



Denetimsiz Öğrenme Temel Bileşen Analizi

M. Bahadır ÇALIŞKAN, M. Furkan ÖLÇER, Ahmet Hakan ÇELİK, Moussa Bane, Yusuf T. ARABACI



İçindekiler

1

Denetimsiz Öğrenme
Ve Boyut İndirgeme

M. Furkan ÖLÇER

2

Temel Bileşen Analizi (PCA)

Ahmet Hakan ÇELİK

3

PCA' nın Çalışma Mantığı

Moussa Bane

4

PCA Algoritması

M. Bahadır Çalışkan

5

PCA Uygulamaları

Yusuf T. ARABACI

6

Sonuç - Sorular





1

Denetimsiz Öğrenme

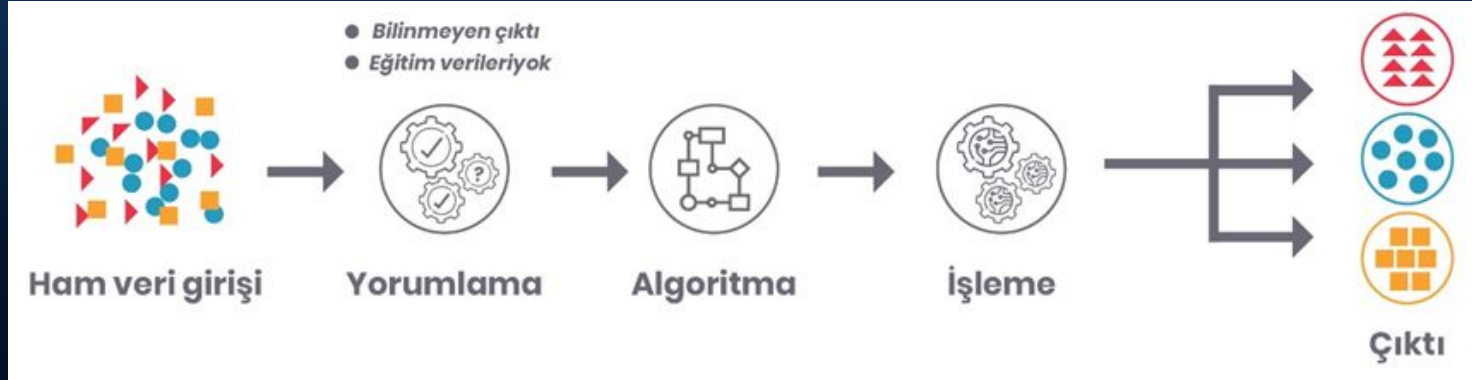
- Boyut indirgeme



Denetimsiz Makine Öğrenmesi (Unsupervised Learning) Nedir ?

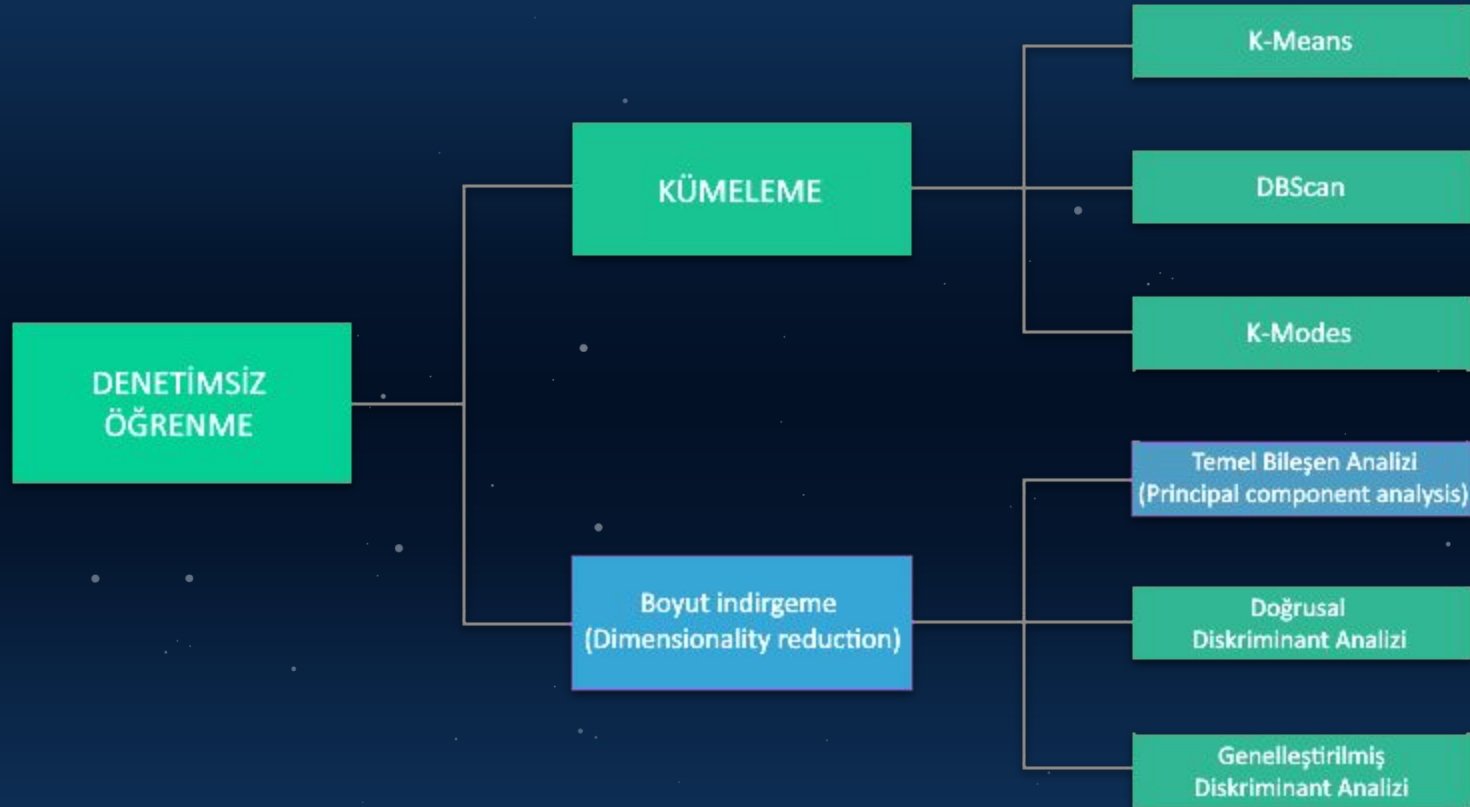
Denetimsiz makine öğrenmesi, algoritmaları eğitmek için kullanılan bilgilerin sınıflandırılmadığı veya etiketlenmediği durumlarda kullanılır. Denetimsiz öğrenme, sistemlerin etiketlenmemiş verilerden gizli bir yapıyı açıklamak için bir işlevi nasıl çıkarabileceğini inceler.





Makine, ham verileri herhangi bir etiketli veri olmadan yorumlayıp ve belirli algoritmalar kullanarak işledikten sonra çıktıları üretmeye çalışır.

Sistem doğru çıktıyı bulamadığında verileri araştırmaya devam eder ve etiketlenmemiş verilerden gizli yapıları açıklamak için veri kümelerinden çıkarımlar yapar.



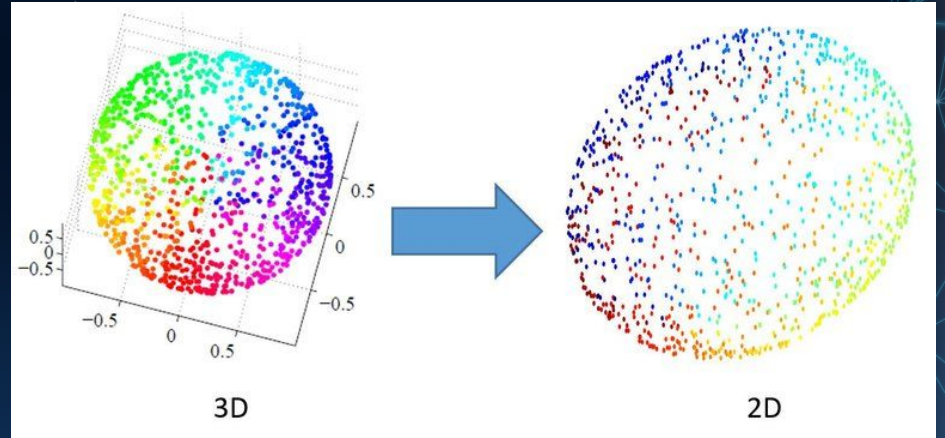
Boyut İndirgeme (Dimensionality Reduction)

Bir veri kümesi için giriş değişkenlerinin veya özelliklerinin sayısına **boyutsallık** denilmektedir. Fazla girdi özelliği genellikle öngörücü modelleme görevini zorlaştırır, daha genel olarak bu durum, boyutsallığın laneti olarak adlandırılmaktadır.

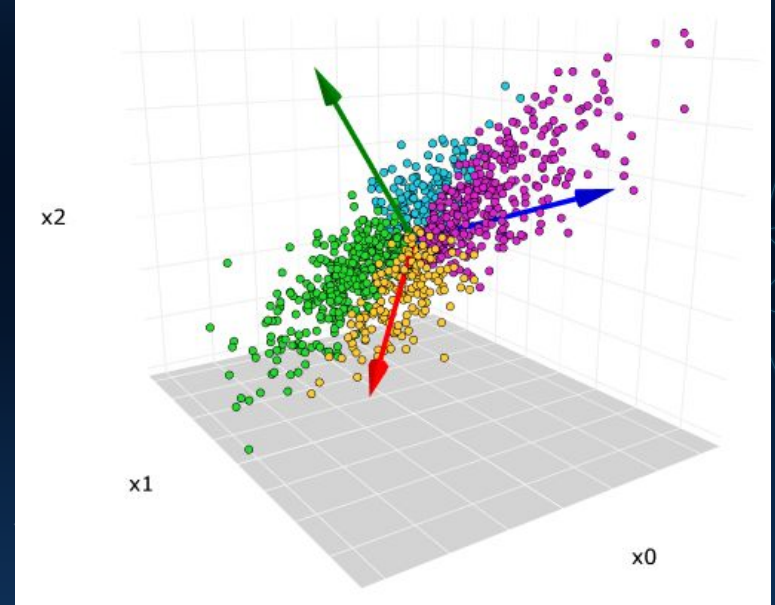
Veri görselleştirme için genellikle yüksek boyutsallık istatistikleri ve boyutsallık azaltma teknikleri kullanılır. Bununla birlikte, bu teknikler, öngörücü modele daha iyi uyması için sınıflandırma veya regresyon veri kümesini basitleştirmek amacıyla uygulamalı makine öğreniminde de kullanılabilir.

Boyut indirgeme yöntemi temel olarak, ele alınan rastgele değişken veya öznelilik sayısını azaltma işlemidir.

Bir dizi temel değişken oluşturularak dikkate alınan rastgele değişkenlerin sayısının azaltılması sağlanır. Bu yönüyle veri bilimi ve makine öğreniminde oldukça önemli bir konumdadır.



- Gerçek hayattaki veriler çok fazla boyuta (özniteliğe) sahip oluyor ve boyut büyüdükçe veri temizlemeden model kurmaya bütün süreçlerde harcamamız gereken **zaman ve kaynaklar** artıyor.
- Ne kadar çok boyuta sahip olunursa görselleştirme de o kadar zorlaştırmaktadır.



- Hemen hemen her veri setinde bazı öznitelikler arasında yüksek korelasyon oluyor ve bu bizim gereksiz bilgiye sahip olmamıza ve modelimizde **overfitting problemine** sebep olabiliyor.

Bu problemleri en aza indirmek için çeşitli algoritmalar kullanılmaktadır.

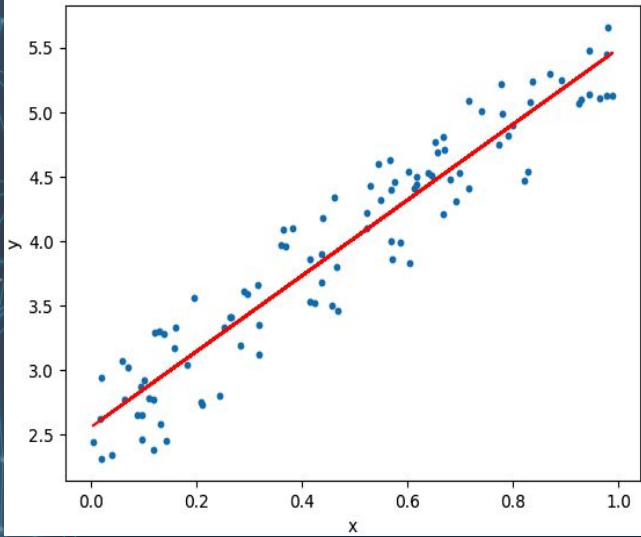
Temel Bileşen Analizi (PCA) yüksek boyutlu bir veri setinin boyutunu azaltmak için kullanılan en yaygın yöntemlerden biridir.



2

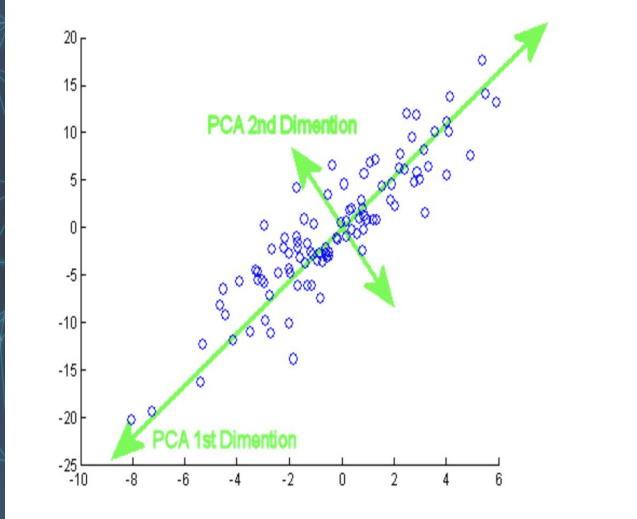
Temel Bileşen Analizi (PCA)

Temel Bileşen Analizi (PCA) Nedir?



Temel Bileşen Analizi (PCA), çok boyutlu uzaydaki bir verinin daha düşük boyutlu bir uzaya izdüşümünü, varyansı maksimize edecek şekilde bulma yöntemidir. Uzayda bir noktalar kümesi için tüm noktalara ortalama uzaklığı en az olan “en uygun doğru” seçilir. Daha sonra bu doğruya dik olanlar arasından yine en uygun doğru seçilerek bu adımlar, yeni bir boyutun varyansı belirli bir eşiğin altına inene kadar tekrarlanır.

Temel Bileşen Analizi (PCA) Nedir?



Bu sürecin sonunda elde edilen doğrular bir doğrusal uzayın tabanlarını oluşturur. Bu taban vektörüne temel bileşen adı verilir. Temel bileşenlerin 3 farklı özelliği vardır ;

1-Verinin temel bileşenleri birbirinden bağımsızdır.

2-Birinci temel bileşen toplam değişkenliği en çok açıklayan bileşendir.

3-Bir sonraki temel bileşen de kalan değişkenliği en çok açıklayan bileşendir.

PCA' nin Amaçları

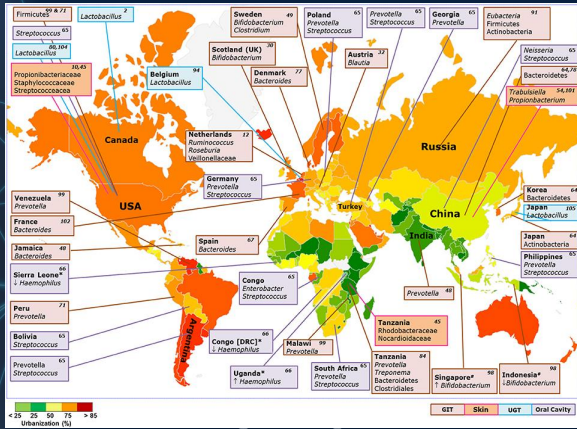
Genel Olarak Temel Bileşen Analizi' nin amaçları ;

- Verilerin boyutunu indirmek (değişken sayısını azaltmak)
- Değişkenler arasındaki ilişki yapısını ortadan kaldırmak
- Diğer istatistiksel analizler için veri toplamak

şeklinde ifade edilebilir.

- GENETİK

Genetik çeşitliliğin coğrafi konum ve etnik kökene göre dağılımı, bir ırkın yaşadığı tarihsel demografik olaylar ve süreçler hakkında geniş bir bilgi kaynağı sağlar. Bununla birlikte kolonizasyon, izolasyon, göç ve karışım gibi süreçlerin doğası ve zamanlaması hakkında çıkarımlar yapmak çok zorlaşabilir. Temel Bileşen Analizi de genetik varyasyonun coğrafi konum ve etnik kökene dağılımındaki yapıyı belirlemek için yaygın olarak kullanılır.



- SAĞLIK

Elektronik Sağlık Hizmeti kayıtları kullanan klinik araştırmalar genellikle çok sayıda değişken sunar. Bu değişkenler sıklıkla birbirleriyle ilişkilidir ve bu da regresyon modellerinde çoklu bağlantıya neden olur. Çoklu bağlantıdan etkilenen tahminlerin büyük standart hataları olabilir ve bu tür tahminler üzerindeki çıkarımı daha az kesin hale getirir.

Bu tip bir sorun klinik çalışmalarda mevcuttur ve bu sorunla başa çıkmak için kullanılan yöntemlerden bir tanesi de Temel Bileşen Analizidir.



- ENERJİ

Günümüzde fosil yakıtlar nedeniyle artan küresel ısınma sorununa karşılık güneş enerjisi benzeri yenilenebilir enerjilere yönelim artmaktadır. Ancak güneş enerjisi sistemlerinin sorunsuz ve de sürekli çalışabilmesi için güneş ışınımının yoğunluğu ile ilgili birkaç dakika önceden bilgi alınmalıdır. Bunun için çeşitli modeller olsa da bu modeller çoğunlukla yüksek hesaplama süresi gerektirir. Hesaplama sürelerini azaltmak amacıyla veri boyutunu küçültülür ve bunun için temel bileşen analizi kullanılır.

Makine Öğrenmesinde Kullanımı

Veri Bilimi çalışmalarında çok sayıda değişken ile çalışılması gerekebilir. Bu durum eğitim(training) süresinin fazla olması, aşırı öğrenme(overfitting) ve çoklu doğrusal bağlantı(multicollinearity) gibi sorunları beraberinde getirir. Hazırlanan modellerin optimum sürede ve performansla çalışması gerekecektir.

Bu problemleri aşmak için değişken seçimi ve boyut indirgeme yöntemleri kullanılabilir. Değişken seçiminde veri setindeki değişken korunur ya da tamamen kaldırılır. Boyut indirgemedede ise mevcut değişkenlerin kombinasyonlarından oluşan yeni değişkenler yaratılarak değişken sayısı azaltılır. Böylece veri setindeki tüm özellikler hala mevcut ancak değişken sayısı azaltılmış olur.



Analizlerde yaşanan bu tip sorunları aşmak için en çok tercih edilen boyut indirgeme yöntemlerinden birisi de Temel Bileşenler Analizidir.

Ayrıca Temel Bileşenler Analizi, yüz tanıma, resim sıkıştırma ve örüntü tanıma gibi alanlarda yaygın olarak kullanılmaktadır.





PCA 'nın Çalışma Mantığı

Principal Component Analysis
(Temel Bileşen Analizi) Çalışma
Mantığı?

Adım Adım PCA Çalışma Mantığı

1. Standardize the range of continuous initial variables...**Standardization**
2. Compute the covariance matrix to identify correlations...**Covariance Matrix Computation**
3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
4. Create a feature vector to decide which principal components to keep... **Feature Vector**
5. Recast the data along the principal components axes

STEP 1: STANDARDIZATION

“Sürekli ilk değişken aralığını standartlaştırma”

Bu adımın amacı, sürekli başlangıç değişkenlerinin aralığını, her birinin analize eşit katkıda bulunacağı şekilde standardize etmektir.

Daha spesifik olarak, PCA'dan önce standardizasyon gerçekleştirmenin kritik olmasının nedeni, PCA'nın başlangıç değişkenlerinin varyansları konusunda oldukça hassas olmasıdır. Yani, başlangıç değişkenlerinin aralıkları arasında büyük farklar varsa, daha geniş aralıklara sahip değişkenler, küçük aralıklara sahip olanlara göre baskın olacaktır (örneğin, 0 ile 100 arasında değişen bir değişken, 0 ile 1 arasında değişen bir değişken üzerinde baskın olacaktır), bu da yanlış sonuçlara yol açacaktır. Dolayısıyla verilerin karşılaştırılabilir ölçeklere dönüştürülmesi bu sorunu önlenebilir.

STEP 1: STANDARDIZATION ⇒ Devam

Matematiksel olarak bu, ortalamayı çıkararak ve her değişkenin her değeri için standart sapmaya bölerek yapılabilir.

$$z = \frac{value - mean}{standard\ deviation}$$

Standardizasyon yapıldıktan sonra, tüm değişkenler aynı ölçeğe dönüştürülmüş olacaktır.

STEP 2: COVARIANCE MATRIX COMPUTATION (KOVARYANS MATRİSİ HESAPLANMASI)

” Korelasyonları belirlemek için kovaryans matrisini hesaplama”

Bu adımın amacı, girdi veri setindeki değişkenlerin birbirlerine göre ortalamadan nasıl farklılaştığını anlamaktır, başka bir deyişle aralarında herhangi bir ilişki olup olmadığını görmektir. Çünkü bazen değişkenler, gereksiz bilgiler içerecek şekilde yüksek oranda ilişkilidir. Dolayısıyla, bu korelasyonları belirlemek için kovaryans matrisini hesaplıyoruz.

STEP 2: COVARIANCE MATRIX COMPUTATION

⇒ Devam

Kovaryans matrisi, başlangıç değişkenlerinin tüm olası çiftleriyle ilişkili kovaryansları girdi olarak içeren bir $p \times p$ simetrik matristir (burada p , boyutların sayısıdır). Örneğin, x , y ve z 3 değişkenli 3 boyutlu bir veri kümesi için kovaryans matrisi, bunun 3×3 matrisi olur:

$$\begin{bmatrix} Cov(x, x) & Cov(x, y) & Cov(x, z) \\ Cov(y, x) & Cov(y, y) & Cov(y, z) \\ Cov(z, x) & Cov(z, y) & Cov(z, z) \end{bmatrix}$$

Covariance Matrix for 3-Dimensional Data

Bir değişkenin kendisi ile kovaryansı onun varyansı olduğundan ($Cov(a, a) = Var(a)$) ana köşegende (Sol üstten sağ alta) aslında her bir ilk değişkenin varyansına sahibiz. Ve kovaryans değişmeli olduğundan ($Cov(a, b) = Cov(b, a)$) kovaryans matrisinin girişleri ana köşegene göre simetriktir, yani üst ve alt üçgen kısımlar eşittir.

STEP 2: COVARIANCE MATRIX COMPUTATION ⇒ Devam

Matris girdileri olarak sahip olduğumuz kovaryanslar, değişkenler arasındaki korelasyonlar hakkında bize ne anlatıyor?

Aslında önemli olan kovaryansın işaretidir:

Pozitif ise: iki değişken birlikte artar veya azalır (korelasyonlu)

Negatif ise: biri azalırken diğeri artar (Ters korelasyonlu)

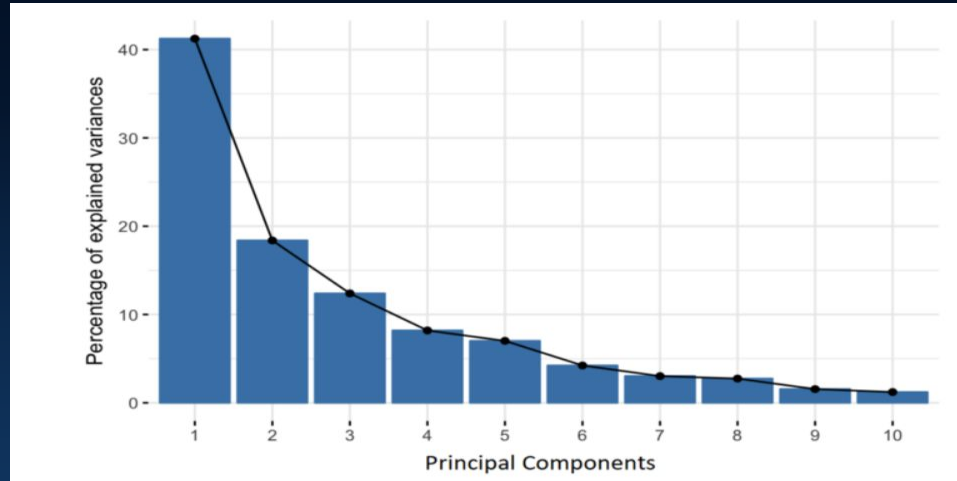
Artık kovaryans matrisinin olası tüm değişken çiftleri arasındaki korelasyonları özetleyen bir tablodan daha fazlası olmadığını bildiğimize göre, bir sonraki adıma geçelim.

STEP 3: ANA BİLEŞENLERİ BELİRLEMEK İÇİN KOVARYANS MATRİSİNİN ÖZVEKTÖRLERİNİ(EIGENVECTORS) VE ÖZDEĞERLERİNİ(EIGENVALUES) HESAPLAMA

Özvektörler ve özdeğerler, verilerin temel bileşenlerini belirlemek için kovaryans matrisinden hesaplamamız gereken lineer cebir kavramlarıdır. Bu kavramların açıklamasına geçmeden önce temel bileşenlerden ne anladığımızı anlayalım.

STEP 3: ⇒ Devam

Temel bileşenler, başlangıç değişkenlerinin doğrusal kombinasyonları veya karışımları olarak oluşturulan yeni değişkenlerdir. Bu kombinasyonlar, yeni değişkenler (yani temel bileşenler) ilintisiz olacak ve ilk değişkenlerdeki bilgilerin çoğu ilk bileşenlere sıkıştırılacak şekilde yapılır. Yani, fikir şu ki, 10 boyutlu veriler size 10 ana bileşen verir, ancak PCA mümkün olan maksimum bilgiyi ilk bileşene, ardından ikinci bileşene maksimum kalan bilgiyi vb.



STEP 3: ⇒Devam

Bilgileri temel bileşenlerde bu şekilde düzenlemek, çok fazla bilgi kaybetmeden boyutsallığı azaltmanıza olanak tanır ve bu, düşük bilgi içeren bileşenleri atarak ve kalan bileşenleri yeni değişkenleriniz olarak kabul etmektir.

Burada anlaşılması gereken önemli bir şey var, ‘temel bileşenlerin daha az yorumlanabilir olduğu ve başlangıç değişkenlerinin doğrusal kombinasyonları olarak oluşturuldukları için herhangi bir gerçek anlamının olmadığıdır’.

Geometrik olarak konuşacak olursak, **ana bileşenler**, verilerin maksimum miktarda varyansı açıklayan yönlerini, yani verilerin çoğunu yakalayan çizgileri temsil eder. Buradaki varyans ve bilgi arasındaki ilişki, bir çizginin taşıdığı varyans ne kadar büyükse, veri noktalarının çizgi boyunca dağılımı o kadar büyük ve bir çizgi boyunca dağılım ne kadar büyükse, o kadar fazla bilgiye sahip olur. Tüm bunları basitçe ifade etmek gerekirse, temel bileşenleri, gözlemler arasındaki farkların daha iyi görülebilmesi için verileri görmek ve değerlendirmek için en iyi açıyı sağlayan yeni eksenler olarak düşünmektir.

Temel bileşenlere sahip olduktan sonra, her bileşenin açıkladığı varyans (bilgi) yüzdesini hesaplamak için her bileşenin özdeğerini özdeğerlerin toplamına böleriz.

STEP 4: FEATURE VECTOR (ÖZELLİK VEKTÖR)

” HANGİ ANA BİLEŞENLERİN TUTULACAĞINA KARAR VERMEK İÇİN
BİR ÖZELLİK VEKTÖRÜ OLUŞTURMA”

Önceki adımda gördüğümüz gibi, özvektörleri hesaplamak ve bunları özdeğerlerine göre azalan düzende sıralamak, temel bileşenleri önem sırasına göre bulmamızı sağlar. Bu adımda yaptığımız şey, tüm bu bileşenleri tutmayı veya daha az anlamlı olanları (düşük özdeğerleri) atmayı seçip kalanlarla Özellik vektörü dediğimiz bir vektörler matrisi oluşturmayı seçmektir.

STEP 4: FEATURE VECTOR \Rightarrow Devam

Dolayısıyla, özellik vektörü, tutmaya karar verdiğimiz bileşenlerin özvektörlerini sütunlar halinde içeren bir matristir. Bu, onu boyut indirgemeye yönelik ilk adım yapar, çünkü n 'nin yalnızca özvektörlerini (bileşenlerini) tutmayı seçersek, son veri kümesi yalnızca p boyuta sahip olacaktır.

Aradığınız şeye bağlı olarak tüm bileşenleri saklamayı veya daha az önemli olanları atmayı seçmek size kalmış. Çünkü, verilerinizi boyutsallığı azaltmaya çalışmadan ilişkisiz yeni değişkenler (temel bileşenler) açısından tanımlamak istiyorsanız, daha az önemli bileşenleri dışarıda bırakmanıza gerek yoktur.

LAST STEP: VERİLERİ ANA BİLEŞENLER EKSENLERİ BOYUNCA YENİDEN DÖKÜMLEME

Önceki adımlarda standardizasyon dışında veriler üzerinde herhangi bir değişiklik yapmadık, sadece ana bileşenleri seçip özellik vektörünü oluşturduk, ancak girdi veri seti her zaman orijinal eksenler cinsindendir (yani, ilk değişkenler).

Son adım olan bu adımda amaç şu ki, kovaryans matrisinin özvektörleri kullanılarak oluşturulan özellik vektörünü kullanmak, verileri orijinal eksenlerden temel bileşenler tarafından temsil edilenlere yeniden yönlendirmektir. Bu, orijinal veri setinin devriğini özellik vektörünün devriğiyle çarparak yapılabilir.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$



4

PCA Algoritması

Principal Component
Analysis(Temel Bileşen Analizi)
Algoritma Adımları



Adım 1 – Veri Seti Elde Et

Öncelikle veri seti elde edilir. Daha sonra elde edilen veri seti X ve Y'ye bölünür. Y doğrulama seti, X ise eğitim seti olarak atanır. Başka bir deyişle X ile öğretilen değerler Y ile teyit edilir.

Adım 2 - Veriler ile Yapını Oluştur

X bağımsız değişkenin 2 boyutlu matrisi alınır. Satırlar veri ögelerini, sütunlar da gerekli ögenin özelliklerini temsil eder. Sütun sayıları boyut sayısı ile eşit miktardadır. Her girişten ilgili sütunun ortalaması çıkarılır, bu işlem her sütun için tekrar edilir.

Adım 3 – Verini Standardize Et

Verilen x sütunlarına bakarak, daha yüksek varyansa sahip özellikler daha düşük varyansa sahip özelliklerden daha mı önemlidir? Yoksa özelliklerin önemi varyanstan bağımsız mıdır? Son durumda önem, özelliğin Y'yi ne kadar iyi tahmin ettiğini ifade eder.

Özelliklerin önemi özelliklerin varyansından bağımsızsa, bir sütundaki her gözlem o sütunun standart sapmasına bölünür.

(Ortalanmış ve standardize edilmiş matrise Z adlandırılması yapılacak.)

```
def standardization(self, data):  
    #In here we subtract the mean and dividing by standard deviation  
    z = (data - np.mean(data, axis = 0)) / (np.std(data, axis = 0))  
    return z
```

Adım 4 - Z'nin Kovaryansını Hesapla

$$\begin{matrix} \begin{bmatrix} x_1^1 - \bar{x}_1 & x_1^2 - \bar{x}_1 & \cdots & x_1^{100} - \bar{x}_1 \\ x_2^1 - \bar{x}_2 & x_2^2 - \bar{x}_2 & \cdots & x_2^{100} - \bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{100}^1 - \bar{x}_1 & x_{100}^2 - \bar{x}_2 & \cdots & x_{100}^{100} - \bar{x}_{100} \end{bmatrix} & \begin{bmatrix} x_1^1 - \bar{x}_1 & x_2^1 - \bar{x}_2 \\ x_1^2 - \bar{x}_1 & x_2^2 - \bar{x}_2 \\ \vdots & \vdots \\ x_1^{100} - \bar{x}_1 & x_2^{100} - \bar{x}_2 \end{bmatrix} \\ 100 \times 2 & 100 \times 2 \end{matrix}$$
$$X_{norm}^T X_{norm}$$

Z matrisinin transpozu alınır (Z^T) ve transpozlu hali kendisi ile çarpılır.

Z kovaryansı = $Z^T * Z$ şeklinde hesaplanır.

Ortaya çıkan matris bir sabite kadar Z'nin kovaryans matrisidir.

Adım 5 - Öz Vektör ve Öz Değeri Hesapla

Özvektörleri ve bunlara karşılık gelen $Z^T Z$ özvektörlerini hesaplanır.

P özvektörler matrisi olmak üzere;

$Z^T Z$ 'nin özdekompozisyonu, $Z^T Z$ 'nin PDP^{-1} 'ye ayrıştırıldığı yerdir.

D ise özdeğeri olan köşegen matrisidir ve köşegen dışında değeri 0'dır.

D'nin köşegenindeki özdeğerler, P'de karşılık gelen değerlerle ilişkilendirilir.

Yani D'nin ilk elemanı λ_1 'dir ve karşılık gelen özvektör P'nin ilk sütunudur.

Bu tüm D değerleri ve P'de denk gelen özvektör karşılıkları için geçerlidir.

PDP^{-1} her zaman bu doğrultuda hesaplanır.

Finding Principal Components

1. find eigenvalues by solving: $\det(\Sigma - \lambda I) = 0$

$$\det \begin{pmatrix} 2.0 - \lambda & 0.8 \\ 0.8 & 0.6 - \lambda \end{pmatrix} = (2.0 - \lambda)(0.6 - \lambda) - (0.8)(0.8) = \lambda^2 - 2.6\lambda + 0.56 = 0$$
$$\{\lambda_1, \lambda_2\} = \frac{1}{2} \left(2.6 \pm \sqrt{2.6^2 - 4 * 0.56} \right) = \{2.36, 0.23\}$$

2. find i^{th} eigenvector by solving: $\Sigma \mathbf{e}_i = \lambda_i \mathbf{e}_i$

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = 2.36 \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} \rightarrow \begin{cases} 2.0e_{11} + 0.8e_{12} = 2.36e_{11} \\ 0.8e_{11} + 0.6e_{12} = 2.36e_{12} \end{cases} \rightarrow e_{11} = 2.2e_{12}$$

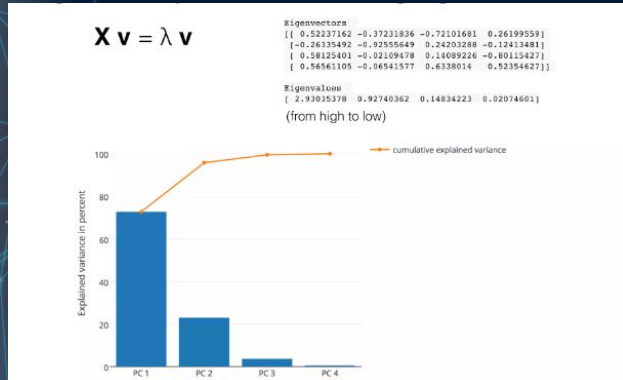
$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix} = 0.23 \begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix}$$

$$\mathbf{e}_1 \sim \begin{bmatrix} 2.2 \\ 1 \end{bmatrix}$$

normalize: $\|\mathbf{e}_i\| = 1$

$$\mathbf{e}_1 = \begin{bmatrix} 0.91 \\ 0.41 \end{bmatrix}$$

Adım 6 - Öz Vektörleri Sırala



Özdeğerler alınıp ($\lambda_1, \lambda_2, \dots, \lambda_p$) büyükten küçüğe sıralanır. Bu şekilde P'deki özvektörler de sıralanmış olur.

(Örneğin λ_2 en büyük özdeğer olsun, o halde P'nin ikinci sütun alınıp en başa konur.)

(Sıralanmış P özvektör matrisine P^* olarak daha sonra ulaşılacak.)

Adım 7 – Yeni Özellikleri Hesapla

$Z^* = ZP^*$ eşitliği hesaplanır. Bu yeni Z^* matrisi X 'in ortalananmış ve standartlaştırılmış halidir ama Z^* 'de her bir gözlem orijinal değişkenlerin kombinasyonudur. Ağırlıklar özvektöre göre hesaplanır.

(P^* içindeki özvektörler bağımsız olduğundan, Z^* 'nin her sütunları da aynı zamanda birbirinden bağımsızdır.)

$$P = \begin{bmatrix} 0.9 & 4.7 & -1.3 & 0.3 \\ -0.4 & 14.0 & 0.4 & -0.1 \\ 1.0 & 0.8 & 0.4 & -1.3 \\ 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix} \rightarrow P^* = \begin{bmatrix} 0.9 & 4.7 \\ -0.4 & 14.0 \\ 1.0 & 0.8 \\ 1.0 & 1.0 \end{bmatrix}$$

$$Z = [-1.0 \quad 1.4 \quad -1.4 \quad -1.3], P^* = \begin{bmatrix} 0.9 & 4.7 \\ -0.4 & 14.0 \\ 1.0 & 0.8 \\ 1.0 & 1.0 \end{bmatrix}$$
$$\rightarrow Z^* = ZP^* = [-4.2 \quad 12.5]$$

Adım 8 - Yeni Setten Önemsiz Özellikleri Çıkar

Yeni setteki özelliklerden hangisi/hangileri tutulmalı bu belirlenir.

Her bir özdeğer kabaca karşılığı olduğu özvektörün önemidir. Bu sebeple açıklanmış olan varyans oranı, tuttuğunuz özelliklerin özdeğerlerinin toplamının, tüm özelliklerin özdeğerlerinin toplamına bölümüdür.

Bunun için 3 farklı yöntem izlenir:

1.Yöntem: Kaç boyut tutulmak istendiğine karar verilir.


2.Yöntem: Her özelliğin varyans oranı hesaplanır. Eşik değeri(Threshold) seçilip eşik değere gelinene kadar özellik eklenir.

3. Yöntem: İkinci metod gibi her özelliğin varyans oranı hesaplanarak başlanır. Özellikler varyans oranına göre sıralanır ve daha fazla özellik tutuldukça açıklanan kümülatif varyans oranı çizilir. Yeni bir özellik eklemenin, önceki özelliğe göre oluşan varyansta önemli bir düşüşe sahip olduğu nokta tespit edilir ve bu noktaya kadar eklenen özellikleri seçerek, kaç özelliğin dahil edileceği belirlenir.



5

PCA Uygulamaları

- Kavramsal bir örnek
 - Uygulamalı bir örnek
 - PCA ile modelleme
- 



PCA Kavramsal Bir Örnek

Temel Bileşen Analizinin kullanımına örnek olarak bir Hisse Senedi Tahminleme modeli verilebilir.

Bir hisse senedi performansının tahmine dayalı bir modelini oluşturmanızın istendiğini varsayalım. Doğru modelleme için stok performansı ile ilgili ölçümlerin tüm boyutlarını göz önünde bulundurmanız gerekir.



PCA Kavramsal Bir Örnek

Hisse seçimi için büyük bir aracı kurum tarafından sağlanan değişkenlerin hızlı bir analizi 30'dan fazla değişken sağlar.

Değişkenlerden bazıları birbiriyle ilişkili olabilir ve bazıları analiz için herhangi bir ek bilgi bile eklemeyebilir. Tüm değişkenler arasındaki korelasyonları sezgisel olarak değerlendirmek finansal okuryazar bir kişi için bile açık olmayabilir.

PCA, bu büyük değişken kümesine uygulanabilir ve orijinal verilerdeki varyansın tamamına yakını temsil eden azaltılmış bir alternatif değişkenler kümesi çıkarabilir. Daha küçük küme daha sonra stok performansı için farklı modelleri değerlendirmek için kullanılabilir.

Değişkenler

- Kapanış fiyatı
- Açılış fiyatı
- Gün içi yüksek
- Gün içi düşük
- Alfa
- Beta
- Finansal oranlar
- İşgücü piyasası
- Konut değişkenleri
- Duygu ölçümleri
- GSYİH ölçümleri
- Enflasyon
- İşsizlik vb.

PCA Uygulamalı Bir Örnek

Görüntü sıkıştırma PCA'nın yaygın bir uygulamasıdır. Şekilde 1200 x 795 piksel çözünürlükte çekilmiş bir ay resmi vardır.

```
moon <- readJPEG("moon.jpg")  
moon_pca <- prcomp(moon, center=FALSE)  
percent_variance = sum((moon_pca$sdev[1:k])^2)/sum((moon_pca$sdev)^2)
```

PCA sonuçları, temel bileşenlerle ilişkili standart sapmaları içeren moon_pca nesnesinde depolanıyor. Kodun son satırındaki k, dikkate alınan temel bileşenlerin sayısını gösterir ve dolayısıyla varyans yüzdesini etkiler.



PCA Uygulamalı Bir Örnek

Şekildeki tablo, k değerine bağlı olarak temel bileşenler tarafından yakalanan yüzde varyansı gösterir (yani, dikkate alınan temel bileşenlerin sayısı)

Bu sonuçlara göre, yalnızca 10 Temel Bileşen seçerek görüntüyü %98 doğrulukla yeniden oluşturabilirsiniz. Bu da, 10/795'lik bir sıkıştırma.

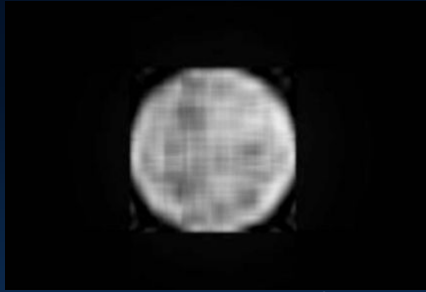
Number of PCs	Percent variance
5	96.3
10	98.1
20	98.9
100	99.9

PCA Uygulamalı Bir Örnek

Orijinal veriler, aşağıdaki kod alıntısı kullanılarak PC'lerden çoğaltılabilir.

```
moon_compressed = moon_pca$x[,1:k] * transpose(moon_pca$rotation[,1:k])
```

5 temel bileşen ve 20 temel bileşen ile yeniden yapılandırılmış görseller aşağıda verilmiştir.



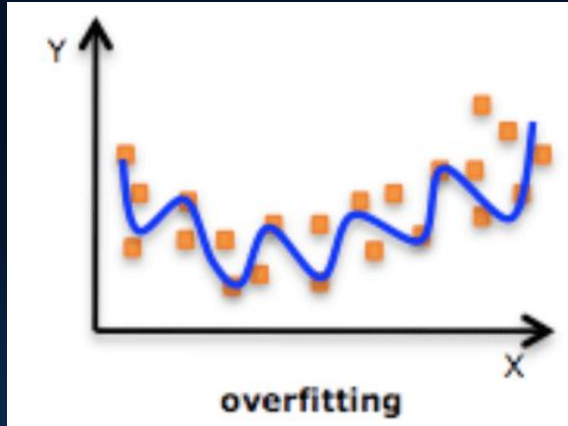
5 PC'li görüntü



20 PC'li görüntü

PCA ile Modelleme

Overfitting

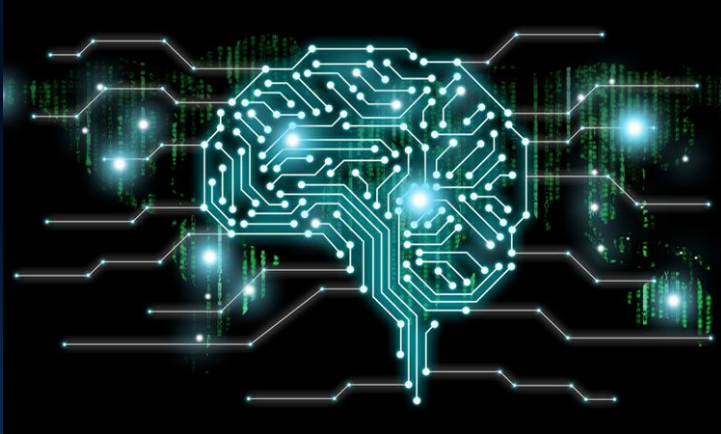


Modelleme yapılırken öznitelik sayısı çok ve yeterli veri olmadığında oluşan modelde overfitting meydana gelebilir. Öznitelik sayısını azaltmak için bazen, öznitelik sayısının azalmasının fazla uydurmayı önleyeceği varsayımı altında PCA kullanılır.

Overfittingi önlemek için PCA kullanmak kötü sonuçlara yol açabilir. Aşırı uydurma sorunu, düzenleştirme (regularization) teknikleri ve daha iyi veri kümeleri kullanılarak ele alınabilir.

PCA ile Modelleme

Model Üretimi



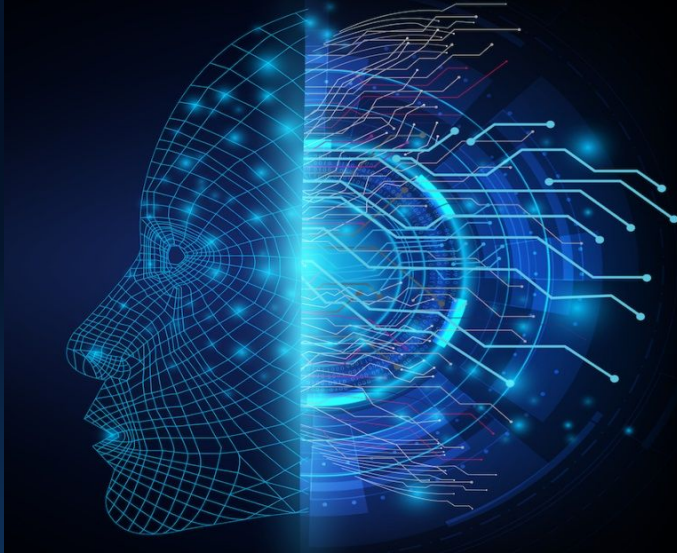
Model oluşturmak için Temel Bileşenlerin kullanılması, yeni girdiler için verilerin tahmininde büyük hatalara neden olabilir.

Bir alternatif, verilerden orijinal faktörleri çıkarmaktır. Ortak faktör analizi, gözlenen değişkenlerin varyansını maksimize edebilen orijinal değişkenlerin kombinasyonunu tanımlayan bir yöntemdir. Bu, orijinal değişkenlerin toplam varyansını maksimize eden PCA'dan farklıdır.

Model oluşturmak için orijinal bileşenlere dayalı modelleme tercih edilmeli ve yalnızca boyutların ölçeğiyle ilgili sorunlar ortaya çıkarsa, Temel Bileşenlere dayalı model oluşturmayı düşünmelisiniz.

PCA ile Modelleme

Model Yorumu



Temel bileşen yükleri, değişkenler için korelasyon faktörünü temsil eder; yüklemenin boyutuna ve işaretine göre değişkenin pozitif ilişkili veya negatif ilişkili olduğu sonucuna varabilirsiniz.

Ancak, ilk temel bileşende düşük yükleme değerine sahip bir değişken, diğer temel bileşenlerde daha yüksek değerlere sahip olabileceğinden, yükleme boyutunu ve yönünü net olarak yorumlayamayız.

Modeli yorumlamaya çalışırken ortak faktör (CF) analizinin kullanılması daha iyi bir yorum sağlayacaktır.



6

SON

Burada onursal üyemiz Hüseyine
bağlanıyoruz...





Teşekkürler!

Sorusu olan var mı?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**



Kaynakça:

https://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/slides/lec12-slides.pdf

https://www.cs.tufts.edu/comp/135/2020f/slides/day22_principal_components_analysis.pdf

<https://www.cs.cmu.edu/~mgormley/courses/10701-f16/slides/lecture14-pca.pdf>

<https://davidrosenberg.github.io/mlcourse/Archive/2017/Lectures/13-PCA-Slides.pdf>

<https://weber.itn.liu.se/~aidvi/courses/06/dm/seminars2008/PCA.pdf>

<http://people.sabanciuniv.edu/berrin/cs512/lectures/11-PCA-FaceReco.pdf>

https://www.math.uci.edu/icamp/courses/math77b/lecture_12w/pdfs/PCA.pdf

<https://bkict-ocw.knu.ac.kr/caster/file/lecture/5DA7C3F03B31B.pdf>

https://alex.smola.org/teaching/cmu2013-10-701/slides/14_PrincipalComp.pdf

https://en.wikipedia.org/wiki/Principal_component_analysis

<http://www.zafercomert.com/IcerikDetay.aspx?zcms=78>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2757795/>