MAKİNE ÖĞRENMESİ VE GENOMİK ANALİZLER

M. Çisel Kemahlı Aytekin

mckemahli@gmail.com

Evrimsel Genombilim Uygulamalı Eğitim - 2024

DERS İÇERİĞİ

Genomik nedir?

Temel Makine Öğrenmesi (ML)

Genombilimde ML

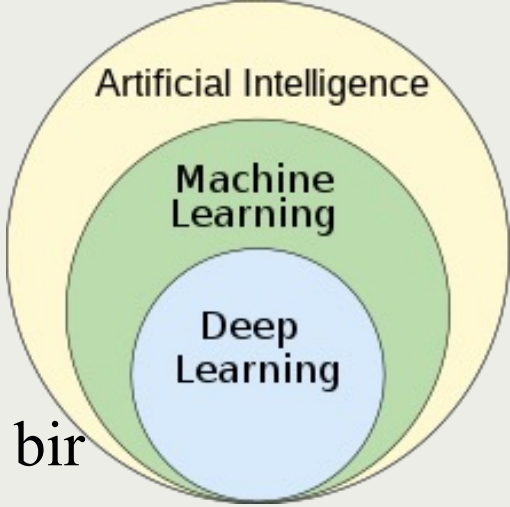
ML algoritmaları

Uygulama

GENOMİK

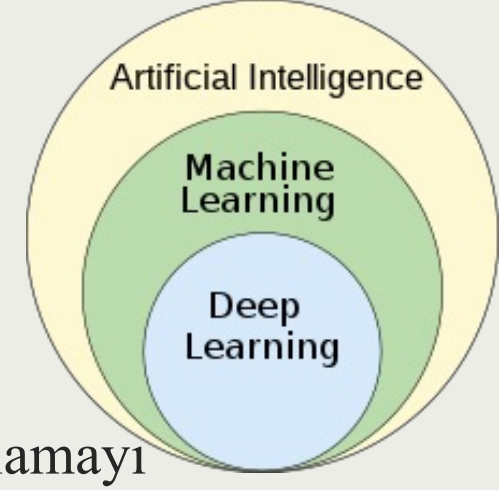
- Bir organizmanın tüm genetik materyalini, yani genomunu inceleyen biyoloji dalıdır.
- Genom, bir organizmanın büyümesini, gelişimini ve işlevlerini yöneten tüm genlerin ve diğer genetik bilgilerinin toplamıdır.
- Genomik, biyolojik araştırmalarda ve tıpta devrim yaratmıştır.
- İnsan Genomu Projesi gibi büyük ölçekli genomik projeler, hastalıkların genetik temellerini anlamamıza ve yeni tedavi yöntemleri geliştirmemize yardımcı olmuştur.

MAKİNE ÖĞRENMESİ



- Makine öğrenmesi (ML), bilgisayarların verilerden öğrenerek ve bu öğrenmeyi kullanarak tahminler veya kararlar almasına olanak tanıyan bir yapay zeka dalıdır.
- İstatistiksel yöntemler ve algoritmalar kullanarak verilerden kalıpları öğrenir ve bu bilgiyi yeni verilere uygulayarak sonuçlar çıkarır.
- Büyük veri analizinde ve otomatik karar verme süreçlerinde yaygın olarak kullanılmaktadır.
- Görüntü tanıma, doğal dil işleme, öneri sistemleri gibi birçok alanda etkili çözümler sunmaktadır.

MAKİNE ÖĞRENMESİ



- 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.
- 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.

ML KONTROL LİSTESİ

Problemi belirlemek

Datayı düzenlemek

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

Algoritmaların kontrolünü yapmak

- Deneme testi yapmak
- Farklı algortimaların doğruluğunu test etmek

Sonuçları iyileştirmek

- Algoritma ayarlama yöntemleri
- Ensemble yöntemleri

Mevcut sonuçları değerlendirmek

GENOMBİLİMDE ML

- Genomik veriler genellikle büyük, karmaşık ve çok boyutludur. DNA dizileri, gen ifadeleri, genotip verileri gibi çeşitli veri türlerini içerir. Bu veriler, biyolojik süreçlerin anlaşılması ve hastalıkların teşhis edilmesi için önemli bilgiler sağlar.
- Genomik verilerin analizi, geleneksel yöntemlerle zor ve zaman alıcı olabilir. Makine öğrenmesi, bu büyük ve karmaşık veri kümelerinden anlamlı kalıplar ve ilişkiler çıkarmak için güçlü bir araçtır.
- Makine öğrenmesi algoritmaları, genomik verilerdeki varyasyonları ve ilişkileri belirleyerek, hastalık risklerinin tahmin edilmesi, genetik varyantların fonksiyonel etkilerinin anlaşılması ve kişiselleştirilmiş tedavi planlarının geliştirilmesi gibi birçok uygulamada kullanılabilir.

GENOMBİLİMDE 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.

GENOMBİLİMDE ML

- **Kanser Teşhisi ve Tedavi Planlaması**
- Makine öğrenmesi, kanser teşhisi ve tedavi planlamasında büyük bir rol oynar.
- DNA dizileme verileri kullanılarak farklı tümör türleri sınıflandırılabilir. Örneğin, meme kanseri ile akciğer kanseri arasındaki farkı belirlemek.
- Gen ifadesi profilleri kullanılarak hastalığın ilerleme olasılığı tahmin edilebilir.

GENOMBİLİMDE ML

- **Genetik Varyantların Tespiti ve Analizi**
- Genetik varyantlar, bir bireyin genomunda meydana gelen küçük değişikliklerdir. Bu varyantların tespiti ve analizi, hastalık risklerinin belirlenmesinde ve kişiselleştirilmiş tıpta önemli bir rol oynar.
- Genetik varyantlar kullanılarak bireylerin belirli hastalıklara yatkınlıkları tahmin edilebilir. Örneğin, BRCA1 ve BRCA2 genlerindeki mutasyonların meme kanseri riskini artırdığı bilinmektedir.
- Varyantların gen ifadesi ve protein fonksiyonu üzerindeki etkileri analiz edilerek biyolojik anlamları çözümlenebilir.
- Farklı popülasyonlarda genetik varyantların dağılımı incelenerek, insan evrimi ve genetik çeşitlilik hakkında bilgi edinilebilir.

GENOMBİLİMDE ML

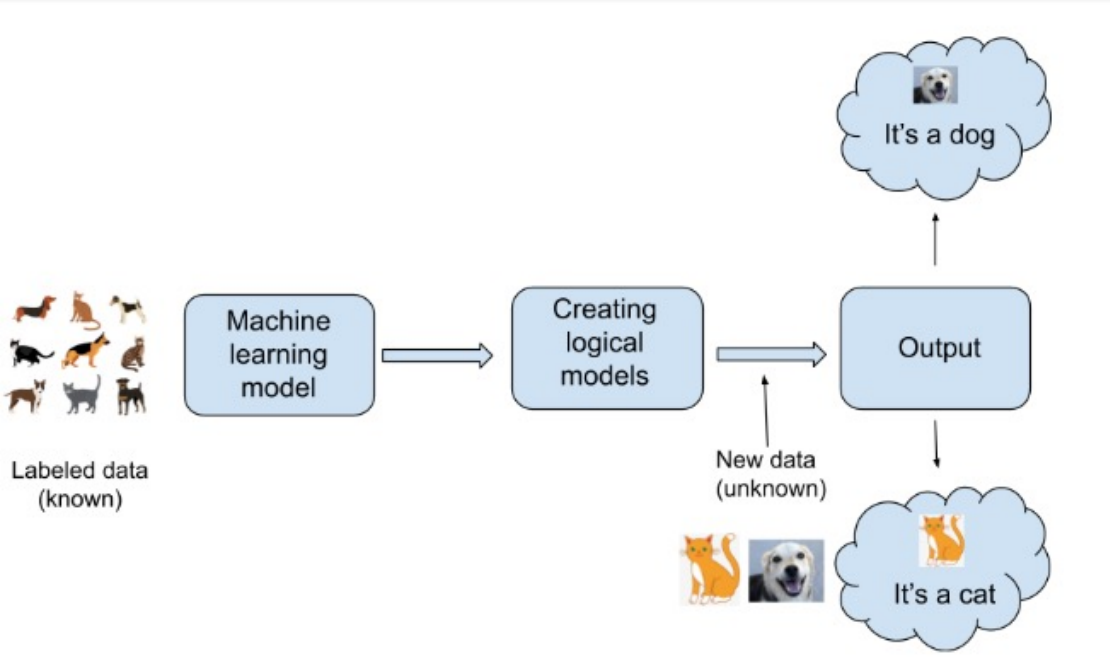
- **Kişiselleştirilmiş Tıp**
- Hastaların genetik, çevresel ve yaşam tarzı bilgilerine dayalı olarak özel tedavi planları geliştirmeyi amaçlar. Bu, hastaların en iyi sonuçları elde etmelerini sağlamak için tedavilerin bireyselleştirilmesini içerir.
- Hastaların genetik profillerine dayalı olarak ilaçların etkinlik ve yan etki profillerinin tahmin edilmesi. Örneğin, belirli genetik varyantların varlığına göre ilaç dozajlarının ayarlanması.
- Bireylerin genetik profillerine dayalı olarak özel diyet ve yaşam tarzı önerileri sunulması. Örneğin, belirli genetik varyantların diyabet riskini artırdığı bireylere özel beslenme planları önerilmesi.

Aklınıza gelen kullanım alanları ne olabilir?

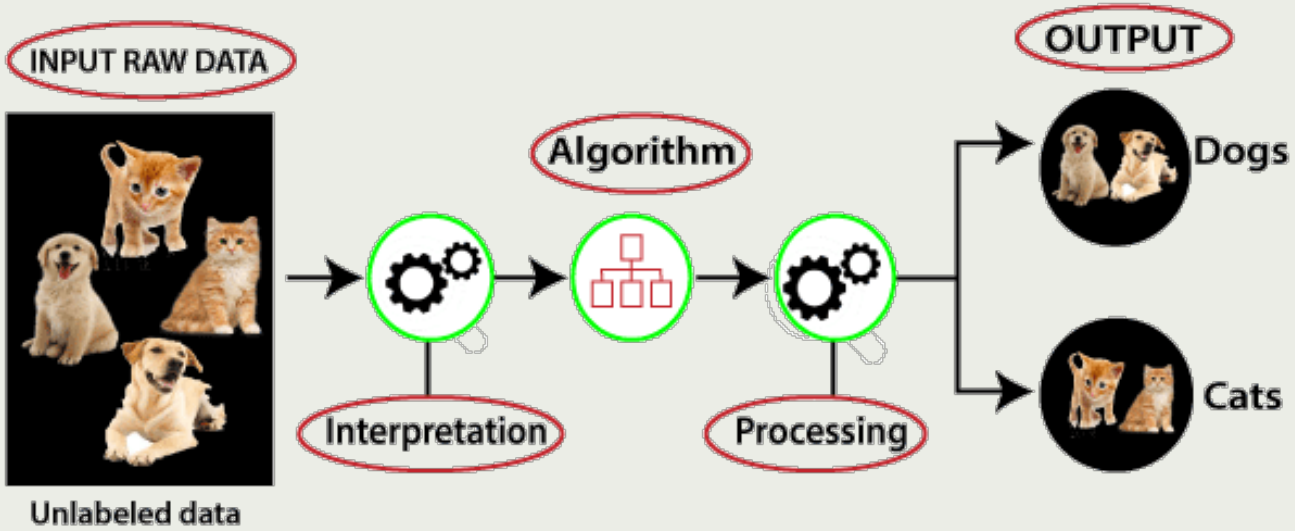
- Genom dizileme
- Tahmine dayalı test
- Farmakogenomik
- Genetik araştırma çalışmaları
- Gen modifikasyonu
- Gen ontolojisi

ML ALGORİTMALARI

Denetimli Öğrenme



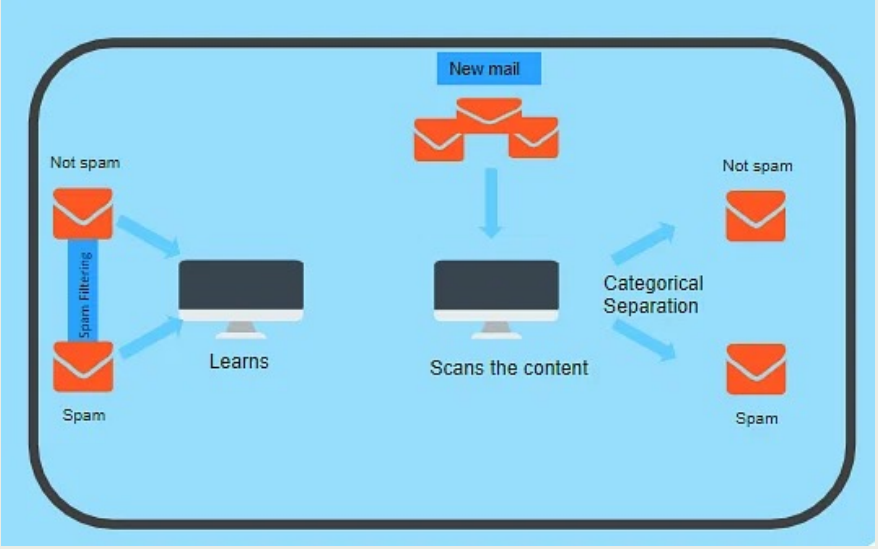
Denetimsiz Öğrenme



ÖĞRENME ADIMLARI

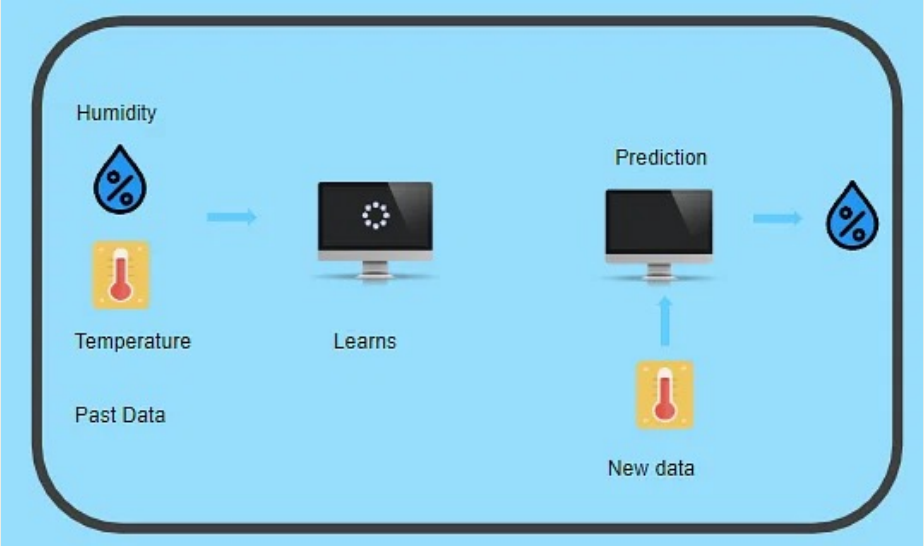
- **Veri Hazırlama**:
  - Verilerin toplanması ve temizlenmesi.
  - Eksik verilerin doldurulması ve veri ön işleme.
- **Özellik Seçimi**:
  - En önemli özelliklerin belirlenmesi ve seçilmesi.
  - Özellik mühendisliği ile anlamlı özellikler çıkarılması.
- **Model Eğitimi**:
  - Seçilen algoritmalarla modelin eğitilmesi.
  - Hiperparametre ayarlamaları ile model performansının optimize edilmesi.
- **Model Değerlendirme**:
  - Modelin performansının değerlendirilmesi ve sonuçların yorumlanması.
  - Performans metriklerinin hesaplanması (örneğin, doğruluk, F1 skoru, ROC eğrisi, siluet skoru, küme içi varyans).
- **Modelin Yorumlanması ve Uygulama**:
  - Modelin sonuçlarının biyolojik anlamının yorumlanması.
  - Modelin gerçek dünya genomik veri setlerine uygulanması.

DENETİMLİ ÖĞRENME



Sınıflandırma

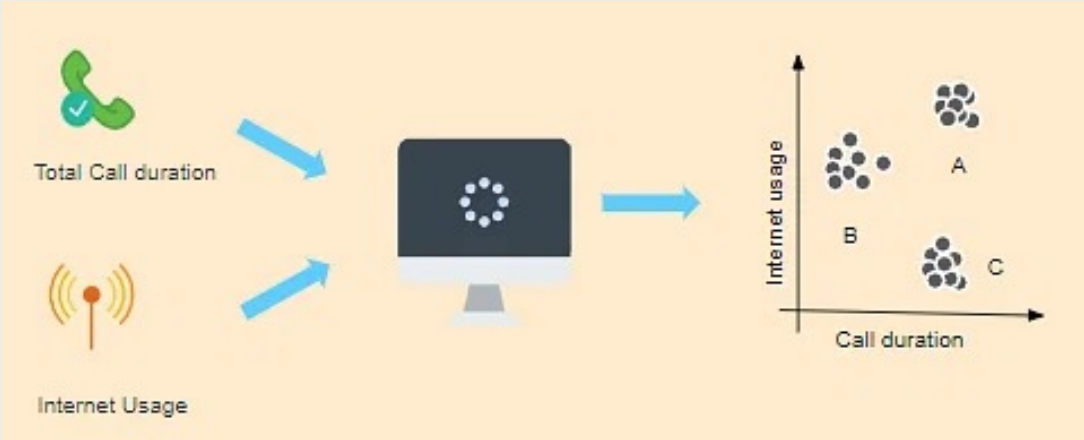
- Karar ağaçları
- Rastgele Orman
- Destekli Vektör Makineleri
- Naïve Bayes
- Doğrusal Ayırma Analizi (LDA)



Regresyon

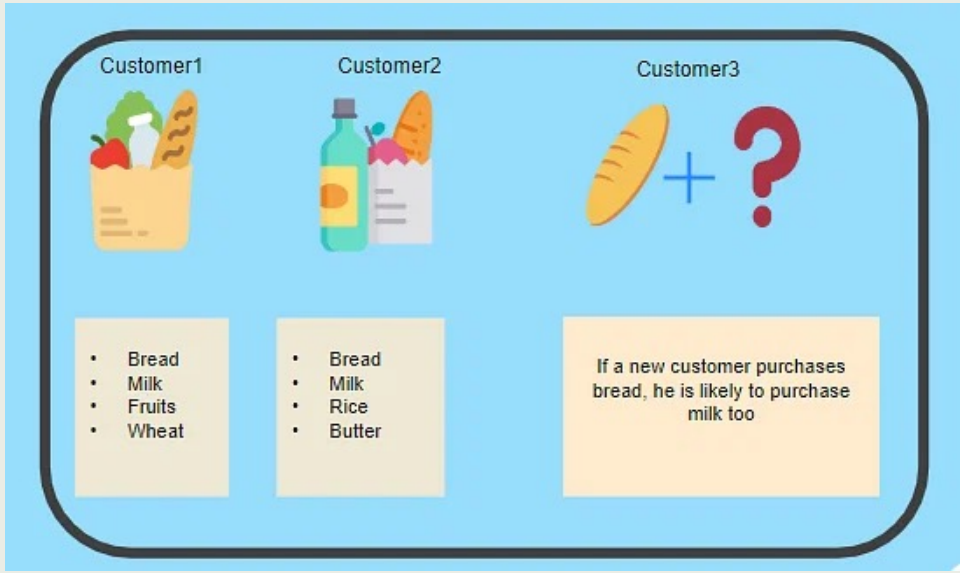
- Lojistik Regresyon
- Lineer Regresyon

DENETİMSİZ ÖĞRENME



Gruplandırma

- K-means kümeleme
- Hiyerarşik kümeleme (Hierarchical clustering)
- KNN (k-ya en yakın komşular)
- PCA (Principle component analysis)
- Sinir ağları (Neural Networks)

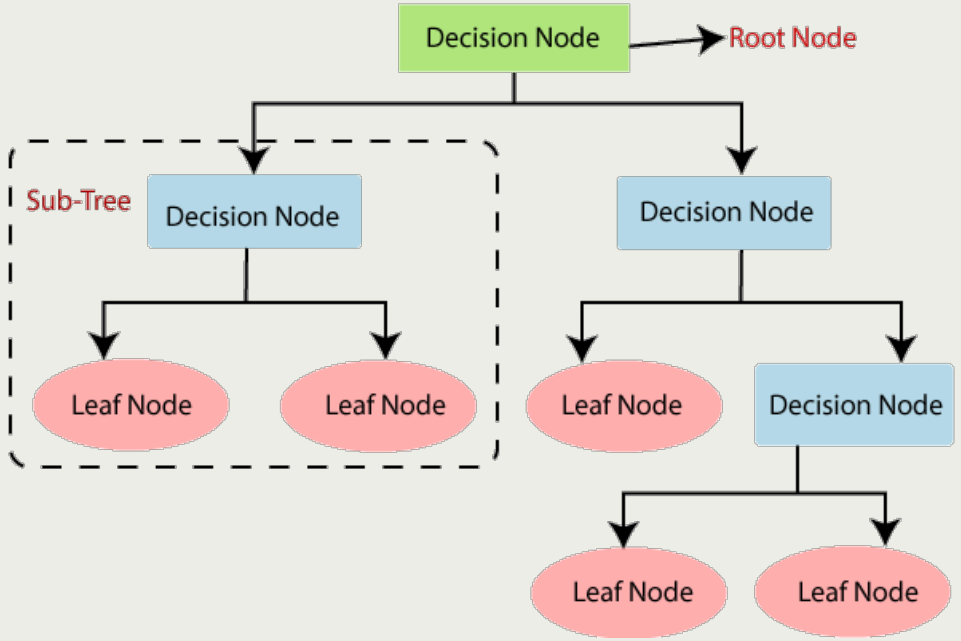


İlişkilendirme

DENETİMLİ ÖĞRENME ALGORİTAMALARI

Karar ağaçları (Decision Tree)

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

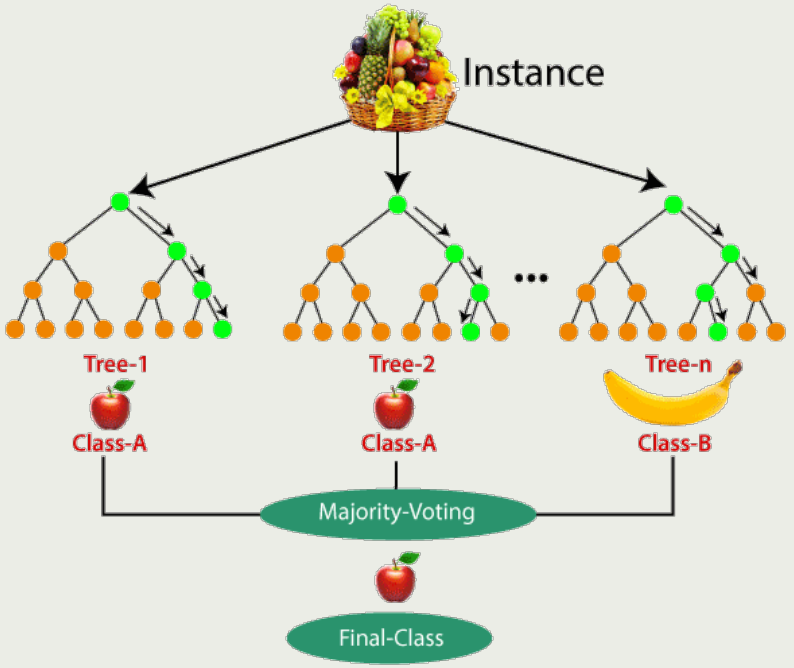


•Genetik verilerde hastalık sınıflandırması.
•Genetik varyantların etkilerinin tahmini.
•Gen ekspresyon verilerinde biyomarker belirleme.

DENETİMLİ ÖĞRENME ALGORİTAMALARI

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.

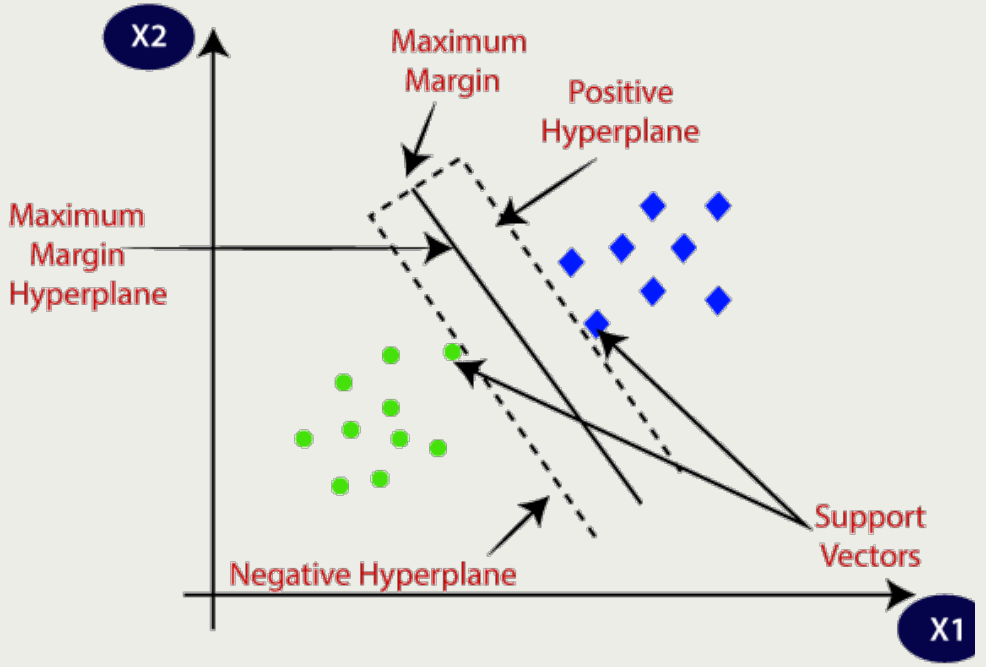


•Genetik verilerin sınıflandırılması ve regresyon analizi.
•Genetik varyantların etkilerinin tahmini.
•Gen ekspresyon verilerinde biyomarker belirleme.

DENETİMLİ ÖĞRENME ALGORİTAMALARI

Destekli Vektör makineleri (Support Vector Machine - SVM)

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

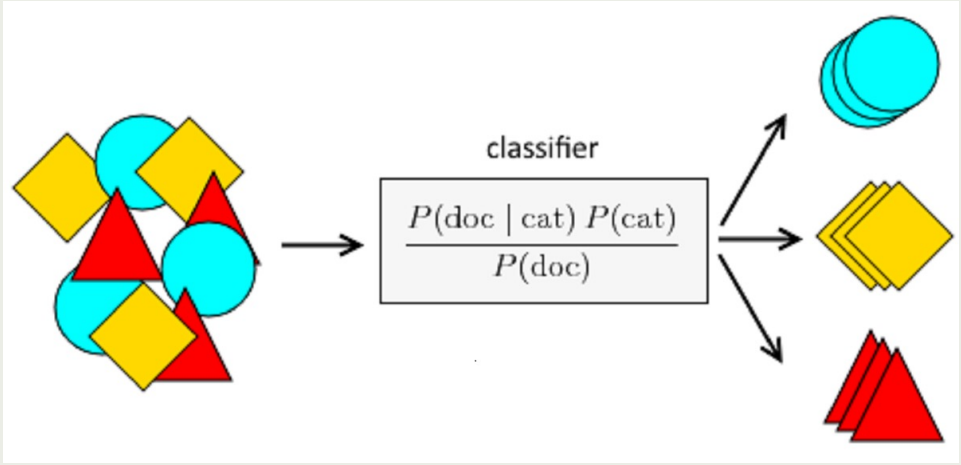


•Genetik mutasyonların farklı kanser türlerine göre sınıflandırılması.

DENETİMLİ ÖĞRENME ALGORİTAMALARI

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.

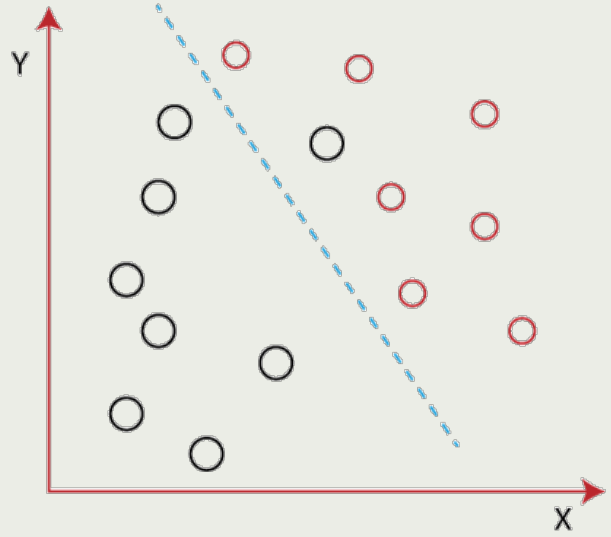


•Genomik verilerden hastalık risklerinin tahmini.

DENETİMLİ ÖĞRENME ALGORİTAMALARI

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.

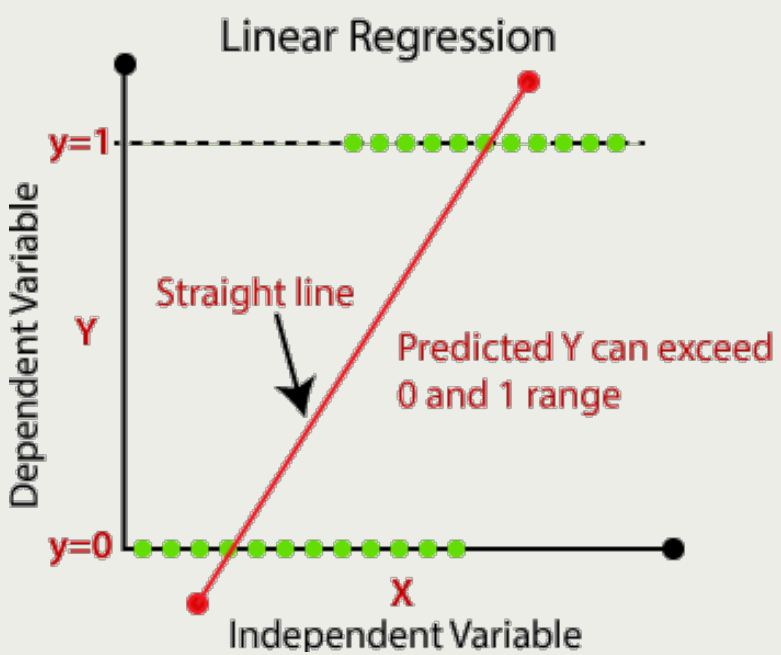


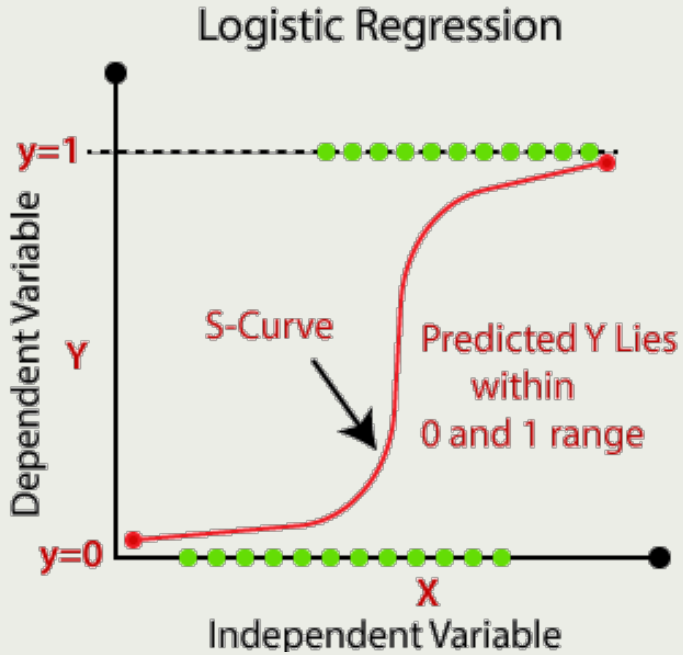
•Genetik verilerin sınıflandırılması (örneğin, hastalık türleri).
•Gen ekspresyon verilerinin analizi.
•Kanser türlerinin ve alt tiplerinin ayrılması.

DENETİMLİ ÖĞRENME ALGORİTAMALARI

Lineer ve Lojistik Regresyon

İşletme alanı
Tahmin stokları





Sınıflandırma
Görüntü İşleme

•Genetik risk faktörlerinin hastalık sonuçları ile ilişkilendirilmesi.

MODEL EĞİTİMİ VE DEĞERLENDİRME

- **Eğitim Veri Seti ve Test Veri Seti Oluşturma**
- Makine öğrenmesi modelleri, genellikle veri setleri eğitim ve test olarak ikiye ayrılarak eğitilir ve değerlendirilir. Eğitim veri seti modeli eğitmek için kullanılırken, test veri seti modelin performansını değerlendirmek için kullanılır.
- **Veri Ayırma (Data Splitting)**: Verilerin belirli bir yüzde ile eğitim ve test setlerine bölünmesi. Yaygın oranlar 70/30 veya 80/20'dir.

- **Model Seçimi ve Parametre Ayarlamaları**
- Farklı makine öğrenmesi algoritmalarının ve modellerin seçilmesi ve bu modellerin hiperparametrelerinin optimize edilmesi, model performansını önemli ölçüde etkiler.
- **Hiperparametre Ayarlamaları (Hyperparameter Tuning)**: Grid search, random search veya Bayesian optimization gibi yöntemlerle modelin hiperparametrelerinin optimize edilmesi.

MODEL EĞİTİMİ VE DEĞERLENDİRME

- **Modelin Doğruluğunu Değerlendirme ve Çapraz Doğrulama**
- Bu yöntemler, modelin genelleme yeteneğini ve overfitting olup olmadığını anlamaya yardımcı olur.
- **Çapraz Doğrulama (Cross-Validation)**: Verilerin farklı bölümlerini eğitim ve test setleri olarak kullanarak modelin performansını değerlendirme yöntemi.
- **Performans Metrikleri**: Modelin doğruluğunu ölçmek için kullanılan metrikler. Örneğin, doğruluk (accuracy), ROC eğrisi (ROC curve), F1 skoru (F1 score).

- **Modelin Sonuçlarının Yorumlanması**
- Modelin çıktılarının ve performans metriklerinin yorumlanması, modelin nasıl çalıştığını ve hangi özelliklerin önemli olduğunu anlamaya yardımcı olur.
- **Önemli Özelliklerin İncelenmesi**: Modelin hangi özelliklere daha fazla önem verdiğini anlamak.
- **Karar Ağaçları ve Model Şeffaflığı**: Modellerin nasıl çalıştığını ve karar süreçlerini açıklamak için kullanılan yöntemler.

GENOMİKTE MAKİNE ÖĞRENMESİNİN FAYDALARI VE GELECEĞİ

1. Daha Hızlı ve Doğru Teşhis
2. Kişiselleştirilmiş Tedavi ve İlaç Geliştirme
3. Büyük Veri Analizi ve Görselleştirme

Gelecekteki Potansiyel Uygulamalar

1. Genomik Tıp ve Kişiselleştirilmiş Sağlık Hizmetleri
2. Genomik Araştırmalar ve Biyolojik Keşifler
3. Kamu Sağlığı ve Epidemiyoloji

KARŞILAŞILAN ZORLUKLAR VE ÇÖZÜM YOLLARI

1. Büyük Veri Yönetimi ve Depolama
2. Veri Kalitesi ve Eksik Veriler
3. Hesaplama Gücü ve Zaman Gereksinimleri
4. Veri Gizliliği ve Güvenlik
5. Yorumlama ve Anlamlandırma

VERİ HAZIRLAMA VE ÖN İŞLEME

- **Veri Temizleme ve Eksik Verilerin Doldurulması:** Genomik veriler sıklıkla eksik veya hatalı olabilir. Bu verilerin analizden önce temizlenmesi ve eksik değerlerin uygun yöntemlerle doldurulması gereklidir.
- **Verilerin Normalizasyonu ve Ölçeklendirilmesi:** Genomik verilerdeki farklılıkları ortadan kaldırmak ve verileri karşılaştırılabilir hale getirmek için normalizasyon ve ölçeklendirme işlemleri uygulanır.
- **Özellik Mühendisliği ve Seçimi:** Makine öğrenmesi modellerinin performansını artırmak için verilerden anlamlı özellikler çıkarmak ve en önemli olanları seçmek önemlidir.

- 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