BLM6103 – Olasılık, Rastgele Değişkenler ve Stokastik Süreçler

2023-2024 Güz Yarıyılı

Ödev – 3

- "Olasılık, R.D. ve Stokastik Prosesler" isimli Google Classroom grubundan güncel duyuruları takip etmek için ackaraca@yildiz.edu.tr mail adresine mail atabilirsiniz.

- Bu ödev, toplam 2 uygulamadan oluşmaktadır. Soru çözümlerini bu kapak sayfasının arkasına ekleyerek pdf haline dönüştürünüz. Tek bir pdf dosyası halinde mail yukarıdaki mail adresine gönderiniz.

- Dönem boyunca 3 ödev, 1 arasınav ve 1 final yapılacaktır. Bu sebeple, ödevler oldukça önemlidir. Eksiksiz ve kopyasız bir şekilde yapmanız önerilmektedir.

- Birbirinin tamamen aynısı olan çözümlerde hak edilen puan kişi sayısına bölünerek hesaplanacaktır. Bu sebeple, kendi çözümlerinizi paylaşmayınız.

- Puanların sorulara göre dağılımı aşağıdaki gibidir:

Problem	Soru 1 & Soru 2	Soru 3	Soru 4	Soru 5
Puanlar	40p	20p	20p	20p

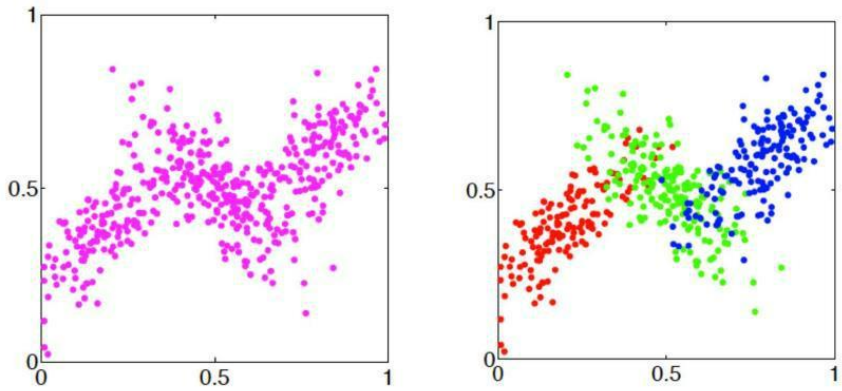
- Bilgisayar üzerinden çözülecek sorular için kod ve ekran çıktıları paylaşılmalıdır. Derste MATLAB üzerinden kodlar paylaşılsa da siz ödevleri farklı programlama dillerinde yapabilirsiniz.

- Soruların tamamı diğer sayfada paylaşılmıştır.

- SON GÖNDERİM TARİHİ 01 OCAK 2024 OLUP ÖDEVLER GOOGLE CLASSROOM ÜZERİNDEN YÜKLENECEKTİR.

KONU: GAUSS KARIŞIM MODELİ KULLANARAK KÜMELEME YAKLAŞIMI

Önbilgi: Kümeleme yaklaşımlarından en sık bilineni k-ortalamalar yöntemidir. Bu yöntemin bazı problemlerini çözebilmek için literatürde beklenti enbüyükleme (Expectation Maximization, EM) yoluyla Gauss Karışım Modeli (Gaussian Mixture Model, GMM) kullanılmaktadır. Bu yöntem aşağıdaki gibi etiketsiz (pembe işaretli) bir veriye uygulandığında sağdaki gibi kümeler oluşturulmaktadır. Buna göre aşağıdaki sorulara cevap veriniz.



1- GMM'in algoritmasını ve denklemlerini yazarak detaylıca paylaşınız.

2- K-Ortalamalar yöntemlerinin algoritmalarını yazarak karşılaştırınız. Karşılaştırmada kullanılan eşitlikleri paylaşarak olasılık dersimiz açısından inceleyiniz. Buna göre, GMM, K-Ortalamalar yöntemindeki hangi sorunlara nasıl çözüm sağlamaktadır?

3- Aşağıda parametreleri paylaşılan çok değişkenli Gauss dağılımı kullanarak 10000 örnek içeren veriler üretiniz ve scatter çizimi olarak yukarıdaki gibi çizdiriniz. Burada, iki küme sayısı varsayımı ile veri oluşturulmaktadır.

   a. $\mu 1 = [2, 1]$, $\mu 2 = [2, -1]$, $\Sigma 1 = \Sigma 2 = [2, -1; -1 , 1]$,
   b. $\mu 1 = [2, 1]$, $\mu 2 = [-2, -1]$, $\Sigma 1 = [2, -1; -1 , 1]$, , $\Sigma 2 = [2, 2; 2 , 3]$

4- Yukarıda elde edilen her iki veri için de kovaryans ve varyans değerlerini ayrı ayrı yorumlayınız.

5- Elde edilen veriler için K-Ortalamalar ve GMM yöntemlerini eşit şartlar altında çalıştırarak sonuçlarını karşılaştırınız. Sonuçlarını olasılıksal açıdan yorumlayınız.

NOT: GMM yönteminin kodunu siz yazmalısınız fakat alt işlemler için bazı fonksiyonlardan yardım alabilirsin.

Başarılar dilerim.

Doç. Dr. Ali Can KARACA

GAUSS KARIŞIM MODELİ KULLANARAK KÜMELEME YAKLAŞIMI

BLM6103 – Olasılık, Rastgele Değişkenler ve Stokastik Süreçler

Recep Furkan Koçyiğit

22501048

furkan.kocyigit@std.yildiz.edu.tr

Bilgisayar Mühendisliği Bölümü

Elektrik Elektronik Fakültesi, Yıldız Teknik Üniversitesi

Özet

Bu ödevde, Gauss Dağılımı kullanılarak verilen ortalama vektörü, kovaryans matrisi ve örneklem sayısına göre veri üretilmiştir. K-Ortalamalar(K-Means) ve Gauss Karışım Modeli (Gaussian Mixture Model, GMM) algoritmaları yazılmış ve oluşturulan veri kullanılarak, eşit şartlarda eğitilip karşılaştırılması istenmiştir. Bu kapsamda sistemin geçekleştirilmesi için programlama dili olarak Python kullanılmıştır.

Giriş

Ödev kapsamında veri kümeleri Gauss dağılımı kullanılarak oluşturulmuştur. Ödev kapsamında verilen ortalama ve kovaryans matrislerine göre oluşturulan veriler yorumlanmıştır. Gauss Karışım Modeli ve K-Ortalamalar algoritmaları yazılmıştır. Daha sonrasında oluşturulan veri kümesindeki verilerin %80'i eğitim ve %20'i test için ayrılmıştır. Her iki algoritma bu verilerle eğitilmiş ve test edilmiştir.

Yöntem

Gauss Dağılımı Kullanılarak Veri Üretimi:

Ödev kapsamında Gauss dağılımı kullanılarak verilen ortalama vektörü ve kovaryans matrislerine göre 10.000 adet veri oluşturulması istenmiştir. 5.000 adet iki adet veri birleştirilerek oluşturulmuştur.

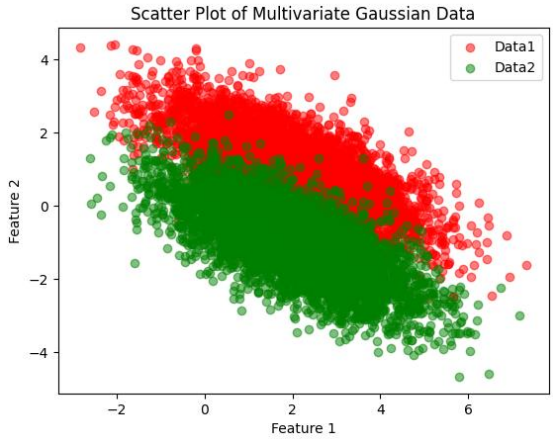
İlk örnek için ortalama vektörleri ve kovaryans matrisleri şu şekildedir:

$$\mu_1 = [2,1]$$

$$\mu_2 = [2,-1]$$

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

Oluşturulan verinin nokta grafiği aşağıdaki gibidir:



Kırmızı ve yeşil verilerin kovaryansları -1'dir. Yani, Feature1 artarken Feature2 azalmaktadır. Kısaca, aralarında ters bir ilişki vardır. Varyanslarına baktığımızda her iki veri için de 2 ve 1 olarak gözükmektedir. Varyansı 2 olan özelliğin varyansı 1 olan özelliğe göre daha geniş olarak yayıldığını bu veriye bakarak söyleyebiliriz. Bu, Feature1 in ortalama etrafında ±√2 genişliğinde, Feature2'nin ise ±1 genişliğinde değişiklik gösterdiği anlamına gelir.

Kırmızı ve yeşil verilerin birleşimi sonucunda kovaryans -1 ve varyanslar 2 bulunmuştur. İlk örnek için Feature1 ve Feature2 arasında ters bir ilişki vardır.

İkinci örnek için ortalama vektörleri ve kovaryans matrisleri şu şekildedir:

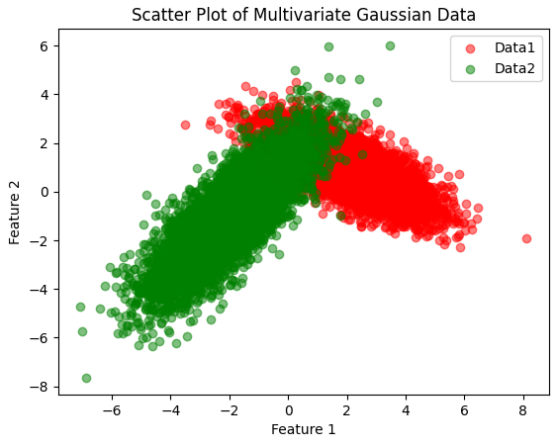
$$\mu_1 = [2,1]$$

$$\mu_2 = [-2,-1]$$

$$\Sigma_1 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 2 & 2 \\ 2 & 3 \end{bmatrix}$$

Oluşturulan verinin nokta grafiği aşağıdaki gibidir:



Kırmızı renkli veri bir önceki parametrelerle aynı şekilde oluşturulmuştur. Kırmızı ve yeşil verilerin kovaryansları 2'dir. Yani, Feature1 artarken Feature2 de artmaktadır. Kısaca, aralarında doğru bir ilişki vardır. Yeşil renkli verinin varyansına baktığımızda varyans değerleri 2 ve 3 olarak gözükmektedir. Feature1 özelliğinin varyansı 3 ve Feature2 özelliğinin varyansı 2 olduğu için Feature1 daha geniş olarak yayıldığını bu veriye bakarak söyleyebiliriz. Bu, Feature1 in ortalama etrafında ±√3 genişliğinde, Feature2'nin ise ±√2 genişliğinde değişiklik gösterdiği anlamına gelir.

Kırmızı ve yeşil verilerin birleşimi sonucunda kovaryans 2.46 bulunmuştur. İlk örnek için Feature1 ve Feature2 arasında doğru bir ilişki vardır. Varyanslar ise 5.9 ve 3 çıkmıştır. Feature1 in kendi içindeki varyansının Feature2 den yüksek olduğunu söyleyebiliriz.

Gauss Karışım Modeli Algoritması:

Gauss Karışım Modeli (GMM), veri setinin birçok Gauss dağılımının bir karışımı olarak modellemek için kullanılan bir istatistiksel modeldir. Bu model, özellikle veri setinde yapıları keşfetmek ve karmaşık dağılımları yakalamak için kullanılır. Beklenti en büyükleme (Expectation Maximization, EM) algoritması, GMM'nin parametrelerini tahmin etmek için kullanılan bir optimizasyon algoritmasıdır.

Gauss dağılım fonksiyonunun olasılık yoğunluk fonksiyonu(pdf) aşağıdaki gibidir:

$$G(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Burada μ ortalamayı, σ^2 varyansı göstermektedir.

Gauss Karışım Model algoritması şu şekildedir:

1) **Her küme için ilk değer ataması:**
   - Ağırlıkların atanması. π_k
   - Ortalama vektörlerinin atanması. μ_k
   - Kovaryans matrislerinin atanması. $\sum_k$
2) **Beklenti Aşaması:**
   - Veri noktalarının her birinin her küme için ait olma olasılıklarını tahmin eder.
   - Bayes kuralını kullanarak her veri noktası için her Gauss kümesinin sorumluluğunun hesaplanması.

$$\gamma_{n,k} = \frac{\pi_k G(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j G(x_n|\mu_j,\Sigma_j)}$$

3) **Maksimumlaştırma Aşaması:**
   - Beklenti aşamasındaki tahminlere dayanarak Gaussian bileşenlerin parametrelerini günceller.
   - Ağırlıkların beklentiye göre güncellenmesi.

$$\pi_k^{new} = \frac{1}{N}\sum_{n=1}^{N} \gamma_{n,k}$$

   - Ortalama vektörlerinin beklentiye göre güncellenmesi.

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} \gamma_{n,k} x_n}{\sum_{n=1}^{N} \gamma_{n,k}}$$

   - Kovaryans matrislerinin beklentiye göre güncellenmesi.

$$\Sigma_k^{new} = \frac{\sum_{n=1}^{N} \gamma_{n,k}(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T}{\sum_{n=1}^{N} \gamma_{n,k}}$$

4) **Tekrar Aşaması:**
   Parametreler önemli ölçüde değişmediği durumda veya maksimum iterasyon sayısına ulaşılana kadar 2. ve 3. aşamaların tekrar edilmesi.

K-Ortalama Kümeleme Algoritması:

K-ortalama kümeleme yöntemi, N adet veri nesnesinden oluşan bir veri kümesini giriş parametresi olarak alarak k adet kümeye bölme amacı güder. Temel hedef, gerçekleştirilen bölümlendirme işlemi sonucunda elde edilen kümelerin içindeki benzerliklerin maksimum, kümeler arası benzerliklerin ise minimum olmasını sağlamaktır.

K-Ortalama kümeleme algoritması şu şekildedir:

1. **Rastgele Başlatma:**
   Rastgele k başlangıç küme merkezinin seçilmesi.

2. **Merkezlere atama:**
   Her veri noktasının en yakın merkeze atanması.

$$J(c^{(i)}, \mu_k) = \min_k \left|\left|x^{(i)} - \mu_k\right|\right|^2$$

   Burada J maliyet fonksiyonu, $c^{(i)}$, $x^{(i)}$ 'ye en yakın ağırlık merkezinin indeksi ve μ_k k-inci ağırlık merkezidir.
3. **Merkezlerin Güncellenmesi:**
   Her kümeye atanan veri noktalarının ortalamasına göre ağırlık merkezlerinin yeniden hesaplanması.

$$\boldsymbol{\mu_k} = \frac{\mathbf{1}}{|\boldsymbol{C_k}|} \sum_{\boldsymbol{i \in C_k}} \boldsymbol{x^{(i)}}$$

   Burada C_k, k'inci kümeye atanan veri noktaları kümesidir. k'inci kümeye atanan veri noktaları kümesidir.
4. **Tekrar Aşaması:**
   Merkezlerin değişmediği durumda veya maksimum iterasyon sayısına ulaşılana kadar 2. ve 3. aşamaların tekrar edilmesi.

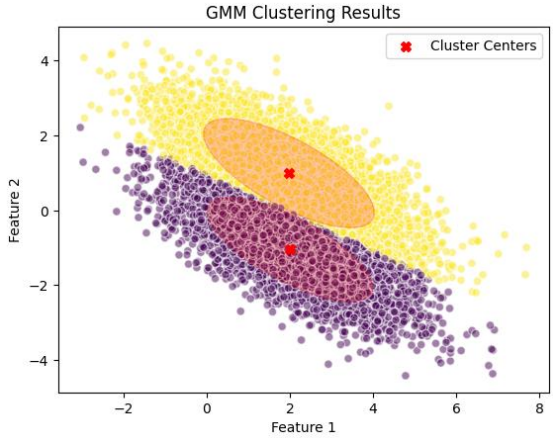
Gauss Karışım Modeli (GMM) ve K-Ortalamalar Karşılaştırması

- GMM, veri noktalarını kümelere olasılık dağılımı aracılığıyla yumuşak bir şekilde atarken, K-Ortalamalar ise bir noktayı bir kümeye direkt yani sert atama yapar. Bu esneklik, GMM'nin karmaşık desenleri ve örtüşen kümeleri daha iyi yakalamasına olanak tanır.
- GMM, özellikler arasındaki kovaryansı modelleyerek, kümeleme özellikler arasındaki ilişkileri ve kümelerin eliptik şekillerini yakalayabilir. K-Ortalamalar ise küresel ve eşit kovaryanslı şekillere dayandığı için bu esnekliği sağlayamaz.
- K-Ortalamalar, küresel kümeleri varsayar, GMM ise kümelerin farklı şekil ve yönlere sahip olabileceğini modelleyebilir.
- GMM, K-Ortalamalara kıyasla aykırı değerlere daha az duyarlıdır. Çünkü sadece uzaklık metriklerine güvenmek yerine tüm olasılık dağılımını dikkate alır.

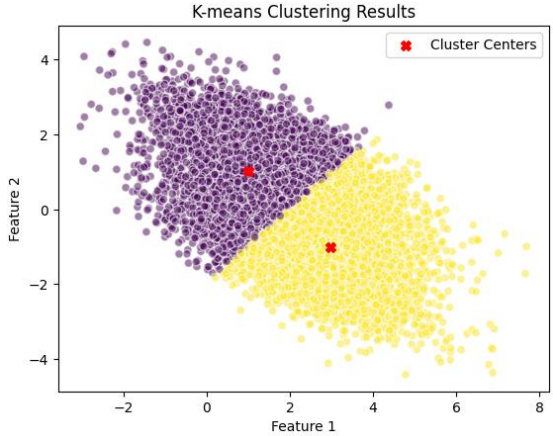
Deneysel Sonuçlar

İlk Veri:

Gauss Karışım Modelini eğittiğimiz ve veriyi iki kümeye ayırmasını istediğimizde sonuçlar aşağıdaki gibidir:



K-Ortalamalar ile elde edilen sonuç ise şu şekildedir:



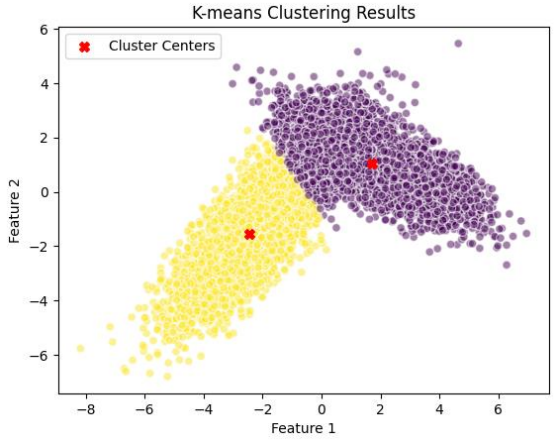
Bu modellerden alınan sonuçlara baktığımızda, veriyi ikiye ayırmaya çalıştıklarında farklı şekilde ayırdıklarını görmekteyiz. Kümeleme merkezlerinin de farklı kullanılan algoritmalara göre farklı noktalarda olduğunu söyleyebiliriz. İlk veriyi Gauss dağılımı için oluşturduğumuz için ilk kümelemenin daha doğru bir sonuç olduğunu söylemek mümkündür.

İkinci Veri:

Gauss Karışım Modelini eğittiğimiz ve veriyi iki kümeye ayırmasını istediğimizde sonuçlar aşağıdaki gibidir:

Bu görsele de baktığımızda ikinci örnek için bizim veriyi verdiğimiz gibi kümelemeye çalıştığını görmekteyiz. Bunun sebebinin bizim veriyi oluştururken ortalama vektörü ve kovaryans matrisi kullanmamız ve GMM'nin de tahmin yaparken bu verileri kullanmasıdır.

K-Ortalamalar ile elde edilen sonuç ise şu şekildedir:



Bu görsele baktığımızda kümelemenin daha doğru olduğu görülmektedir. Veriyi olasılıksal değil, küme merkezlerine göre ayırdığımızda daha doğal bir bölünme görülmektedir.

Sonuç

Bu ödev kapsamında, Gauss dağılımı kullanılarak iki farklı örneklem oluşturulmuş, Gauss Dağılım Modeli(GMM) ve K-Ortalamalar algoritmaları yazılmıştır. Oluşturulan veriler kovaryanslarına göre yorumlanmış, modellerin başarıları karşılaştırılmıştır. Bu algoritmaların başarısının kullanılan veriye göre değişiklik göstereceği görülmüştür. GMM'nin K-Ortalamalara göre hesaplama maliyetinin daha fazla olduğu ve GMM'nin K-Ortalamaların yakalayamayacağı eliptik şekilleri yakalamada daha başarılı olduğu görülmüştür.