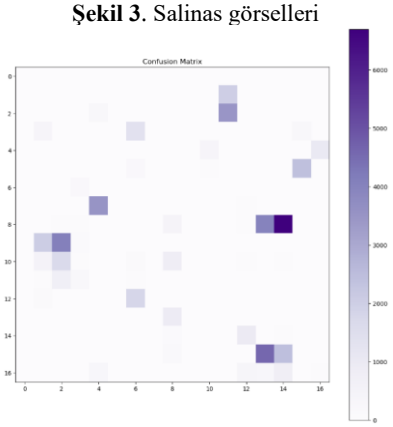
- N değeri verilen algoritma ile saptandıktan sonra, Ödev3'te yazılmış olan K-means kümelendirme algoritması ile kümelendirilip görselleştirilmiştir.

- Ardından salinas_gt.mat dosyasına göre gerekli metrik değerleri hesaplanıp kıyaslanmıştır

IV. SONUÇLAR



Şekil 3. Salinas görselleri



Şekil 4. Confusion Matrix

Yapılan çalışma sonucu Clustering Error Rate: 0.48719218029953915 olarak bulunmuştur. PCA ile boyut indirgeme sonrasında, gerçek verilere göre kaybedilen veriler değerlendirildiğinde yaklaşık %50lik başarılı bir kümelendirme işlemi yapıldığı söylenebilmektedir.
Varyans değerinin %95'lik korunması durumunda N=2 olarak belirlenmektedir.

KAYNAKÇA

[1] E. Alpaydin, Introduction to Machine Learning, 3. bs. Cambridge, MA, USA: MIT Press, 2014.
[2] E. Alpaydin, Introduction to Machine Learning, 3. bs. Cambridge, MA, USA: MIT Press, 2014.