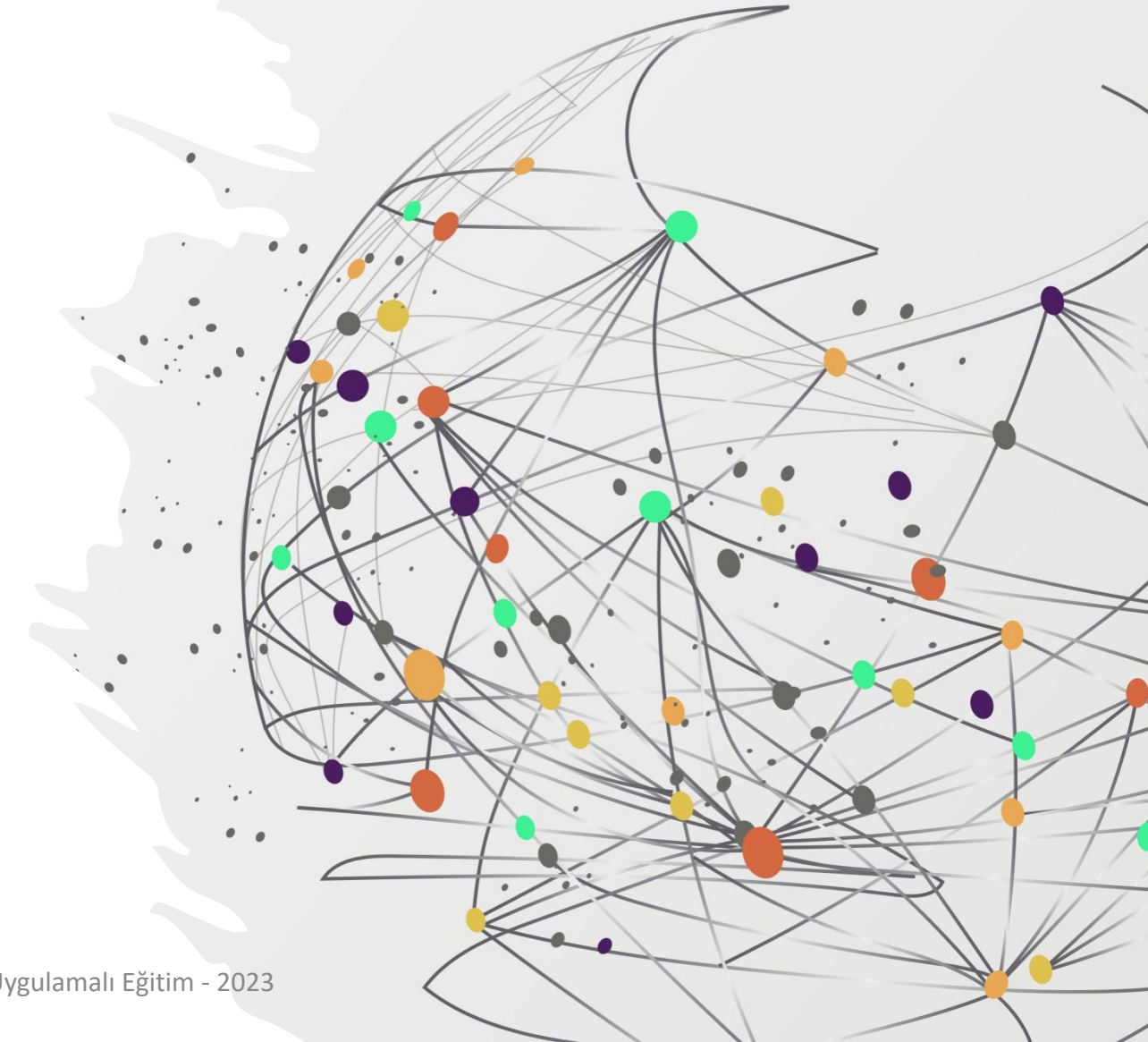
Makine Öğrenmesi ve Genombilim

M. Çisel Kemahlı Aytekin

mckemahli@gmail.com

@ciselkemahli



Ders İçeriği

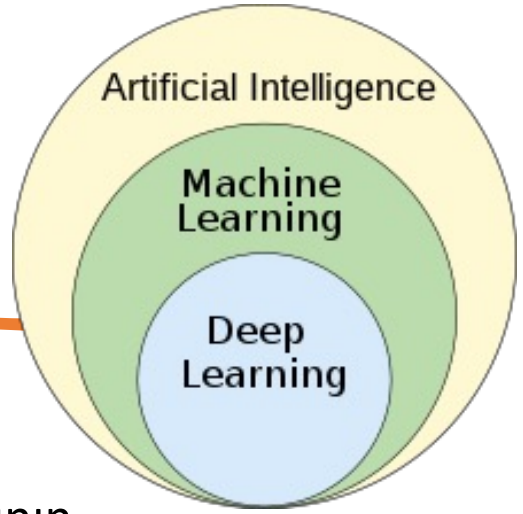
Temel Makine Öğrenmesi (ML)

Genombilimde ML

ML algoritmaları

Uygulama

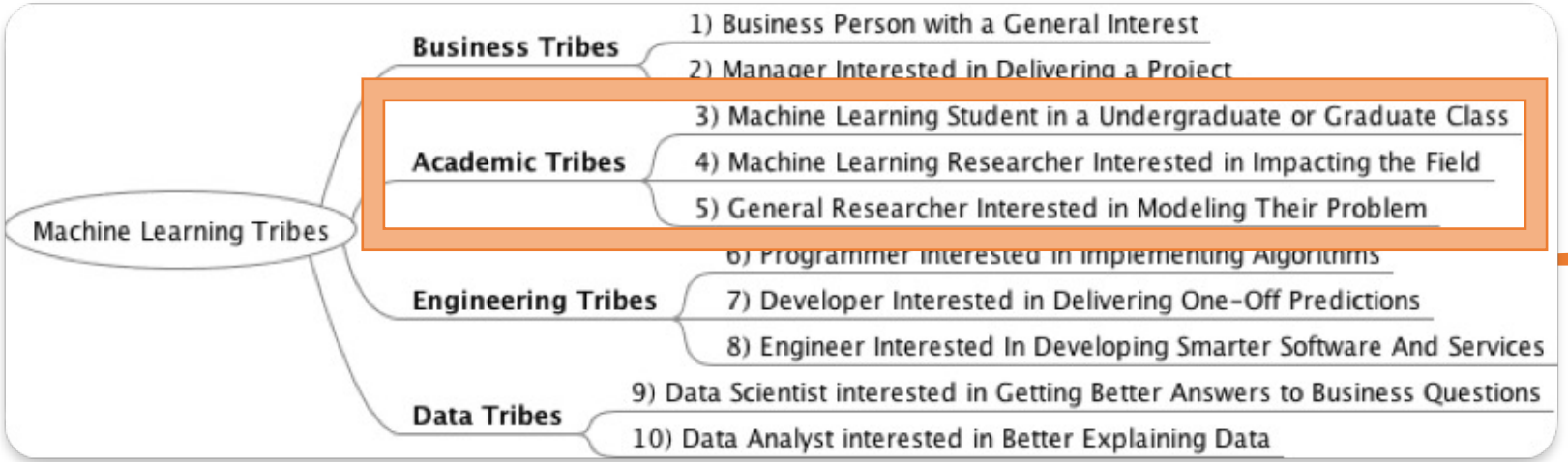
Makine öğrenimi nedir?



- Öğrenen programlarla ilgili bir bilgisayar bilimi alanı
- Makine öğrenimi alanı, deneyimle otomatik olarak gelişen bilgisayar programlarının nasıl oluşturulacağı sorusuyla ilgilenir. (Machine Learning, 1997.)
- Bu, aşağıdakiler gibi çeşitli öğrenme türlerini kapsayabilir:
  - Organizma popülasyonlarının evrimsel zaman içinde çevrelerine uyum sağlamayı nasıl "öğrendiğini" araştırmak için kod geliştirmek.
  - Beyindeki bir nöronun diğer nöronlardan gelen uyarana yanıt olarak nasıl "öğrendiğini" araştırmak için kod geliştirmek.
  - Karıncaların evlerinden besin kaynaklarına giden en uygun yolu nasıl "öğrendiklerini" araştırmak için kod geliştirmek.
- Geçmiş verilerdeki kalıpların nasıl "öğrenileceğini" araştırmak için kod geliştirmek.

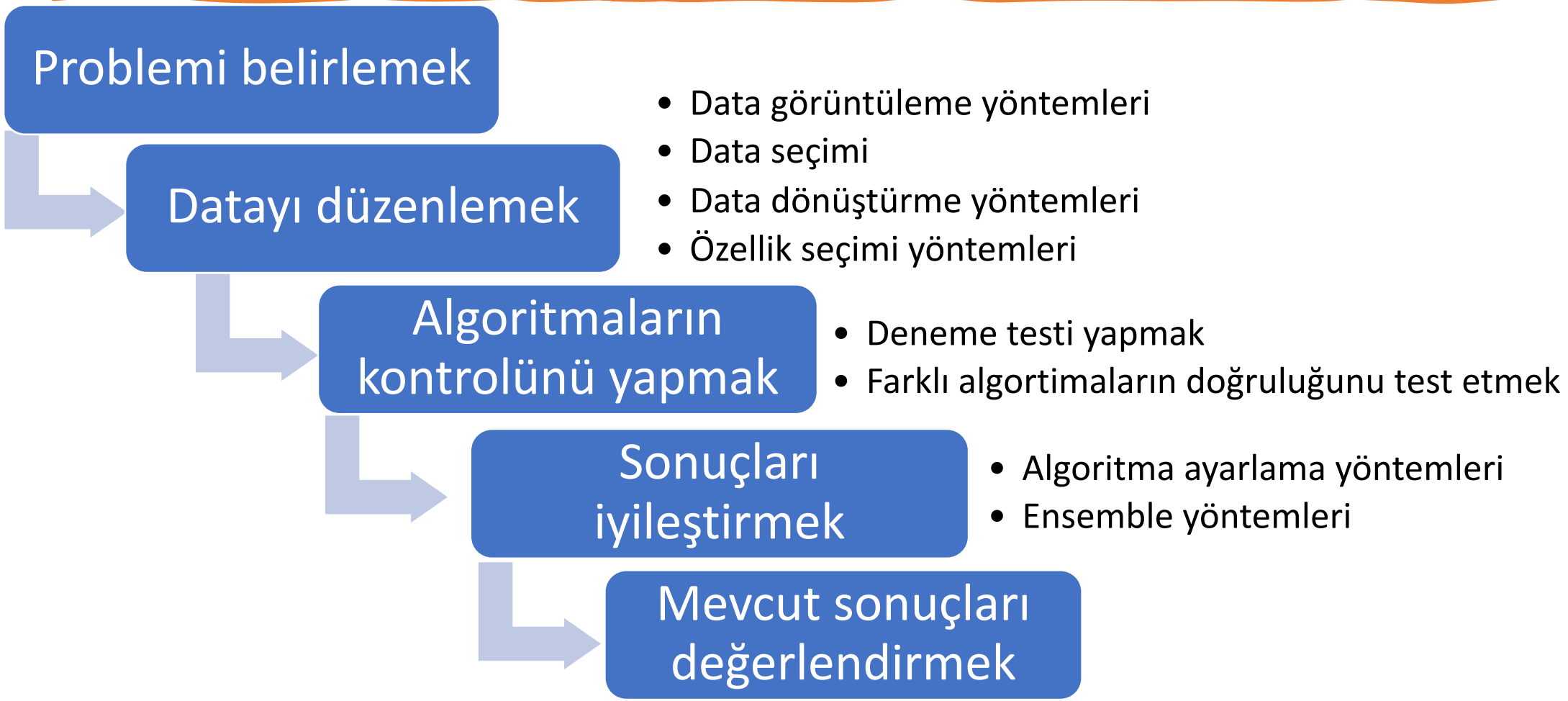
Makine öğrenimi nedir?

- Algoritmalar ve verilerle dolu büyüleyici ve güçlü bir çalışma alanı
- Makine öğrenimiyle ilgilenen pek çok farklı alan var ve her birinin farklı ihtiyaçları var. Makine öğreniminden ne istediğinizi anlamak ve bireysel çalışmanızı bu ihtiyaçlara göre uyarlamak önemlidir.
- Bunu yapmazsanız, kolayca tavşan deliğine düşebilir ve kaybolabilir, ilginizi kaybedebilir ve aradığınızı bulamayabilirsiniz.



- Araştırmacı, makine öğrenimiyle bir araç olarak ilgilenebilir. Kendi verilerini kullanarak tanımlayıcı veya tahmine dayalı bir model oluşturmak amaç olabilir.
- Genellikle model doğruluğuyla daha az ilgilenirler ve modelin açıklanabilirliğiyle daha çok ilgilenirler.
- Bu nedenle, doğrusal regresyon ve lojistik regresyon gibi istatistiklerden ödünç alınan daha basit ve iyi anlaşılan yöntemler tercih edilir.

Makine öğrenimi kontrol listesi



- Data görüntüleme yöntemleri
- Data seçimi
- Data dönüştürme yöntemleri
- Özellik seçimi yöntemleri

- Deneme testi yapmak
- Farklı algortimaların doğruluğunu test etmek

- Algoritma ayarlama yöntemleri
- Ensemble yöntemleri

Makine öğrenmesine nasıl başlamalıyım?

- **Adım 1:** Zihninizi hazırlayın.
  Makine öğrenimi uygulayabileceğinize inanın.
- **Adım 2:** Bir Süreç Seçin.
  Sorunları çözmek için sistemik bir süreç kullanın.
- **3. Adım:** Bir Araç Seçin.
  Seviyeniz için bir araç seçin ve bunu sürecinizle ilişkilendirin.
- **Adım 4:** Veri Kümeleri Üzerinde Pratik Yapın.
  Üzerinde çalışmak ve süreci uygulamak için veri kümelerini seçin.
- **Adım 5:** Bir Sonuç Oluşturun.
  Sonuçları toplayın ve becerilerinizi gösterin.

Genom bilimde ML

- ML araçlarının genomikte kullanımı henüz erken bir aşamada olmasına rağmen, araştırmacılar, belirli şekillerde yardımcı olan programlar geliştirmekten zaten yararlanmıştır.
  - Bir sıvı biyopsiden birincil kanser türünü belirlemek için makine öğrenimi tekniklerini kullanma.
  - Bir hastada belirli bir kanser türünün nasıl ilerleyeceğini tahmin etmek.
  - Makine öğrenimini kullanarak hastalığa neden olan genomik varyantları iyi huylu varyantlara kıyasla belirleme.
  - CRISPR gibi gen düzenleme araçlarının işlevini iyileştirmek için derin öğrenmeyi kullanma.
- Bunlar, ML yöntemlerinin genomik verilerdeki gizli kalıpları tahmin etmeye ve tanımlamaya yardımcı olmasının yalnızca birkaç yoludur. Bilim insanları ayrıca halk sağlığı çabalarına yardımcı olmak için grip ve SARS-CoV-2 virüslerinin genomlarında gelecekteki varyasyonları tahmin etmek için ML'yi kullanıyor.

Genom bilimde ML

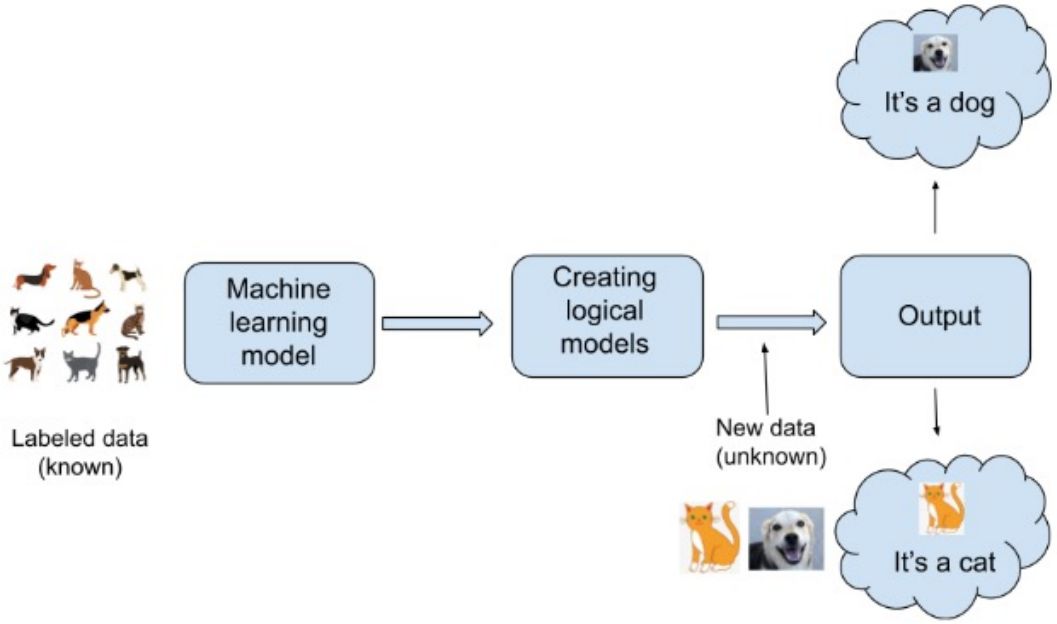
Aklınıza gelen kullanım alanları ne olabilir?

Genom bilimde ML

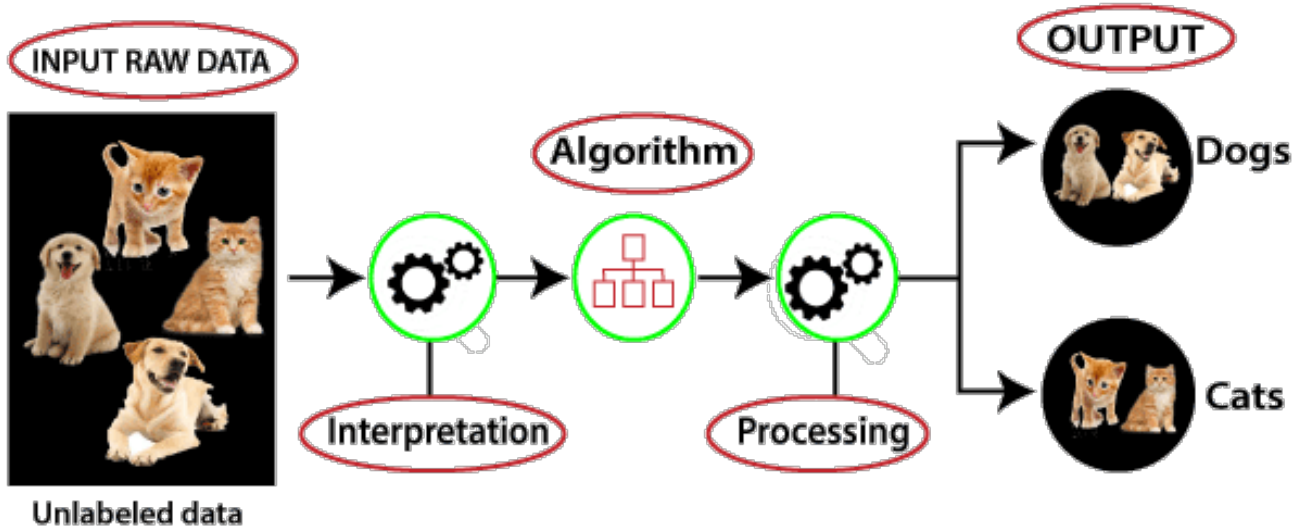
- Genom dizileme
- Tahmine dayalı test
- Farmakogenomik
- Genetik araştırma çalışmaları
- Gen modifikasyonu
- Gen ontolojisi

Makine öğrenmesi algoritmaları

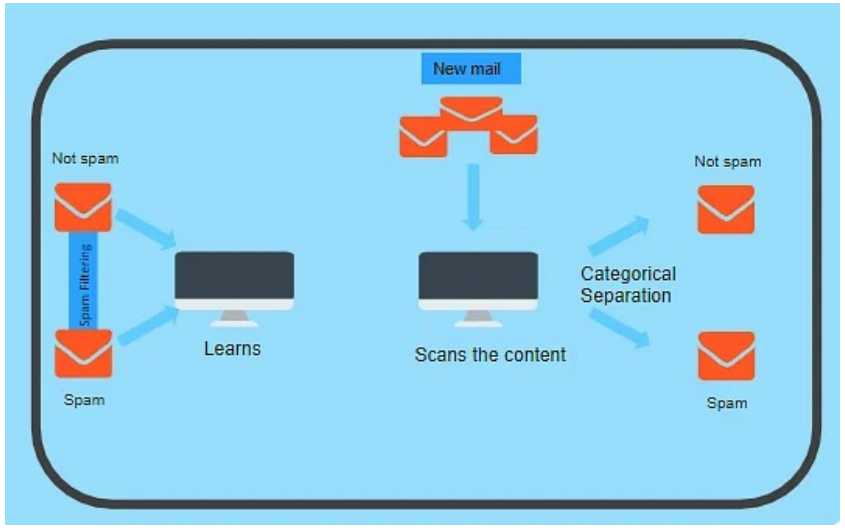
Gözetimli Öğrenme

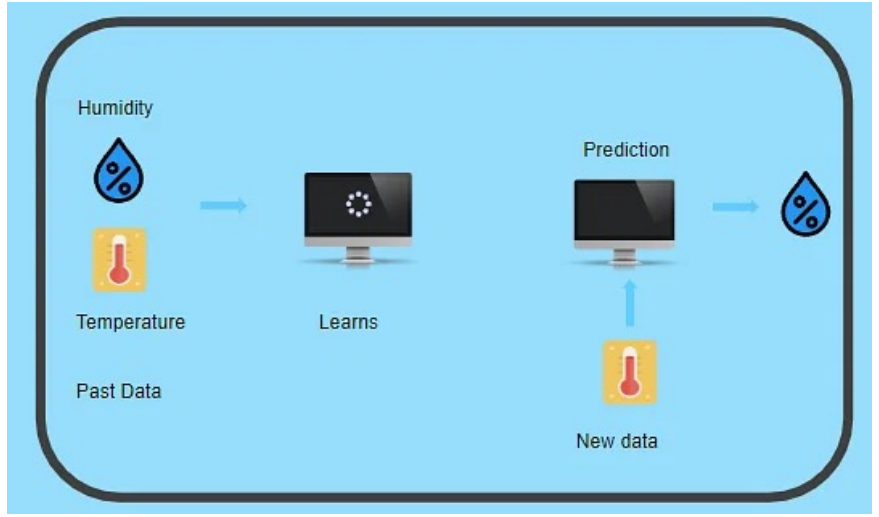


Gözetimsiz Öğrenme



Gözetimli Öğrenme





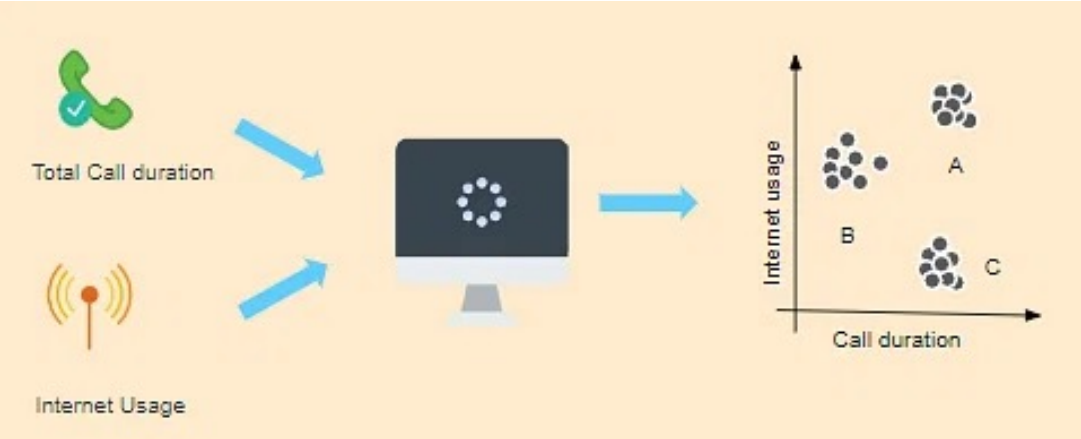
Sınıflandırma

- Rastgele Orman
- Karar ağaçları
- Lojistik Regresyon
- Destekli Vektör Makineleri

Regresyon

- Lineer Regresyon
- Regresyon Ağaçları
- Lineer Olmayan Regresyon
- Bayesian Lineer Regresyon

Gözetimsiz Öğrenme



Gruplandırma

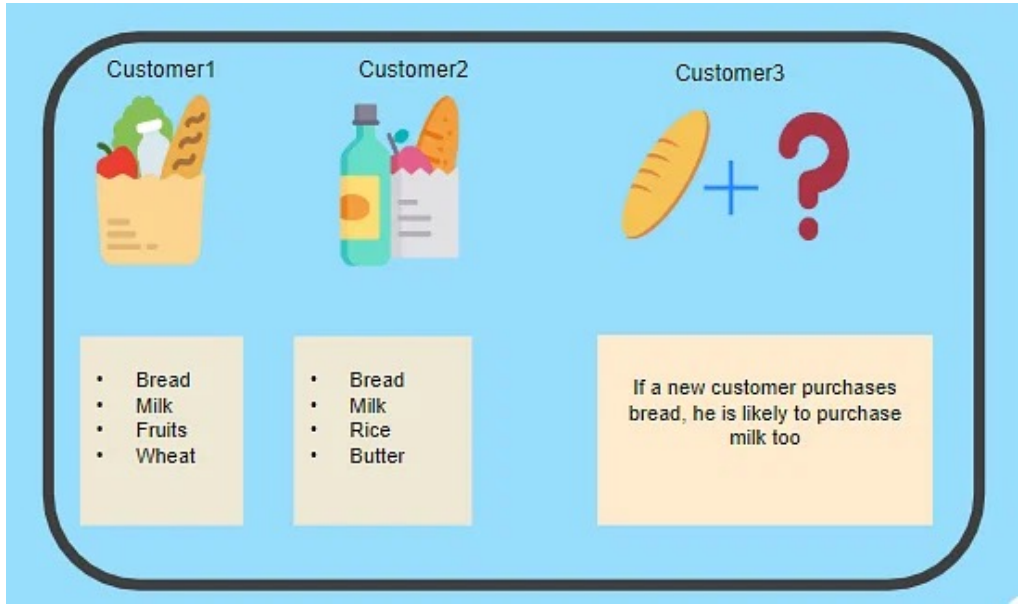
K-means kümeleme

Hiyerarşik kümeleme (Hierarchical clustering)

KNN (k-ya en yakın komşular)

PCA (Principle component analysis)

Sinir ağları (Neural Networks)



İlişkilendirme

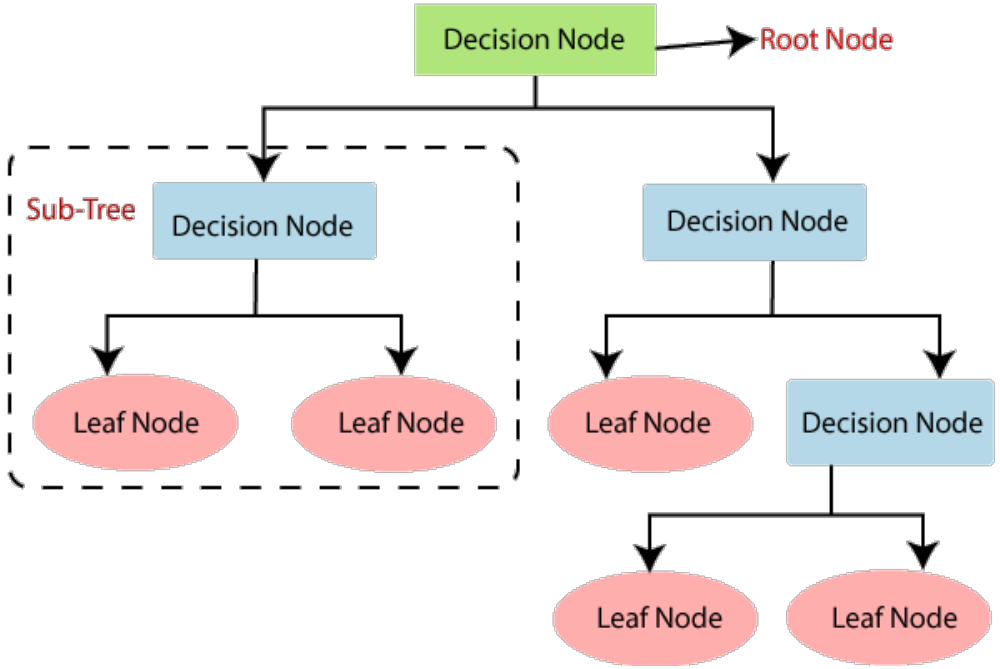
Gözetimli öğrenme aşamaları

1. Kullanılacak eğitici veri setine karar verme.
2. Bir eğitici setin verisini toparlama. Gerçek dünya örnekleri olmasına dikkat edilmelidir.
3. Öğrenilen fonksiyonun girdi özelliğini belirleme. Elde edilen fonksiyonun doğruluğu buna bağlıdır.
4. Öğrenilen fonksiyonun yapısını ve buna karşılık gelen öğrenme algoritmasını belirleyin.
5. Tasarımı tamamlayın. Toplanan eğitim setinde öğrenme algoritmasını çalıştırın.
6. Öğrenilen işlevin doğruluğunu değerlendirin. Parametre ayarı ve öğrenmeden sonra ortaya çıkan fonksiyonun performansı, eğitim setinden ayrı bir test setinde ölçülmelidir.

Gözetimli öğrenme algoritmaları

Karar ağaçları (Decision Tree)

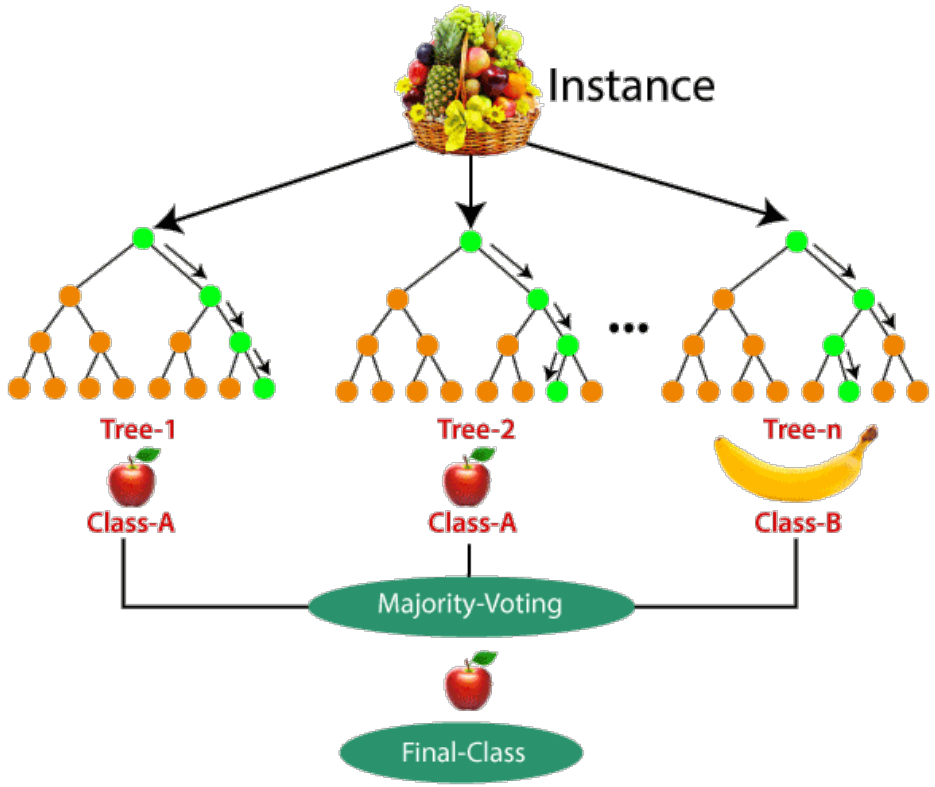
- Bir sınıflandırma veya regresyon karar ağacı, bir dizi gözlem hakkında sonuçlar çıkarmak için tahmine dayalı bir model olarak kullanılır.



Gözetimli öğrenme algoritmaları

Rastgele Orman (Random Forest)

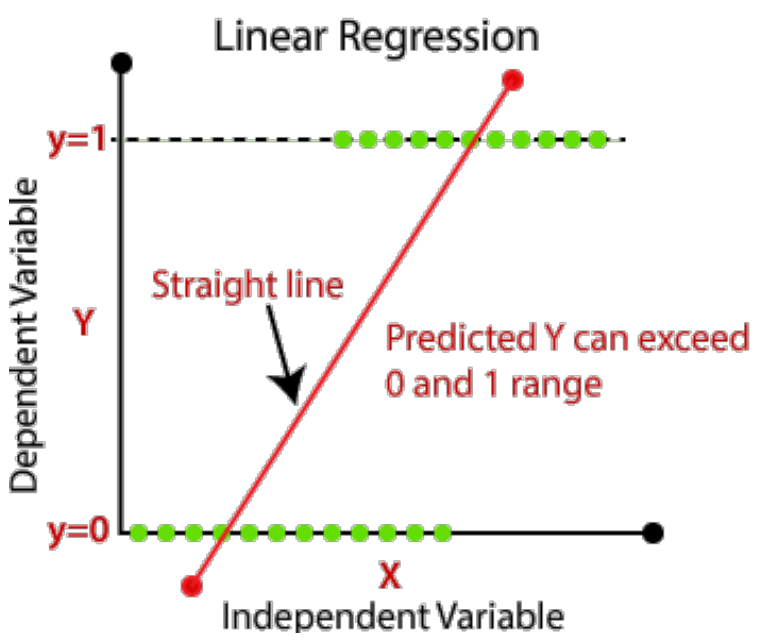
- Sınıflandırma için en yaygın kullanılan ML algoritması
- Her ağaç, toplam eğitim verisinin kabaca 2/3'ü (tam olarak %66) ile eğitilir. Durumlar, orijinal verilerden değiştirilerek rastgele çizilir. Bu örnek, ağacı büyütmek için eğitim seti olacaktır.
- Her ağaç için, kalan (%34) verileri kullanarak, yanlış sınıflandırma oranı - torbadan çıkma (Out-of-bag OOB) hata oranını hesaplanır. Sınıflandırma için genel OOB hata oranını belirlemek için tüm ağaçlardan hatalar toplanır.
- Rastgele orman algoritmasında iki parametre önemlidir:
  - Ormanda kullanılan ağaç sayısı (ntree)
  - Her ağaçta kullanılan rastgele değişken sayısı (mtry).

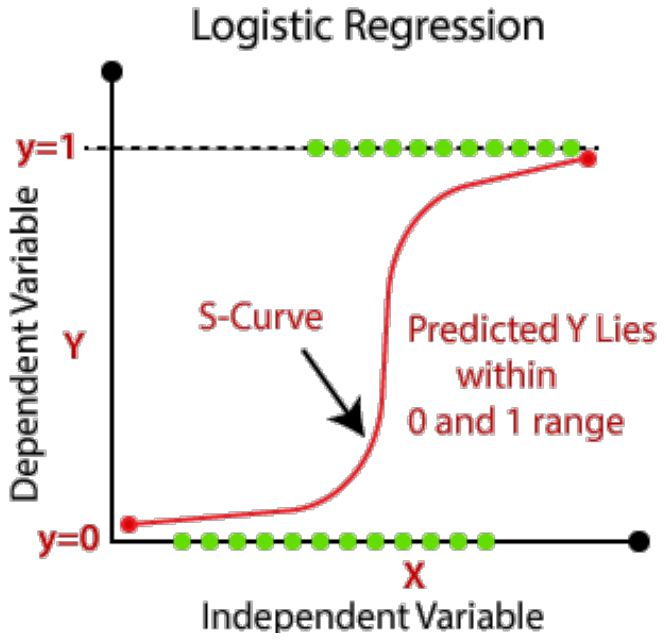


Gözetimli öğrenme algoritmaları

Lineer ve Lojistik Regresyon

İşletme alanı
Tahmin stokları



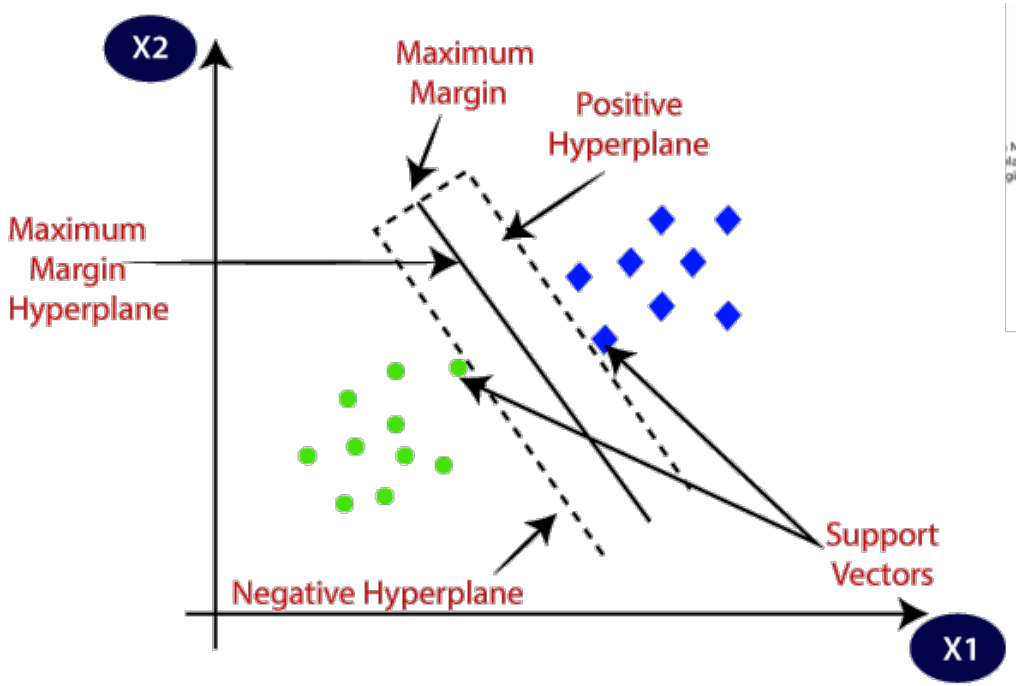


Sınıflandırma
Görüntü İşleme

Gözetimli öğrenme algoritmaları

Destekli Vektör makineleri (Support Vector Machine - SVM)

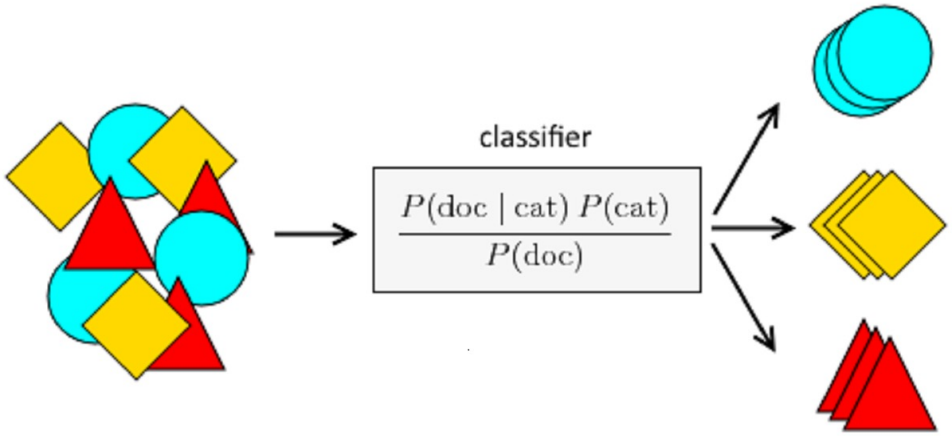
- Sınıflandırma için en yaygın kullanılan ML algoritması
- Regresyon için de kullanılabilir



Gözetimli öğrenme algoritmaları

Naïve Bayes Sınıflandırma

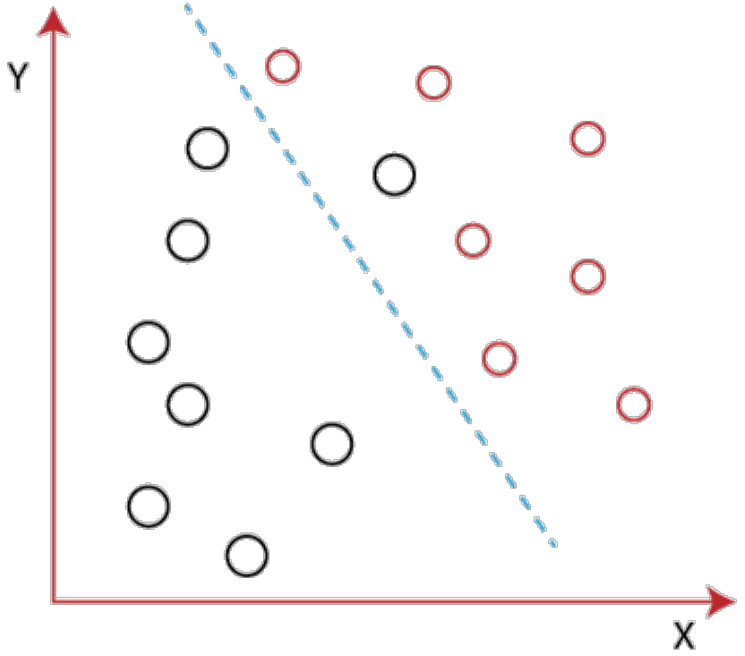
- Olasılığa dayalı bir sınıflandırıcıdır, yani bir nesnenin olasılığı temelinde tahmin yapar.
- Tüm özelliklerin bağımsız veya ilgisiz olduğunu varsayar, bu nedenle özellikler arasındaki ilişkiyi öğrenemez.
- Medikal data sınıflandırması ve gerçek-zaman tahminleri konularında kullanılabilir.



Gözetimli öğrenme algoritmaları

Lineer Diskriminant Analizi

- 2 boyutlu düzlemi boyutsal olarak 1 boyutlu düzleme indirger.
- Çoklu sınıflar arasındaki ayrılabilirliği maksimize etmeyi amaçlamaktadır.
- PCA ile benzerdir ama data arasındaki en fazla değişime odaklanmak yerine kategorileri oluşturup bunlar arasındaki farklara odaklanır.



Gözetimli öğrenme algoritmaları

Gözetimli Öğrenme

- **Karar ağaçları (Decision Tree)**
- **Destekli Vektör makineleri (Support Vector Machine - SVM)**
- **Rastgele Ormanlar (Random Forest- RF)**
- **Naïve Bayes**
- **Doğrusal regresyon**
- **Lojistik regresyon**
- **Lineer diskriminant analizi**

Uygulama

Sorular ??

Hadi başlayalım!

Works cited

- Algoritmaların resimleri https://www.javatpoint.com/ sitesinden alınmıştır.
- https://machinelearningmastery.com/start-here/
- https://archive.ics.uci.edu -> Example databases
- http://alexkychen.github.io/assignPOP/index.html