YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Mekatronik Mühendisliği Uygulamalarında Yapay Zekâ

Ders 9 – Özellik Belirleme (Feature selection and extraction)

Prof. Dr. Erhan AKDOĞAN

Giriş:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- Makine öğreniminde özellik seçimi ve özellik çıkarımı (feature selection ve feature extraction) veri ön işleme süreçlerinde kullanılan iki farklı yöntemdir. İkisi de modelin performansını artırmak ve veri boyutunu azaltmak için kullanılır, ancak farklı amaçları ve yaklaşımları vardır.

Özellik Seçimi (Feature Selection):

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Özellik seçimi, mevcut özellikler (değişkenler) arasından en anlamlı olanlarını seçmek anlamına gelir. Burada yeni özellikler oluşturulmaz; sadece mevcut özelliklerin bir alt kümesi alınır.

Avantajları:

•**Yorumlanabilirlik:**Verinin anlamını ve özelliklerin etkisini korur, bu da model sonuçlarının yorumlanabilirliğini artırır.

•**Basitlik:** Modeli daha hızlı çalıştırır ve daha az karmaşıklık sağlar.

•**Aşırı Öğrenmeyi Azaltır:** Daha az gereksiz veriyle çalıştığı için aşırı öğrenme riskini azaltır.

Dezavantajları:

•**Ayırt Edici Güç Azalabilir:** Bazı durumlarda önemli olabilecek özellikler göz ardı edilebilir.

•**Doğruluk Kaybı:** Özelliklerin azaltılması, modelin performansını düşürebilir.

Kullanım Örnekleri:

•İstatistiksel yöntemler (örneğin, **ANOVA**, **chi-square testi**).

•Makine öğrenimi tabanlı yöntemler (örneğin, **recursive feature elimination**).

Özellik Çıkarımı (Feature Extraction):

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Özellik çıkarımı, mevcut özelliklerden daha anlamlı ve özet özellikler oluşturma sürecidir. Bu yeni özellikler, genellikle orijinal özelliklerin matematiksel bir dönüşümüdür.

Avantajları:

- **Yüksek Ayırt Edicilik:** Yeni oluşturulan özellikler, genellikle daha güçlü ve anlamlıdır.
- **Boyut Azaltma:** Verinin boyutunu önemli ölçüde azaltabilir.
- **Aşırı Öğrenmeyi Kontrol Eder:** Özellikle karmaşık modellerde aşırı öğrenmeyi önleyebilir.

Dezavantajları:

- **Yorumlanabilirlik Kaybı:** Özelliklerin dönüşümü, verinin anlamını karmaşıklaştırabilir.
- **Maliyet:** Özellik çıkarımı için gerekli olan hesaplamalar pahalı olabilir (örneğin, büyük veri setlerinde).

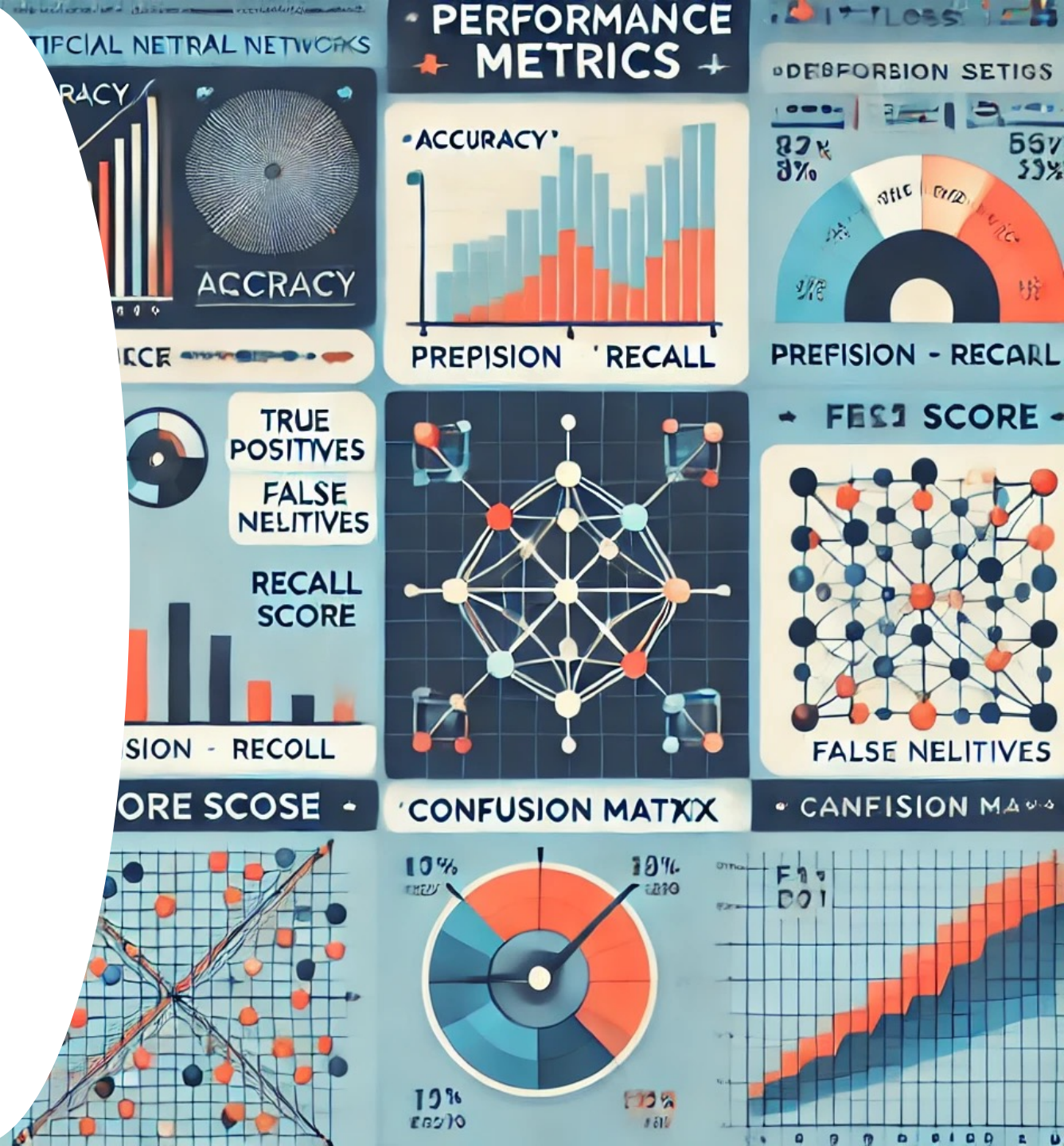
Kullanım Örnekleri:

- **Principal Component Analysis (PCA):** Çok boyutlu veriyi daha az boyuta indirir.
- **T-SNE, Autoencoders:** Karmaşık dönüşümlerle anlamlı özellikler oluşturur.

Özellik Seçimi

"Doğru özellikleri seç, veriyi sadeleştir, modeli güçlendir!

Makine öğreniminde başarı, karmaşıklığı değil, anlamı yakalamaktan geçer!"



Özellik seçimini kullanmanın başlıca nedenleri:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- Makine öğrenimi algoritmasının daha hızlı ilerlemesini sağlar.
- Bir modelin karmaşıklığını azaltır ve yorumlanmasını kolaylaştırır.
- Doğru alt küme seçildiğinde bir modelin doğruluğunu artırır.
- Ezberlemeyi, minimuma takılmayı azaltır, genelleştirmeyi artırır.

Özellik seçim yöntemleri:

1. Bilgi kazancı (Information Gain)
2. Sinyal Gürültü Oranı (signal to noise ratio)
3. Alt Küme Seçiciler (Wrappers)
4. Filtre Metodları (Filter methods)
5. Gömülü Metodlar (Embedded methods)

1. Bilgi Kazancı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- **Bilgi kazancı**: Verilen bir özelliğin sınıflandırmada ne kadar etkili olduğunu gösteren ölçüttür. 0 ile 1 arasında değişir.
- **Entropi:** Bir veri kümesi içinde belirsizlik ve rasgeleliği ölçmek için kullanılır. Bu değer ne kadar büyük ise belirsizlik o kadar yüksektir.

Entropi

$$H(S) = -\sum_{t=1}^{n} p_i * log_2(p_i)$$

H **: Entropi**
S **: Kaynak**
p **: Olasılık**

Bilgi Kazancı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

$$IG(D) = H(D) - \sum_{i=1}^{n} P(D_i)H(D_i)$$

***IG*: Bilgi kazancı**
***D*: veri kümesi**
***P*: ağırlık olasılığı**
***H*: Entropi**

Örnek:

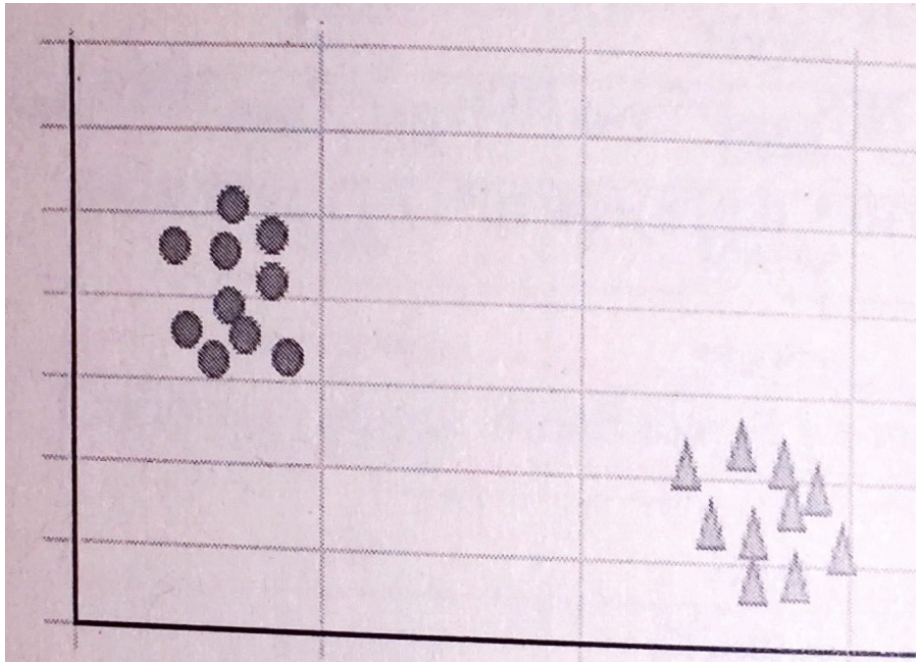
YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

YAŞ	MEZUNİYET	ŞİRKET SAHİBİ	KARAR
orta	lise	E	İYİ
orta	üniv	E	FAKİR
yaşlı	lise	E	ZENGİN
genç	lise	H	İYİ
genç	üniv	E	ORTA
genç	lise	H	İYİ
yaşlı	üniv	H	İYİ
yaşlı	üniv	E	ZENGİN
yaşlı	lise	E	İYİ
orta	üniv	E	FAKİR

Hangi özelliğin bilgi kazancı daha yüksektir?

2. Sinyal Gürültü Oranı:

- Sınıflar arası uzaklıklar fazla, sınıf içi uzaklıklar az olduğunda özellik seçiminde kullanılan bir yöntemdir.
- Herbir özellik için bu oran ayrı ayrı hesaplanır.
- Yüksek değerler yüksek ilişkiye(korelasyon) işaret eder.



İki sınıflı örneklem kümesinin koordinat düzlemindeki görüntüsü

Sinyal gürültü oran formülü

$$S_i = \frac{m_1 - m_2}{d_1 - d_2}$$

S_i : i. özelliğin sinyal gürültü oranı
m_1 : 1.sınıfdaki özelliklerin ortalaması
d_1 : 1.sınıfdaki özelliklerin standart sapması

3. Alt Küme Seçiciler (Wrappers)

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- Herbir özellik için ayrı bir bilgi edinme yerine özellikler birlikte değerlendirilerek sınıflandırma yapılır ve özellik alt uzayları tespit edilir.
- Sınıflandırma başarısı yüksektir.
- Hesaplama karmaşıklığı içerir. Çalışma hızları yavaştır.
- Özellik seçiminin her adımında sınıflandırıcıya ihtiyaç duyar.

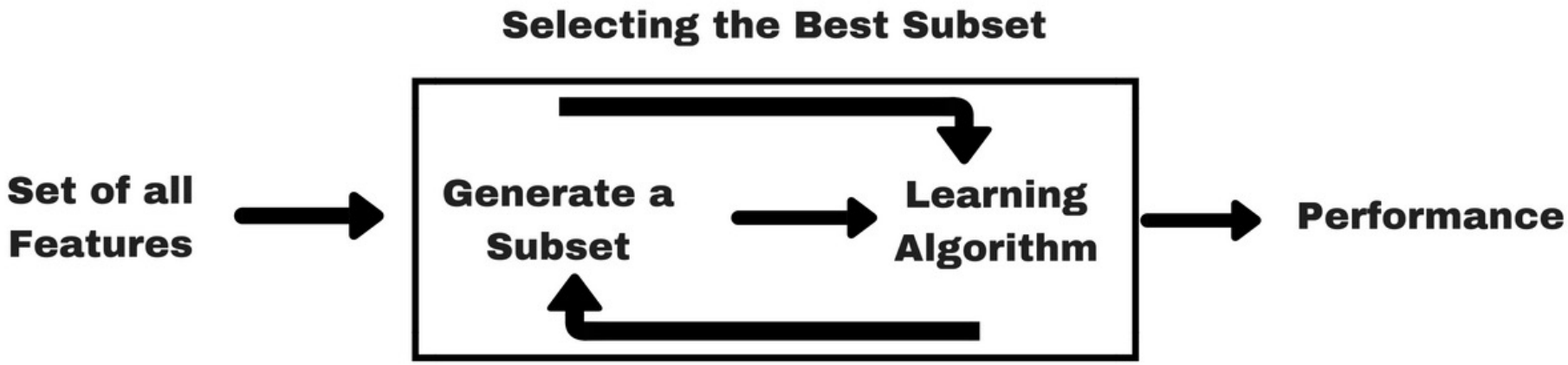
Alt Küme Seçiciler (devam)

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- Bu yöntemde, bir alt özellik kümesi kullanılmaya ve bunları kullanarak bir model geliştirilmeye(eğitilmeye) çalışılır.
- Önceki modelden alınan çıkarımlara dayanarak alt kümeden özellikler eklemeye veya kaldırmaya karar verilir.
- Problem aslında bir arama problemine indirgenmiştir.
- Bu yöntemler genellikle hesaplama açısından çok pahalıdır.
- Bu yöntem ile en iyi özellik çıkarımı için “Boruta” metodu kullanılabilir. Detaylı bilgi için

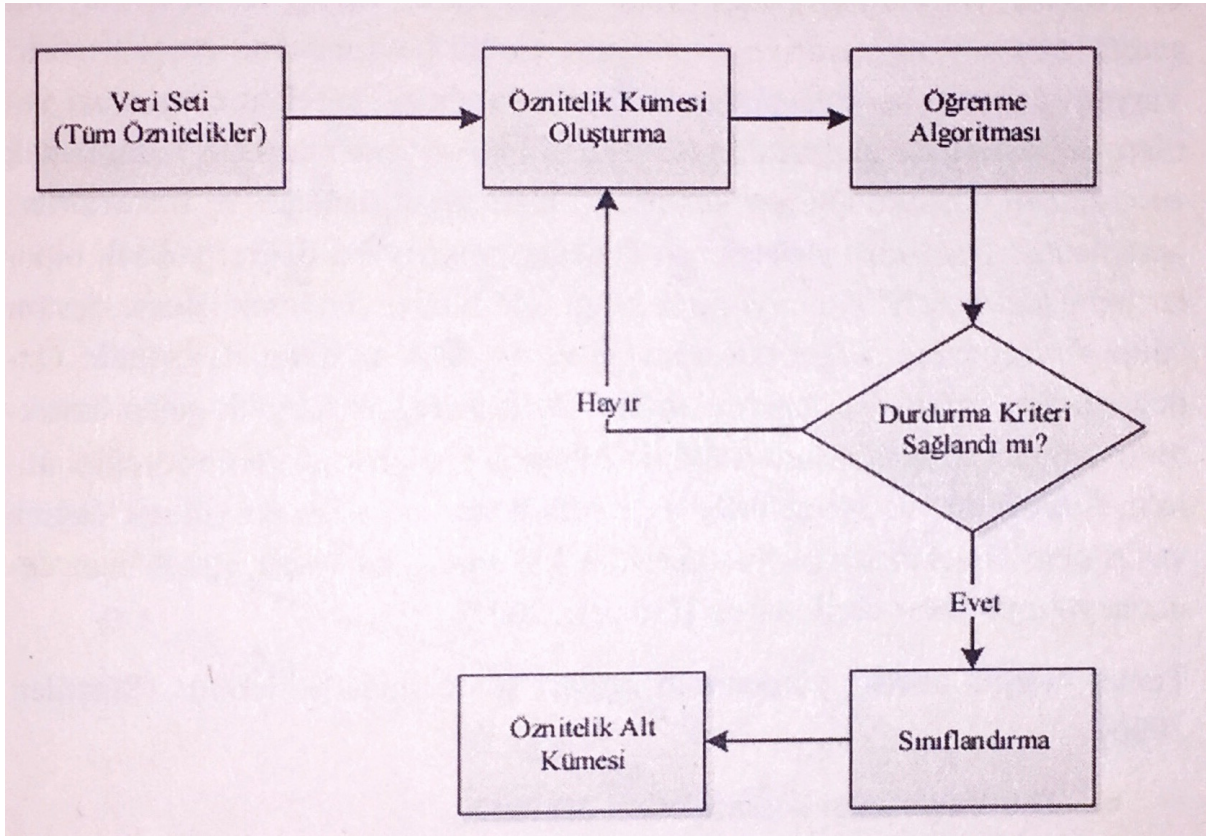
https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/

Alt küme seçim diyagramı:



https://www.analyticsvidhya.com

Alt küme seçim diyagramı:



Alt küme seçicilere ait akış diyagramı(Kaya, 2014)

Alt küme seçicilerde yaygın örnekler

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- İleriye doğru özellik seçimi,
- Geriye doğru özellik eleme,
- Özyinelemeli (recursive) özellik eleme, vb.

İleriye doğru özellik seçimi

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- İleriye doğru seçim, modelde hiçbir özelliğe sahip olmadan başladığımız yinelemeli(iteratif) bir yöntemdir.
- Her yinelemede, yeni bir değişkenin eklenmesi, modelin performansını iyileştirmeyene kadar modeli en iyi şekilde geliştiren özellik eklenmeye devam edilir.

Geriye Doğru Eleme:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- Tüm özellikler ile başlanır
- Modelin performansını artıran her yinelemede en az önemli olan özellik kaldırılır.
- Özelliklerin kaldırılmasında hiçbir gelişme gözlenmeyene kadar bu işlem devam eder.

Özyinelemeli (recursive) Özellik Elemesi:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- En iyi performans gösteren özellik alt kümesi bulunmaya çalışılır.
- Bir optimizasyon algoritmasıdır.
- Tekrarlı olarak model oluşturur ve her yinelemede en iyi veya en kötü performans özelliği elenir.
- Tüm özellikler bitene kadar bir sonraki modeli atılan özellikler ile yapılandırır.
- Daha sonra, elemelerinin sırasına göre özellikleri sıralar.

4. Filtre Metodları:

- Genellikle önişleme adımlarında kullanılır.
- Özellik seçimi herhangi makine öğrenmesi adımından bağımsızdır.
- Özellikler, çıktı değişken ile özellik korelasyonları için çeşitli istatistiksel testlerdeki puanlara dayanarak seçilir.

İstatiksel Metodlar:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Pearson Korelasyonu

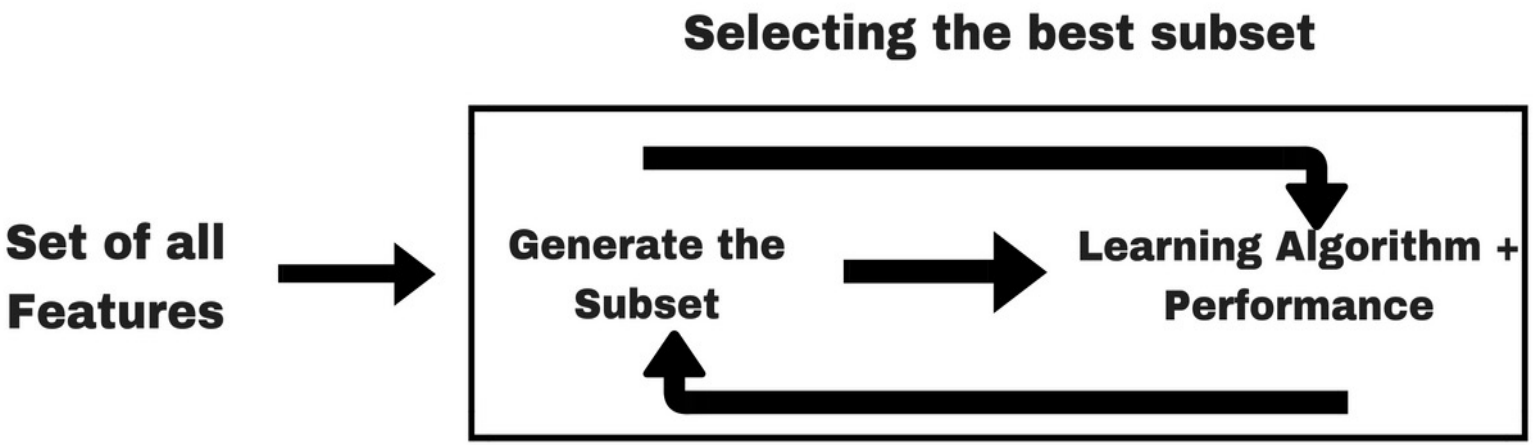
LDA: Linear discriminant analysis

ANOVA: Varyansın analizi için kullanılır.

Chi-Square

5. Gömülü metotlar

- Gömülü yöntemler, filtre ve wrapper yöntemlerinin niteliklerini birleştirir.
- Kendi yerleşik özellik seçim yöntemlerine sahip olan algoritmalar tarafından uygulanır.



Özellik Çıkarımı

Özellik çıkarım Yöntemleri

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

1. İstatistiksel Özellik Çıkarımı
2. Frekans Tabanlı Özellik Çıkarımı
3. Temel İşlem Tabanlı Özellik Çıkarımı
4. **Boyut Azaltma Yöntemleri (PCA, LDA, t-SNE, UMAP)**
5. Gözetimli Özellik Çıkarımı
6. Sinyal Tabanlı Özellik Çıkarımı
7. Metin ve Doğal Dil İşleme Özellikleri (TF-IDF, N-gram, Embedding)
8. Görüntü Özellikleri (HOG, SIFT, SURF)
9. Derin Öğrenme Tabanlı Özellik Çıkarımı
10. Uzaysal ve Bağlamsal Özellik Çıkarımı
11. Veri Zenginleştirme ve Özellik Kombinasyonu
12. Transfer Öğrenme ile Özellik Çıkarımı
13. Mutual Information
14. Genetik Algoritmalar
15. Otomatik Kodlayıcılar (Autoencoders)

1. İstatistiksel Özellik Çıkarımı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Ham verilerin temel istatistiksel özetlerini çıkarır:

- **Ortalama (Mean)**: Veri dağılımının merkezi eğilim ölçüsü.
- **Standart Sapma (Standard Deviation)**: Verinin yayılım ölçüsü.
- **Medyan, Çeyrekler (Median, Quartiles)**: Merkezi değer ve dağılım bilgisi.
- **Çarpıklık ve Basıklık (Skewness, Kurtosis)**: Veri dağılımının simetrisi ve sivriliği.
- **Minimum ve Maksimum Değerler**: Veri aralığını temsil eder.

Kullanım Alanı: Zaman serisi analizi, sinyal işleme.

2. Frekans Tabanlı Özellik Çıkarımı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Zaman alanındaki veriyi frekans alanına dönüştürerek analiz eder:

- **Fourier Dönüşümü (FFT)**: Ana frekans bileşenlerini analiz eder.
- **Wavelet Dönüşümü**: Zaman ve frekans bilgisini birlikte sağlar.
- **Spektral Enerji (Spectral Energy)**: Frekans bantlarındaki enerji yoğunluğu.
- **Spektral Pikler (Spectral Peaks)**: Önemli frekans bileşenlerini tespit eder.
- **Kullanım Alanı:** Ses işleme, görüntü analizi, biyomedikal sinyal işleme.

3. Temel İşlem Tabanlı Özellik Çıkarımı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Ham veriden çeşitli matematiksel işlemlerle özellikler oluşturur:

- **Polinom Özellikleri (Polynomial Features):** Mevcut özelliklerin polinomlarını çıkarır.
- **Logaritmik ve Üstel Dönüşümler:** Veriyi normalize eder veya doğrusal olmayan ilişkileri ortaya çıkarır.
- **Oranlar ve Farklar:** Özellikler arasındaki ilişkileri temsil eder.

Kullanım Alanı: Finansal analiz, regresyon problemleri.

4. Boyut Azaltma Yöntemleri

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Boyut azaltma ile daha anlamlı alt özellik kümeleri çıkarır:

- **Temel Bileşenler Analizi (PCA):** Varyansı maksimize eden bileşenleri çıkarır.
- **Doğrusal Ayırt Etme Analizi (LDA):** Sınıflar arasındaki ayrımı maksimize eden bileşenleri seçer.
- **T-SNE ve UMAP:** Görselleştirme ve boyut azaltmada kullanılır.
- **Kullanım Alanı:** Görselleştirme, sınıflandırma, yüksek boyutlu veri analizi.

5. Gözetimli Özellik Çıkarımı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Etiketli verilerden anlamlı özellikler çıkarır:

- Mutual Information: Özellikler ile sınıf etiketleri arasındaki ilişkiyi ölçer.
- ReliefF Algoritması: Sınıf ayrımına en fazla katkıda bulunan özellikleri seçer.
- Sınıf Ayrımı Bazlı Analiz: Sınıflar arasındaki mesafeyi maksimize eder.
- Kullanım Alanı: Sınıflandırma, gözetimli öğrenme.

6. Sinyal Tabanlı Özellik Çıkarımı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Sinyal işleme teknikleri kullanılarak özellikler elde edilir:

- Enerji Özellikleri: Sinyalin toplam enerjisi veya güç spektrumu.
- Tepe Noktalar (Peaks): Sinyaldeki önemli olaylar.
- Zarf Analizi (Envelope Analysis): Sinyal genliğindeki değişiklikler.
- Kullanım Alanı: Titreşim analizi, sağlık verisi analizi.

7. Metin ve Doğal Dil İşleme Özellikleri

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Metin verisinden anlamlı özellikler çıkarır:

- N-gramlar: Kelime veya harf gruplarının sıklığı.
- TF-IDF (Term Frequency-Inverse Document Frequency): Belirli kelimelerin önemini ölçer.
- Embedding Teknikleri: Word2Vec, GloVe, BERT gibi modellerle metni sayısal vektörlere dönüştürür.
- Sentiment Analizi Özellikleri: Duygu skorları çıkarılır.
- Kullanım Alanı: Metin sınıflandırma, duygu analizi.

8. Görüntü Özellikleri

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Görsellerden anlamlı özellikler çıkarır:

- **Kenar Tespiti (Edge Detection):** Görsellerdeki önemli kenarları belirler.
- **Histogram Özellikleri:** Renk dağılımını analiz eder.
- **Özellik Algılayıcılar: HOG** (Histogram of Oriented Gradients), SIFT, SURF gibi algoritmalarla görüntüden anlamlı özetler çıkarılır.
- **Kullanım Alanı:** Görüntü sınıflandırma, yüz tanıma.

9. Derin Öğrenme Tabanlı Özellik Çıkarımı

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Derin öğrenme modelleri, ham veriden otomatik olarak özellikler çıkarır:

- **Konvolüsyonel Sinir Ağları (CNN):** Görsellerden uzamsal özellikler öğrenir.
- **Yinelemeli Sinir Ağları (RNN):** Zaman serisi verisinden veya metinden ilişkiler çıkarır.
- **Transfer Öğrenme:** Önceden eğitilmiş bir modelden özellikler alınır.
- **Otomatik Kodlayıcılar (Autoencoders):** Düşük boyutlu anlamlı özellikler çıkarır.
- **Kullanım Alanı:** Görüntü, metin ve zaman serisi analizi.

10. Uzaysal ve Bağlamsal Özellikler

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Verinin uzaysal ve bağlamsal ilişkilerini analiz eder:

- Coğrafi Veri Analizi: Konum tabanlı ilişkiler.
- Hough Dönüşümü: Şekil tespiti.
- Komşuluk İlişkileri: Özellikler arasındaki mesafeler.
- Kullanım Alanı: Görüntü işleme, coğrafi veri analizi.

11. Veri Zenginleştirme ve Özellik Kombinasyonu

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Ham veriyi zenginleştirerek veya mevcut özelliklerden yeni özellikler türeterek daha anlamlı bilgiler elde edilir:

- **Zenginleştirme:** Dış veri kaynaklarından alınan bilgiler.
- **Özellik Kombinasyonu:** İki veya daha fazla özelliğin matematiksel kombinasyonu.
- **Kullanım Alanı:** Veri analitiği, makine öğrenimi.

Özellik Çıkarım Yöntemleri:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

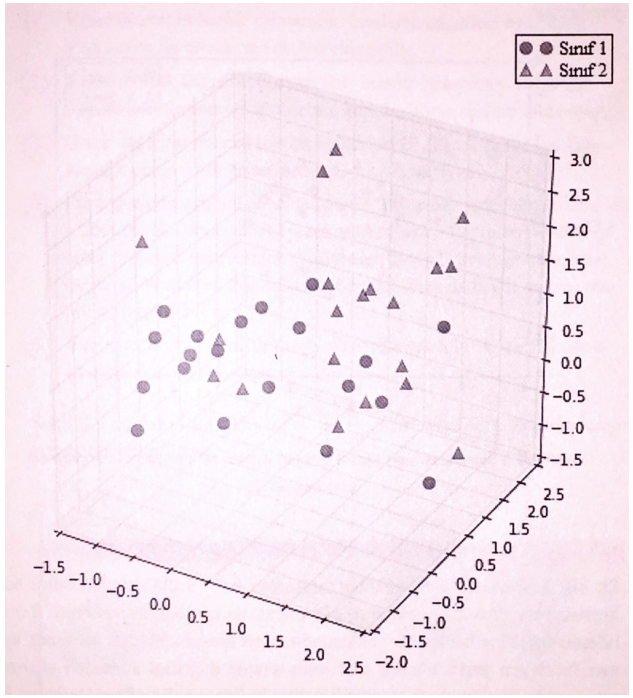
1. Temel Bileşen analizi (Principal Component Analysis (PCA))
2. Doğrusal Ayırt eden Analizi

1. Temel Bileşen Analizi:

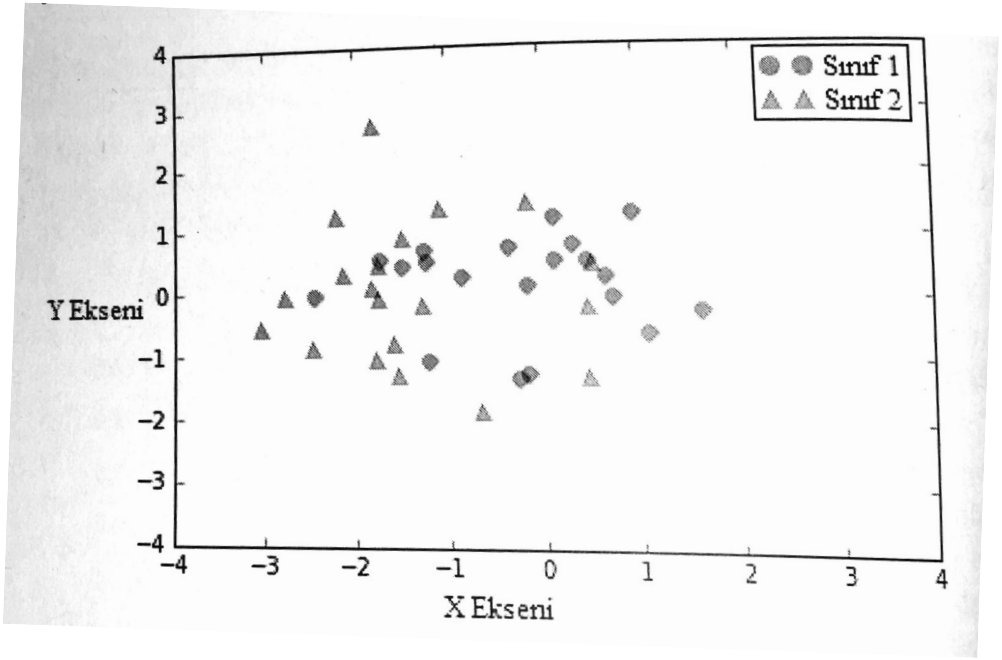
YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- Temel bileşen analizi birbiri ile ilişkili olan birden fazla değişkeni bulunan bir veri kümesinin boyutunu azaltmak için geliştirilmiş bir yöntemdir.
- Amaç: verinin temel yapısının bulunması ve boyutunun azaltılmasıdır.
- Boyut azaltma varyans ile yapılır.
- Bunun için eldeki verilerle kovaryans matrisi hesaplanır.
- Bundan sonra buna göre özdeğer (eigen value) ve özvektörler (eigen vectors) hesaplanır.
- Sayısal değeri yüksek özdeğerler ile işleme devam edilir.
- Özdeğerin sıfır olması ilgili özdeğere karşılık gelen özvektörün kendisi dışındaki özvektörlerin bileşenleri oalrak gösterilecebileceğini ifade eder.
- En yüksek değere sahip özvektörlerden hangisinin kullanılacağı deneme yanılma yolu ile belirlenir.

Örnek:



3-B Veri



İndirgenmiş 2-B Veri

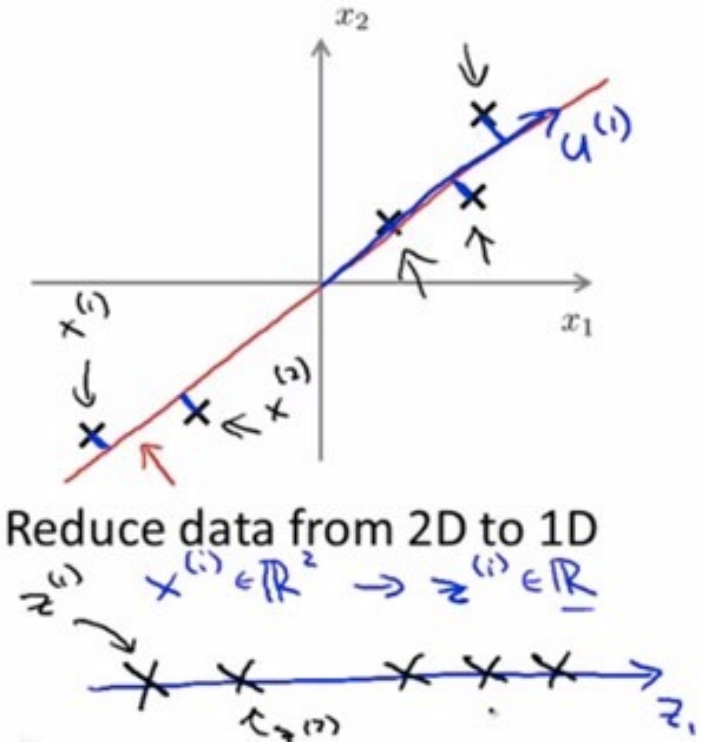
Bir başka deyişle temel bileşen analizi

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

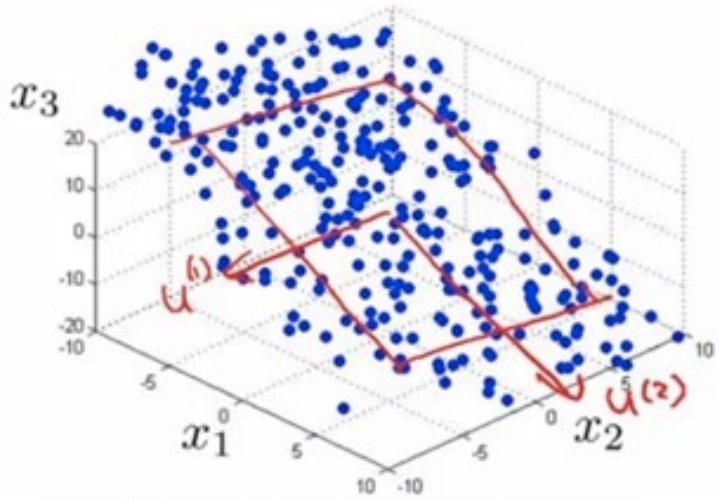
- Bir PCA analizinin ana amacı: verideki paternleri tanımlamaktır.
- PCA değişkenler arasındaki korelasyonu tespit etmeyi amaçlamaktadır.
- Değişkenler arasında güçlü bir korelasyon varsa, boyutsallığı azaltma mantıklıdır.
- Özetle, PCA'nın tamamı şu şekildedir: Yüksek boyutlu verilerde maksimum varyansın yönlerini bulmak ve bilgiyi korurken daha küçük boyutlu bir alt uzaya bunları yansıtmak.

Veri indirgeme:

Principal Component Analysis (PCA) algorithm



Reduce data from 2D to 1D



Reduce data from 3D to 2D

Kaynak: Machine Learning Lectures by Prof. Andrew NG at Stanford University

Temel Bileşen Analizi Adımları:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

1. Verileri normalize et
2. Kovaryans hesabı yap
3. Özdeğer(Eigenvalue) ve Özvektörleri(Eigen Vector) Hesapla
4. Bileşenleri seç ve bir özellik vektörü haline dönüştür
5. Temel Bileşen haline getir

2. Doğrusal Ayırteden Analizi:

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

- Veri kümesinin herbir elemanın ortalama değerleri hesaplanır. Veri sayısı ile ortalama vektörü elde edilir.

- Veririnin ortak kovaryans matrisi hesaplanır. (Matrisin boyutu her bir veri vektörü için d kabul edilir ise d * d şeklinde olur.)

- Ortak kovaryans (değişinti) matrisi için öz vektörleri (E1, E2,… Ed) ve bunlara karşılık gelen öz değerleri $(\lambda_1, \lambda_2, \ldots\ldots, \lambda_d)$ hesaplanır.

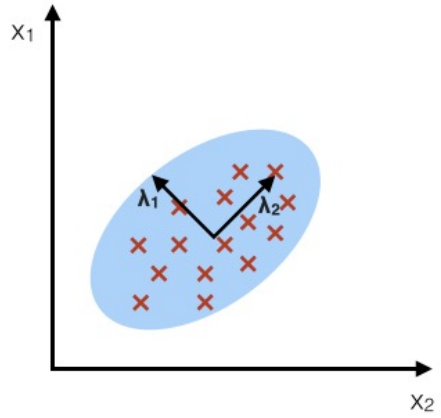
2. Doğrusal Ayırteden Analizi(Devam)

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

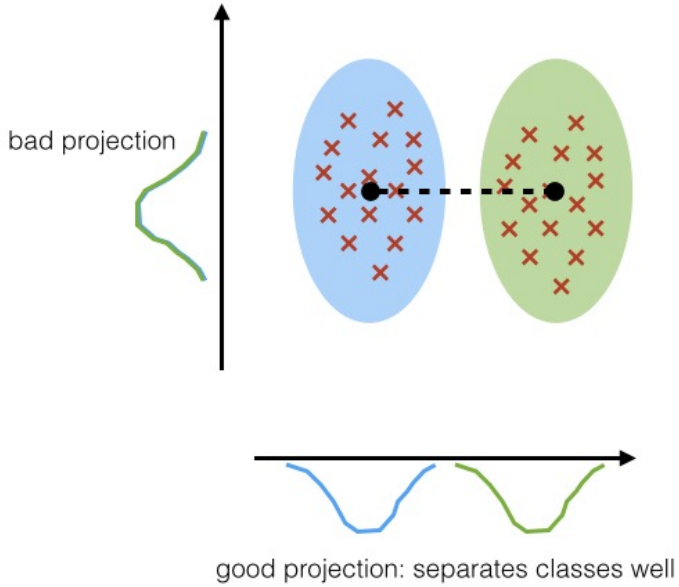
- Özdeğerlerin azalma sırasına karşılık gelen öz vektörler sıralanır.
- k adet büyük öz vektörü seçecek şekilde d*k boyutlu bir W matrisi ortak kovaryans matrisinden oluşturulur.
- Veri vektörlerini yeni alt uzaya dönüştürmek için W matrisi dönüşüm matrisi olarak kullanılır.

Karşılaştırma:

PCA:
component axes that maximize the variance



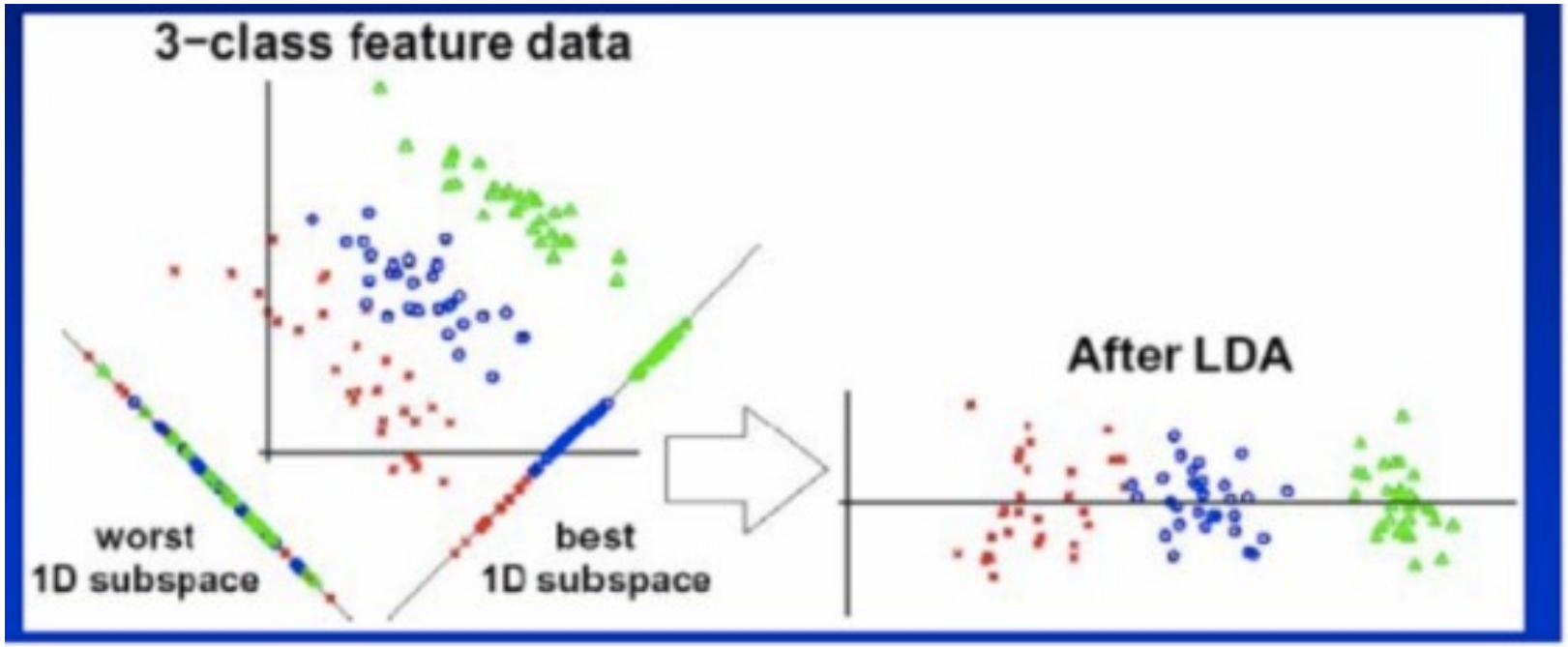
LDA:
maximizing the component axes for class-separation



Temel Bileşen Analizi Varyansı maksimize eden bileşen ekseni

Doğrusal Ayırteden Analiz
Sınıf ayrımı için bileşen eksenlerinin maksimize edilmesi

Üç sınıfa ait verilerin 1 boyuta indirgenmesi



YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

Referanslar

- M. Bilgin, *Makine Öğrenmesi: Teorisi ve Algoritmaları*, A. Yılmaz, Ed., 1st ed. Istanbul, Turkey: [Yayınevi Adı], 2024. ISBN: 978-605-9594-25-7.
- Machine Learning Lectures by Prof. Andrew NG at Stanford University
- https://www.analyticsvidhya.com

YILDIZ TECHNICAL UNIVERSITY
BIOMECHATRONICS
LABORATORY

TEŞEKKÜRLER

Prof. Dr. Erhan AKDOĞAN

eakdogan@yildiz.edu.tr

www.biomech.yildiz.edu.tr