Principal Component Analysis (PCA)

Atıl Samancıoğlu

1 Giriş

Principal Component Analysis (PCA), çok boyutlu veri kümelerinde bulunan korelasyonları kullanarak daha az sayıda boyutta, veri içindeki varyansı koruyacak şekilde yeni bileşenler üretmeyi amaçlayan bir **boyut indirgeme** yöntemidir.

PCA ne işe yarar?

- Curse of Dimensionality'yi (Boyutlanma Laneti) azaltır,
- Modelin eğitim süresini ve hesaplama maliyetini düşürür,
- Aşırı öğrenmeyi (overfitting) engeller,
- Veriyi 2D veya 3D'ye indirerek görselleştirme imkânı sunar.

2 Curse of Dimensionality Nedir?

Boyut sayısı (özellik sayısı) arttıkça:

- Modelin ezberleme riski artar (overfitting),
- Anlamsız ya da gürültülü değişkenler modelin doğruluğunu düşürebilir,
- Hesaplama süresi ve karmaşıklık artar,
- Görselleştirme imkânı ortadan kalkar.

Örnek: Bir konut fiyat tahmini problemi düşünelim. Eğer modelinize 10 anlamlı değişkenin yanında 90 gereksiz değişken eklerseniz, model doğru sinyalleri gürültüden ayırt etmekte zorlanacaktır. Bu da hem doğruluğu hem de eğitim süresini olumsuz etkiler.

3 Boyut İndirgeme Yöntemleri

1. **Feature Selection (Özellik Seçimi):** Sadece önemli değişkenleri tutar.
2. **Feature Extraction (Özellik Çıkarımı):** Var olan değişkenlerden yeni bileşenler üretir. PCA bu kategoridedir.

4 Feature Extraction: PCA'nin Temel Fikri

PCA, birbirine bağlı (korelasyonlu) özellikler arasında yeni eksenler (**Principal Components**) tanımlar. Bu eksenler:

- Veri içindeki varyansı maksimum yakalayan doğrultulardır,
- Birbirine diktir (orthogonal),
- Enformasyon kaybını minimumda tutarak boyut indirgeme sağlar.

Hedef

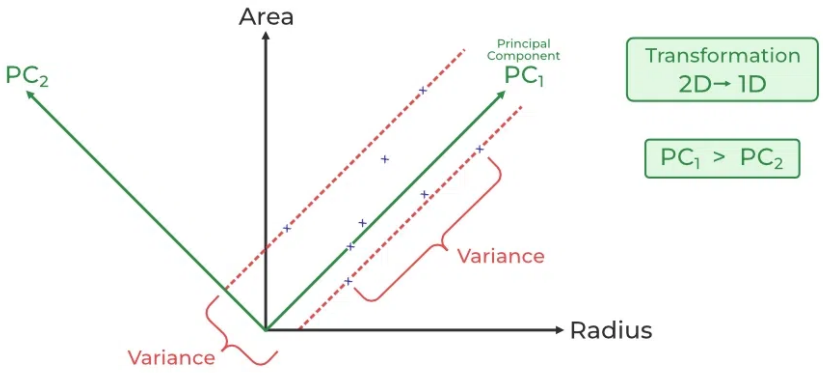
Orijinal n boyutlu veriyi, en fazla varyans bilgisini taşıyan $k < n$ boyutlu bir alt uzaya projekte etmek.

5 Geometrik Sezgi

İki boyutlu bir örnek:

- X: Oda büyüklüğü
- Y: Oda sayısı

Bu iki özellik birbiriyle güçlü bir şekilde ilişkilidir. PCA, bu ortak varyansı maksimum tutacak şekilde yeni bir eksen (örneğin "evin genel büyüklüğü") tanımlar.



Buradaki temel fikir:

- PCA1: En fazla varyansı yakalayan yön,
- PCA2: PCA1'e dik, kalan varyansı yakalayan yön.

Eğer 2D veriyi 1D'ye indirmek istiyorsak sadece PCA1 tutulur.

6 Eigenvector ve Eigenvalue Nedir?

PCA'nın temelinde yatan kavramlardan ikisi **eigenvector (özvektör)** ve **eigenvalue (özdeğer)**'dir. Bu terimler ilk bakışta soyut görünse de, PCA'da amaçları oldukça nettir: verideki **en fazla bilgi taşıyan yönleri** bulmak.

Vektör ve Doğrultu Sezgisi

Bir veri kümesi içindeki gözlemleri birer vektör gibi düşünebiliriz. Bu vektörlerin uzayda nasıl yayıldığını (hangi doğrultuda en çok dağıldıklarını) bilmek, verinin yapısını anlamak açısından önemlidir.

İşte tam bu noktada **eigenvector**, verinin yayılma (spread) yönünü; **eigenvalue** ise bu yöndeki yayılmanın (varyansın) miktarını gösterir.

Matematiksel Tanım

Bir matris A için, aşağıdaki denklem sağlanıyorsa:

$$A\vec{v} = \lambda\vec{v}$$

burada:

- A: Kovaryans matrisi (verinin yapısını özetler),
- $\vec{v}$: Eigenvector (doğrultu),
- λ: Eigenvalue (o doğrultudaki veri yayılımı).

Bu denklem diyor ki: "Eğer vektör $\vec{v}$, matris A ile çarpıldığında sadece ölçekleniyorsa (yönü değişmeden sadece boyu değişiyorsa), o zaman $\vec{v}$ bir **eigenvector**, ölçek katsayısı λ ise **eigenvalue**'dur."

Görsel Sezgi

`https://wiki.pathmind.com/eigenvector`

PCA Açısından Anlamı

PCA algoritması:

- Kovaryans matrisini oluşturur,
- Bu matrisin eigenvector'lerini ve bunlara karşılık gelen eigenvalue'leri bulur,
- En büyük eigenvalue'ya sahip olan eigenvector, verinin en fazla varyans gösterdiği doğrultudur (PC1),
- Diğer bileşenler (PC2, PC3...) daha az varyansı temsil eder.

Neden Eigenvector ve Eigenvalue Kullanıyoruz?

- Verideki **en anlamlı yönleri** otomatik olarak bulur,
- Bilginin en yoğun olduğu doğrultular seçilir,
- Boyut indirimi sırasında **bilgi kaybı minimumda tutulur**,
- Sadece korelasyonlara bakmak yerine, tüm verinin geometrik yapısını analiz eder.

Bu sayede PCA, yüksek boyutlu veri setlerinden özet ama anlamlı temsiller çıkarmada son derece etkili bir yöntem haline gelir.

7 Matematiksel Temel: PCA Adımları

1. Verinin Standartlaştırılması

Her özelliğin ortalaması 0, standart sapması 1 yapılır. Aksi halde yüksek skala değeri olan özellikler varyansı domine eder.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

2. Kovaryans Matrisi Hesaplama

Kovaryans, iki değişkenin birlikte nasıl değiştiğini gösterir:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Örnek bir kovaryans matrisi:

$$\begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{bmatrix}$$

3. Eigen Decomposition (Öz Vektör ve Değerleri)

$$A\mathbf{v} = \lambda\mathbf{v}$$

Burada:

- A: Kovaryans matrisi
- $\mathbf{v}$: Eigenvector (yeni eksen yönü)
- λ: Eigenvalue (bu eksenin varyansı ne kadar yakaladığı)

4. Principal Components Seçimi

- Tüm eigenvalue'lar büyükten küçüğe sıralanır,
- En yüksek k tanesi seçilir,
- Bunlara karşılık gelen eigenvector'lar alınır.

5. Veri Projeksiyonu

Veri bu yeni eksenlere projekte edilir:

$$Z = X \cdot W$$

Burada:

- X: Orijinal verimiz
- W: Seçilen eigenvector'lar matrisi (boyutu $d \times k$)
- Z: Yeni boyut indirgenmiş veri (boyutu $n \times k$)

8 PCA'nin Avantajları

- Bilgi kaybı olmadan boyut indirgeme (en fazla varyans korunur),
- Daha hızlı model eğitimi,
- Gürültüden arındırılmış veri,
- Görselleştirilebilir hale gelmiş veri (örneğin 100 boyuttan 2 boyuta),
- Korelasyonlu değişkenlerin ortadan kaldırılması.

9 Örnek: Neden PCA?

Durum: 2 özellikli bir veri setimiz var:

- X1: Oda büyüklüğü
- X2: Oda sayısı

Her ikisi de yüksek korelasyonlu.
Çözüm: PCA ile bu iki özelliği birleştirip tek bir bileşen (örneğin "ev büyüklüğü") oluştururuz. Böylece:

- Veri boyutu azalır,
- Fazla bilgi kaybı olmadan model eğitimi yapılabilir.

10 Sonuç

PCA:

- Korelasyonlu değişkenleri birleştirerek bilgi özetler,
- Veri boyutunu düşürür,
- Model performansını ve eğitim süresini iyileştirir,
- Varyansı korumaya odaklı olduğu için anlamlı dönüşümler yapar.

İdeal olarak veri setinizdeki toplam varyansın büyük bir kısmını (örneğin