daire, simge, sembol, logo, meneviş mavisi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**BURSA ULUDAĞ ÜNİVERSİTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ**

**2023 – 2024 Eğitim Öğretim Yılı Güz Dönemi**

**ALGORİTMA ANALİZİ**

**PROJE KONUSU**

Makine Öğrenmesi Algoritmaları (Logistic Regression ve KNN)

**PROJE ÜYELERİ**

**AD NO**

ALİ CAN AYHAN 032190050

RÜZGAR İREN 032190040

BEDİRHAN ÇAT 03210038

**Algoritma nedir, nerede kullanılır?**

Algoritma, belirli bir görevi tamamlamak veya belirli bir sorunu çözmek için adım adım uygulanacak prosedürler veya izlenecek bir dizi kuraldır. Bir bilgisayar programı, ayrıntılı bir algoritma örneğidir.

Algoritmalar aslında hayatımızın her evresinde yer alıyor. Kek pişirme tarifi, bir matematik problemini çözmek için kullandığımız yöntem ve çamaşır yıkama işlemi de birer algoritma örneği olarak sayılabilir.

**Algoritmalar nasıl çalışır?**

Örnek olarak bir arama motoru algoritmasını ele alalım. Arama motoru, girdi olarak anahtar sözcükler ile aramaları ele alır ve ilgili web sayfalarını bulmak için veri tabanını taradıktan sonra sonuçları size gösterir. Bu işlem sizin için anlık olsa da arka tarafta gerçekleşen milyonlarca veri adımı ve işlem bulunmaktadır.

Algoritmaları kolayca görselleştirebilirsiniz. Girdi, sırayla ele alınması gereken adımlara ve sorulara yol açar. Akış şemasının her bölümü tamamlandığında ise üretilen sonuç çıktıdır.

Şifreleme algoritmaları verileri korumak için farklı eylemlere başvurabilir. Örneğin ABD Savunma Bakanlığı, Veri Şifreleme Standardı (Defense's Data Encryption Standard - DES) isimli gizli bir anahtar algoritmasını kullanarak verileri şifreler ve şifreleri çözer. Algoritma yeterince karmaşık olduğu sürece anahtarı olmayan hiç kimse verilerin şifresini çözemez.

**Algoritmalar ne için kullanılır?**

Algoritma bir sonuca ulaşmak için takip edilmesi gereken işlem basamaklarını belirlemede kullanılır. Örneğin, 3 ve 5 sayılarının ortalamasını bulmak istediğinizde önce (3+5) işlemini yaptıktan sonra sırada “çıkan sonuç/2” işlemini yapmanız gerekmektedir.

Bilgisayar biliminde algoritmalar yazılımdaki her yerdeler.

Örneklendirecek olursak, bilgisayarda offline olarak yapılan tüm işlemler, oynanan oyunlar, çalıştırılan programlar, web siteleri, mobil uygulamalar, robotlar… Kısacası sanal ortamdaki her işlem algoritmalar ile gerçekleşiyor.

**Günlük Hayatta Algoritmalar**

Günlük yaşamda biz farkında olmasak da algoritmalar hayatımızın büyük bir çoğunluğunda karşımıza çıkar. Örneğin; bir mobilyanın kurulum detaylarına ilişkin yayınlanan kurulum şeması adım adım çözüm yöntemlerini sunarak algoritmalar sistemine başvurur. Aynı şekilde bir yemek tarifi de adımlarla sunduğu çözüm yöntemlerinin akabinde bir algoritma mantığını günlük yaşama yansıtır.

Algoritmalar günlük yaşamı kolaylaştıran bir sistemler bütünüdür. Problemlerin aşamalar şeklinde çözüm yöntemlerinin irdelenmesi ve mantıksal düşünce çerçevesi sorunlar karşısında yapıcı çözümlerin bulunabilmesine olanak sağlar.

**Logistic Regression Nedir?**

Logistic regresyon, özellikle ikili sınıflandırma problemlerini çözmek için kullanılan istatistiksel bir yöntemdir. Temel olarak, bağımsız değişkenlerle bağımlı bir değişken arasındaki ilişkiyi modellemek için kullanılır. Ancak, bu ilişkiyi lineer bir denklemle ifade etmek yerine, logistic regresyon olasılık değerlerini kullanır.

Örneğin, web sitesi ziyaretçinizin alışveriş sepetindeki ödeme düğmesine tıklayıp tıklamayacağını tahmin etmek istediğinizi varsayalım. Lojistik regresyon analizi, web sitesinde harcanan zaman ve sepetteki ürün sayısı gibi geçmiş ziyaretçi davranışlarına bakar. Geçmişte, ziyaretçiler sitede beş dakikadan fazla zaman geçirdiyse ve sepete üçten fazla ürün eklediyse ödeme düğmesine tıkladıklarını belirler. Lojistik regresyon işlevi bu bilgiyi kullanarak daha sonra yeni bir web sitesi ziyaretçisinin davranışını tahmin edebilir

Logistic regresyon, sigmoid fonksiyonu kullanarak çıktıyı sınırlar ve bu sayede olasılık değerlerini elde eder. Eğitim süreci, modelin belirli bir veri setine ne kadar iyi uydurulduğunu ölçer ve optimal parametreleri bulmaya çalışır.

Sonuç olarak, logistic regresyon, ikili sınıflandırma problemlerinde yaygın olarak kullanılan bir yöntemdir ve özellikle istatistiksel analiz ve makine öğrenmesi alanlarında önemli bir rol oynar.

**Lojistik regresyon neden önemlidir?**

Lojistik regresyon, yapay zeka ve makine öğrenimi (AI/ML) alanında önemli bir tekniktir. ML modelleri, insan müdahalesi olmadan karmaşık veri işleme görevlerini gerçekleştirmek için eğitebileceğiniz yazılım programlarıdır. Lojistik regresyon kullanılarak oluşturulan ML modelleri, kuruluşların iş verilerinden eyleme dönüştürülebilir öngörüler elde etmelerine yardımcı olur. Bu bilgileri operasyonel maliyetleri azaltmak, verimliliği artırmak ve daha hızlı ölçeklendirmek amacıyla tahmine dayalı analiz için kullanabilirler. Örneğin, işletmeler, çalışanların elde tutulmasını artıran veya daha kârlı ürün tasarımına yol açan kalıpları ortaya çıkarabilir.

Diğer ML tekniklerine göre lojistik regresyon kullanmanın bazı avantajları aşağıda verilmiştir.

### ****Basitlik****

Lojistik regresyon modelleri matematiksel olarak diğer ML yöntemlerine göre daha az karmaşıktır. Bu nedenle, ekibinizdeki hiç kimsenin derinlemesine ML uzmanlığı olmasa bile bunları uygulayabilirsiniz.

### ****Hız****

Lojistik regresyon modelleri, bellek ve işlem gücü gibi daha az hesaplama kapasitesine ihtiyaç duydukları için büyük hacimli verileri yüksek hızda işleyebilir. Bu da onları, ML projelerine başlayan kuruluşların hızlı kazançlar elde etmesi için ideal kılar.

### ****Esneklik****

Lojistik regresyon, iki veya daha fazla sınırlı sonucu olan soruların yanıtlarını bulmak için kullanılabilir. Ayrıca verileri önceden işlemek için de kullanılabilir. Örneğin, banka işlemleri gibi çok çeşitli değerlere sahip verileri lojistik regresyon kullanarak daha küçük, sınırlı bir değer aralığında sıralayabilirsiniz. Daha sonra daha doğru analiz için diğer ML tekniklerini kullanarak bu küçük veri kümesini işleyebilirsiniz.

### ****Görünürlük****

Lojistik regresyon analizi, geliştiricilere dahili yazılım süreçlerinde diğer veri analizi tekniklerinden daha fazla görünürlük sağlar. Hesaplamalar daha az karmaşık olduğundan sorun giderme ve hata düzeltme de daha kolaydır

## Lojistik regresyon uygulamaları nelerdir?

Lojistik regresyon birçok farklı sektörde birkaç gerçek dünya uygulamasına sahiptir.

### ****Üretim****

İmalat şirketleri, makinelerde parça arızası olasılığını tahmin etmek için lojistik regresyon analizini kullanır. Daha sonra gelecekteki arızaları en aza indirmek için bu tahmine dayalı olarak bakım programları planlarlar.

### ****Sağlık hizmetleri****

Tıbbi araştırmacılar, hastalarda hastalık olasılığını tahmin ederek önleyici bakım ve tedaviyi planlar. Aile öyküsünün veya genlerin hastalıklar üzerindeki etkisini karşılaştırmak için lojistik regresyon modelleri kullanırlar.

### ****Finans****

Finansal şirketlerin dolandırıcılık için finansal işlemleri analiz etmesi ve kredi başvurularını ve sigorta uygulamalarını risk açısından değerlendirmesi gerekir. Lojistik regresyon modellerinin yüksek riskli veya düşük riskli ve dolandırıcılık olan ya da olmayan gibi ayrı sonuçları olduğundan bu sorunlar lojistik regresyon modeli için uygundur.

### ****Pazarlama****

Çevrimiçi reklamcılık araçları, kullanıcıların bir reklama tıklayıp tıklamayacağını tahmin etmek için lojistik regresyon modelini kullanır. Sonuç olarak pazarlamacılar, farklı kelimelere ve resimlere verilen kullanıcı yanıtlarını analiz edebilir ve müşterilerin etkileşimde bulunacağı yüksek performanslı reklamlar oluşturabilir.

## **Lojistik regresyon modeli nasıl çalışır**?

Logistic regresyon modeli, bağımlı değişkenin (genellikle ikili sınıflandırma durumunda) olasılık değerini tahmin etmek için kullanılır.

**Logistic Regression Formülleri:**

Logistic regresyon modelinin matematiksel formülü, bağımlı değişkenin olasılık değerini hesaplamak için kullanılır. Temel logistic regresyon formülü şu şekildedir;

**yazı tipi, metin, çizgi, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu**



Burada;

P(Y=1) ,bağımlı değişkenin 1 sınıfına ait olma olasılığını temsil eder.

**e** , doğal logaritma tabanı (Euler sayısı) olan matematiksel sabittir.

**z**, bağımsız değişkenlerin doğrusal kombinasyonunu temsil eder.

Bu formül, bağımsız değişkenlerin ağırlıkları (β katsayıları) ile doğrusal bir kombinasyonunu alır ve bu değeri sigmoid fonksiyonuna geçirir. Sigmoid fonksiyonu, sonsuz aralıktaki değerleri 0 ile 1 arasına sıkıştıran bir matematiksel fonksiyondur.

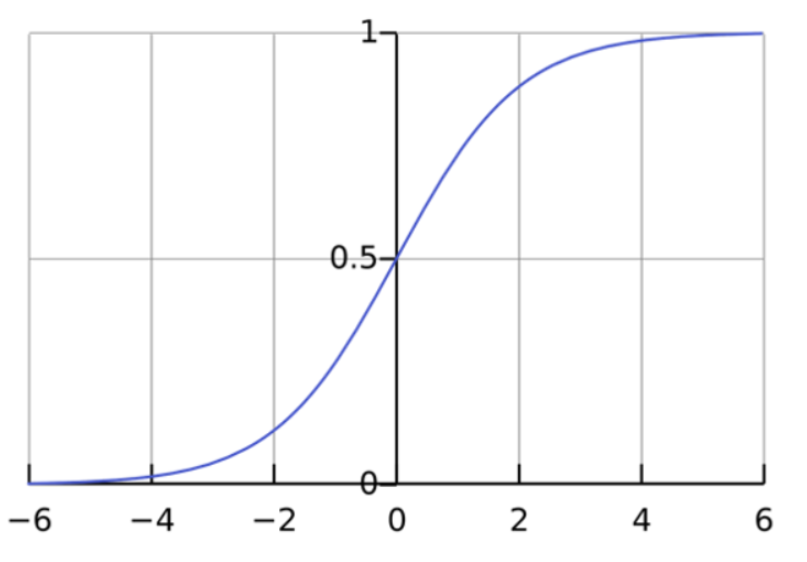
Sigmoid fonksiyonu genellikle şu şekilde ifade edilir:

yazı tipi, beyaz, simge, sembol, metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Z doğrusal regresyon denkleminin sonucu Sigmoid fonksiyona yazıldığında, Sigmoid fonksiyonu 0 ve 1 arasında bir olasılık değeri üretir ve bu olasılık değerine göre 1 veya 0 olarak bir sınıf ataması yapılır. Sınıflandırma probleminde her zaman 1 sınıfına ait olma olasılığı hesaplanır.

Sigmoid Fonksiyonunun Grafiği;

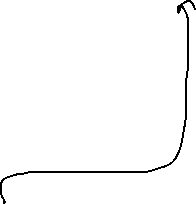
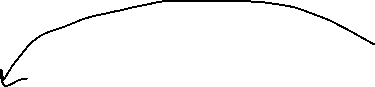


Burada degişken X değerlerine karşılık P(Y=1) olasılığının örnek bir grafik üzerinde incelenmiştir.

**Sigmoid fonksiyonu ve bağımlı değişkenin birinci sınıfının gerçekleşme olasılığı;**

çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin içeren bir resim

Açıklama otomatik olarak oluşturuldu



Klasik regresyon yöntemlerinde bağımlı değişkenin 0 ve 1'lerden oluştuğu durumlarda, elde edilen değerlerin 0 ve 1 arasında olması her iş probleminde mümkün olmamaktadır. **Sigmoid fonksiyonu tam olarak bu aşamada devreye girerek, z doğrusal formunun değerlerini 0 ve 1 arası değerlere eşlemektedir.**

Özetlexᵢ bağımsız değişkenlerinin değerlerine göre Sigmoid fonksiyonu ile yapılan eşleme sonucunda oluşan değerler, bağımlı değişkenin 1. sınıfının gerçekleşmesi olasılığına denk gelir. Örneğin bağımlı değişkenin hayatta kalma durumunu belirten bir değişken olduğu durumda hayatta kalma durumu 1, kalamama durumu ise 0 olarak ifade edilir. Bu işlemin ardından genellikle bir olasılık eşik değeri belirlenir ve bu eşik değerden yüksek değerler 1, düşük değerler ise 0 olarak kabul edilir.,

Lojistik Regresyondaki bu dönüşüm işlemi, y**apay ve derin öğrenme** sınıflandırma problemlerinin temelini oluşturmaktadır.

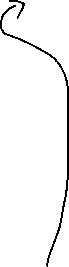
**Parametrelerin Tahmin Edilmesi**

**Logistic Regression Için Gradient Descent**

*Gradient descent* (gradyan azalma), rastgele alınan değişkenlerle başlayarak global minimum değerine ulaşmayı amaçlayan en popüler optimizasyon algoritmalarından biridir. Lojistik regresyon kapsamında, *gradient descent*ile iterasyonlarda kullanılmak üzere *cost*fonksiyonu üzerinden elde edilen J(θ) değerlerine ulaşılır

**çizim, diyagram, metin, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu**



**metin, yazı tipi, beyaz, el yazısı içeren bir resim

Açıklama otomatik olarak oluşturuldu**



Yukarıdaki formülde hₒ(xᶦ) tahmin edilen değerleri, yᶦ ise gerçek değerleri temsil etmektedir. Bu problem özelinde kayıp fonksiyonu, *log loss*ifadesidir. Her bir gözlem birimi tek tek gezilerek J(θ) hesaplanır. *Log loss* fonksiyonu içerisindeki hₒ(xᶦ) ifadesi ise bağımlı değişkenin 1 sınıfının tahmin edilen gerçekleşmesi olasılığıdır. Dolayısıyla -yᶦlog(hₒ(xᶦ)) ifadesi içerisinde değişkenin 1 sınıfının tahmin edilen gerçekleşme olasılığı varken, (1-yᶦ)log(1-hₒ(xᶦ)) ifadesi içerisindeki (1-hₒ(xᶦ)) kısmında 1 sınıfının tahmin edilen gerçekleşmemesi (yani 0 sınıfının tahmin edilen gerçekleşmesi) olasılığı gözlenir.

Ayrıca fonksiyonun yapısı gereği yᶦ gerçek değerleri 0 veya 1 olacağından toplam sembolünün sağındaki ifadelerde ya -yᶦlog(hₒ(xᶦ)) ifadesi ya da (1-yᶦ)log(1-hₒ(xᶦ)) ifadesi sıfır olur.

Bu ifade genel olarak çapraz entropi fonksiyon (*cross-entropy loss function*) olarak bilinir. Entropi ne kadar yüksek olursa çeşitlilik de o kadar fazla olacaktır. ***Dolayısıyla gerçek değer ve tahmin edilen değer açısından entropinin düşük, yani çeşitliliğin az olması istenir.* *Bir sınıfa ait olmanın, tahmin edilen gerçekleşme olasılığı ne kadar yüksekse entropi o kadar düşük olacaktır.***

Sonuç olarak gerçek değerler ile gerçek değerlerin gerçekleşmesi olasılığı ifadeleri birbirine ne kadar yakın ise ***loss*değeri**(kayıp değeri) o kadar küçük olacaktır.

metin, el yazısı, yazı tipi, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bir iterasyonda belirli bir ağırlık setinin kullanılması sonucunda, bir hata değeri elde edilir. Ardından *cost*fonksiyonunun ilgili ağırlıklara göre kısmi türevi alınır. Bu işlem sonucunda elde edilen değerin yönüne göre bu yönün negatifine gidilerek parametrenin değeri güncellenir ve hata değerlerinin düşmesi gözlemlenir.

**Sınıflandırma problemlerinde başarı değerlendirme**

Aşağıdaki örnek müşteri terk veri seti incelendiğinde, müşterilerin terk etme durumu bağımlı değişken olan *Churn* altında 1 ve terk etmeme durumu 0 ile ifade edilmiştir. Bu kapsamda kurulmak istenen model, özellikleri verilen müşterilerin terk etme/etmeme durumunu tahmin etmek üzerinedir. Bu modelde başarı, ilgili gözlem birimi için tahmin edilen ve gerçek değerin eşleşmesidir.

metin, sayı, numara, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bir sınıflandırma modelinin başarısının ölçülebilmesi için çeşitli başarı metrikleri bulunmaktadır. *Confusion matrix* (karmaşıklık matrisi,  *accuracy*(doğruluk), *precision*, *recall*, F1 skoru ve *ROC curve* (ROC eğrisi) bunlardan bazılarıdır.

**Confusion matrix (karmaşıklık matrisi)**

Model kurulduktan sonra ortaya çıkan tahmin sonuçlarında tahmin edilen değerlerin gerçek değerler ile eşleştiği gözlem birimleri:

* ***1 sınıfı*** için ***True Positive*** (TP)
* ***0 sınıfı*** için ***True Negative*** (TN)

Tahmin edilen değerin ve gerçek değerin eşleşmediği gözlem birimleri:

* ***1 sınıfı*** için ***False Positive*** (FP) (İstatistikte 1. tip hata olarak betimlenir)
* ***0 sınıfı*** için ***False Negative*** (FN) (İstatistikte 2. tip hata olarak betimlenir)

olarak ifade edilir. Yani **doğru yapılan işlemlerde True, yanlış yapılan işlemlerde ise False** ifadesi kullanılır.

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Elde edilen gerçek değerler ile tahmin değerleri karşılaştırmasına göre model başarısı aşağıdaki metrikler ile sayısal olarak ifade edilebilir:

***Accuracy*(doğruluk):**Doğru sınıflandırma oranıdır. Hesaplamak için (TP+TN)/(TP+TN+FP+FN) formülü kullanılır.

İncelenen sınıflandırma problemi dengeli sınıf dağılımına sahipse (her sınıfta ait benzer sayıda veri bulunuyorsa) *accuracy* metriği kullanılabilir. Sınıf dağılımının dengesiz olması durumunda, sınıf sayısı düşük olan sınıfı yakalama problemi ile karşılaşılabilmektedir. Sınıf dağılımlarının dengesiz olması durumunda ***accuracy*** yerine ***precision*** ve ***recall*** metrikleri incelenir.

***Precision:***Pozitif sınıf (1) tahminlerinin başarı oranıdır, yani tahmin edilen pozitif sınıfların (1 olarak tahmin edilen sınıfların) **gerçekte ne kadarının pozitif olduğunu** gösterir. 1. tip hata ile ilgilenilir. Tahminlerin başarısına odaklanır. Hesaplamak için TP/(TP+FP) formülü kullanılır.

***Recall:***Pozitif sınıfın (1) doğru tahmin edilme oranıdır.Tahmin edilen pozitif sınıfların **ne kadarının doğru tahmin edildiğini** gösterir. 2. tip hata ile ilgilenir. Gözden kaçırmaların maliyeti hakkında bilgi verdiğinden önemli bir ölçüdür. Gerçekleri yakalama başarısına odaklanır. Hesaplamak için TP/(TP+FN) formülü kullanılır.

***F1 Score:*** Precision ve recall değerlerinin harmonik ortalamasıdır. Hem Precision hem de Recall değerinin etkisini tutmaktadır. Hesaplamak için 2\*(Precision\*Recall)/(Precision+Recall) formülü kullanılır.

**ROC Curve**

metin, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Herhangi bir model kurulmadan, ortalama bir değer tahmin edilmesi durumunda elde edilecek başarının *random classifier doğrusu* gibi olması beklenir. *Random classifier* doğrusunun üstünde kalan ROC eğrileri, *perfect classifier*’a doğru içbükeyleştikçe başarı oranı artmaktadır. Bu eğriler olası eşik değerlere (*threshold*) göre oluşan karmaşıklık matrisi üzerinden, her bir olası eşik değerine karşılık gelen *True Positive* ve *False Positive* değerleri hesaplanarak oluşturulur.

ROC eğrisi ile *random classifier* arasındaki alana ***Area Under Curve (AUC)*** ismi verilmektedir. Bu metrik ROC eğrisinin sayısal bir şekilde ifade edilişidir. AUC, tüm olası sınıflandırma eşikleri için toplu bir performans ölçüsüdür. AUC değerinin yüksek olması sınıflandırmanın iyi yapıldığı anlamına gelir.

**KNN Algoritması nedir?**

K-en yakın komşu algoritması (veya KNN), basitliği nedeniyle en çok kullanılan öğrenme algoritmalarından biridir. KNN veya K-en yakın komşu Algoritması, birbirine yaklaşan her veri noktasının aynı sınıfa girmesi prensibiyle çalışan denetimli bir öğrenme algoritmasıdır. Buradaki temel varsayım, birbirine yakın olan şeylerin birbirine benzediğidir. Çoğunlukla KNN Algoritması, yorumlama kolaylığı ve düşük hesaplama süresi nedeniyle kullanılmaktadır.

KNN (K-Nearest Neighbors) Algoritması iki temel değer üzerinden tahmin yapar;

* **Distance (Uzaklık):**Tahmin edilecek noktanın diğer noktalara uzaklığı hesaplanır. Bunun için Minkowski,Öklid ve Manhattan uzaklık hesaplama fonksiyonları kullanılır.

metin, yazı tipi, diyagram, el yazısı içeren bir resim

Açıklama otomatik olarak oluşturuldu

* **K (komuşuluk sayısı):** En yakın kaç komşu üzerinden hesaplama yapılacağını söyleriz. K değeri sonucu direkt etkileyecektir. K 1 olursa overfit etme olasılığı çok yüksek olacaktır. Çok büyük olursa da çok genel sonuçlar verecektir. Bu sebeple optimum K değerini tahmin etmek problemin asıl konusu olarak karşımızda durmaktadır. K değerinin önemini aşağıdaki grafik çok güzel bir şekilde göstermektedir. Eğer K=3 ( düz çizginin olduğu yer) seçersek sınıflandırma algoritması ? işareti ile gösterilen noktayı, kırmızı üçgen sınıfı olarak tanımlayacaktır. Fakat K=5 (kesikli çizginin olduğu alan) seçersek sınıflandırma algoritması, aynı noktayı mavi kare sınıfı olarak tanımlayacaktır.

ekran görüntüsü, renklilik içeren bir resim

Açıklama otomatik olarak oluşturuldu

**K-en yakın komşu algoritması nasıl çalışır?**

K-en yakın komşu (k*-nearest neighbors,* KNN) algoritması, gözlemlerin birbirlerine olan benzerlikleri üzerinden tahminlerin yapıldığı gözetimli makine öğrenmesi modellerinde regresyon ve sınıflandırma problemlerinde kullanılan bir algoritmadır.

Bu algoritma kapsamında tahminde bulunmak istediğimiz gözlem birimine en yakın K adet farklı gözlem birimi tespit edilir ve bu K adet gözlem biriminin bağımlı değişkenleri üzerinden ilgili gözlem için tahminde bulunulur. Konuyu regresyon ve snıflandırma problemleri kapsamında birer örnekle detaylandıralım.

**Regresyon problemlerinde KNN**

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bağımlı ve bağımsız değişkenin bir arada olduğu yukarıdaki örnek veri setinde, Y bağımlı değişkeni sayısal bir değişkendir. Dolayısıyla bir regresyon problemi ile karşı karşıyayız.

metin, diyagram, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Herhangi bir X1 ve X2 değerine sahip gözlem birimi için Y bağımlı değişken tahmini yapılmak istendiğinde, öklid ya da benzeri bir uzaklık hesabı ile en yakın K adet gözlem birimi hesaplanarak bu gözlem birimlerinin bağımlı değişkenlerinin (Y değerlerinin) ortalaması alınır.

**Sınıflandırma problemlerinde KNN**

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bağımlı ve bağımlı değişkenin bir arada olduğu yukarıdaki örnek veri setinde, Y bağımlı değişkeni sayısal görünmesine karşın *binary encode*edilmiş bir değişkendir. Dolayısıyla bir sınıflandırma problemi ile karşı karşıyayız.

metin, diyagram, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Regresyon problemindekine benzer şekilde herhangi bir X1 ve X2 değerine sahip gözlem için Y bağımlı değişken tahmini yapılmak istendiğinde önce uzaklık hesabı yapılarak en yakın K gözlem birimi belirlenir. Ardından regresyon probleminden farklı olarak en yakın K adet gözlemin Y değerlerinin en sık gözlenen frekansı, tahmin edilen sınıf olarak belirlenir.

### Bir Örneği İnceleyelim:

Verilen görüntüde, iki veri sınıfımız var. Kareleri temsil eden A Sınıfı ve üçgenleri temsil eden B Sınıfı.

Problem, KNN algoritması kullanılarak iki sınıftan birine yeni bir girdi veri noktası atamaktır.

ilk adım, En Yakın Komşuların sayısını temsil eden 'K' değerini tanımlamaktır.

ekran görüntüsü, diyagram, çizgi, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

“k” değeri 6 ise o veri noktasına en yakın 6 Komşuyu, “k” değeri 5 ise o veri noktasına en yakın 5 Komşuyu arar.

'K' = 4'ü ele alalım, bu, algoritmanın veri noktasına en yakın olan dört komşuyu dikkate alacağı anlamına gelir.

Şimdi 'K' = 4'te bir üçgen ve iki kare en yakın komşular olarak görülebilir. Böylece, 'K' = 4'e dayalı yeni veri noktası A Sınıfına atanacaktır.

ekran görüntüsü, diyagram, daire, grafik içeren bir resim

Açıklama otomatik olarak oluşturuldu

## **KNN nerede kullanılır?**

KNN, hem sınıflandırma hem de regresyon tahmin problemlerinde kullanılır. Ancak endüstriyel amaçlarla uygulandığında, bir tekniğin kullanılabilirliği belirlenirken değerlendirilen tüm parametreleri karşıladığı için çoğunlukla sınıflandırmada kullanılır.

1. Tahmin Gücü
2. Hesaplama Süresi
3. Çıktıyı Yorumlama Kolaylığı

## Günlük problemlerde nasıl kullanılır?

Basitliğine rağmen KNN, diğer güçlü sınıflandırıcılardan çok daha iyi çalışır ve ekonomik tahmin ve veri sıkıştırma, Video Tanıma, Görüntü Tanıma, El Yazısı Algılama ve Konuşma Tanıma gibi yerlerde kullanılır.

## **KNN Algoritmasının bazı önemli kullanımları**

KNN Algoritması, bankacılık sisteminde, bir kişinin temerrüde düşen bir kişiyle benzer özelliklere sahip olup olmadığını tahmin ederek, bir kişinin kredi onayına uygun olup olmadığını tahmin etmek için kullanılır. KNN, benzer özelliklere sahip kişilerle karşılaştırarak kişilerin kredi notlarının hesaplanmasına da yardımcı olur.

## **KNN Kullanan Şirketler**

Amazon veya Netflix gibi E-ticaret ve eğlence şirketlerinin çoğu, satın alınacak ürünleri veya izlenecek filmleri/şovları önerirken KNN'yi kullanır.

Bu tavsiyeleri nasıl yapıyorlar? Peki, bu şirketler web sitelerinde daha önce satın aldığınız ürünler veya izlediğiniz filmler gibi kullanıcı davranışları hakkında veri toplar ve KNN uygular.

Şirketler, mevcut müşteri verilerinizi girecek ve bunları benzer ürünleri satın almış veya benzer filmler izlemiş olan diğer müşterilerle karşılaştıracaktır.

Algoritmanın bu veri noktasını nasıl sınıflandırdığına bağlı olarak ürünler ve filmler size önerilecektir.

### KNN'nin Avantajları

* Hızlı hesaplama
* Basit algoritma – yorumlamak için
* Çok yönlü – sınıflandırma ve regresyon için kullanışlıdır
* Yüksek doğruluk
* Veriler hakkında varsayım yok - ek varsayımlar yapmaya veya bir model oluşturmaya gerek yok.

### KNN'nin Dezavantajları

Doğruluk, verilerin kalitesine bağlıdır

Tahmin, büyük verilerle yavaşlar

Büyük veri kümeleri için uygun değil

Tüm eğitim verilerini saklama ihtiyacı, bu nedenle yüksek bellek gerektirir

Tüm eğitimi sakladığı için hesaplama açısından pahalı olabilir.

**Kaynaklar**

* Miuul, [Makine Öğrenmesi](https://miuul.com/makine-ogrenmesi)
* Veri Bilimi Okulu, [Lojistik Regresyon](https://www.veribilimiokulu.com/category/makine-ogrenmesi/siniflandirma-classification/lojistik-regresyon/)
* TowardsDataScience, [Logistic Regression — Detailed Overview](https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc)
* ScienceDirect, [Logistic Regression](https://www.sciencedirect.com/topics/computer-science/logistic-regression" \l ":~:text=Logistic%20regression%20is%20a%20process,/no,%20and%20so%20on)
* AWS, https://aws.amazon.com/tr/what-is/logistic-regression/
* MachineLearning,https://www.coursera.org/learn/machine-learning?source=post\_page-----4e1d2951c5c1--------------------------------
* MakineÖğrenimiNotlarım6-LogisticRegression, <https://alitunacanonar.medium.com/makine-öğrenimi-notlarım-6-logistic-regression-472092ff9ee6>
* Alpar, R. (2011). Uygulamalı çok değişkenli istatistiksel yöntemler. Ankara: Detay Yayıncılık.
* Atasoy, D. (2001). Lojistik regresyon analizinin incelenmesi ve bir uygulaması. (Yüksek lisans tezi, Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Sivas).
* Barak, A., Karahan, S. ve Saraçbaşı, O. (2005, Mayıs). Ordinal lojistik regresyon modelleri. 4. İstatistik Kongresi'nde sunulan bildiri, Belek-Antalya.
* Cook D., Dixon P., Duckworth W. M., Kaiser M. S., Koehler K., Meeker W. Q. & Stephenson W. R. (2001).Binary response and logistic regression analysis, in Part of the Iowa State University NSF/ILI project. Beyond Traditional Statistical Methods www.public.iastate.edu/ ~stat415/stephenson/stat415\_chapter3.pdf. adresinden edinilmiştir.
* Cox, R.,(1983). Some remarks on overdispersion. Biometrika,70, 269-274. http://www.jstor.org/stable/2335966 adresinden edinilmiştir.
* Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). Sosyal bilimler için çok değişkenli istatistik- spss ve lisrel uygulamaları. Ankara: Pegem Akademi Yayınları.
* Çokluk. Ö. (2010). Lojistik regresyon analizi: kavram ve uygulama. Kuram ve Uygulamada Eğitim Bilimleri,10(3),1357-1407. http://www.edam.com.tr/kuyeb/pdf/tr/3e2b1f84ce847e4fef09b68db9b1a420kFULL.pdf adresinden edinilmiştir.
* Dean, C. B. (1992). Testing for overdispersion in poisson and binomial regression models. Journal of American Statistical Association, 87(418), 451-457. http://www.jstor.org/stable/2290276 adresinden edinilmiştir.
* Elhan, A. H. (1997). Lojistik regresyon analizinin incelenmesi ve tıpta bir uygulaması (Yüksek Lisans Tezi Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara).
* Field, A. (2009). Discovering Statistics using SPSS. London: Sage.
* Garson, D. G. (2012). Ordinal Regression. Asheboro: Statistical Associates Publishing.
* Gürcan, M. (1998). Lojistik regresyon analizi ve bir uygulama. (Yüksek lisans tezi, Ondokuz Mayıs Üniversitesi, Fen Bilimleri Enstitüsü, Samsun)
* Hair, J. F, Black, W. C , Babin, B., Anderson, R. E., Tatham, R. L. (2006). Multivariate data analysis. Upper Saddle River, NJ: Prentice-Hall.
* Herrington, R. ve Starkweather, J. (2012). Logistic Regression (Binary). Research and Statistical Support. 5 http://www.unt.edu/rss/class/Jon/SPSS\_SC/Module9/M9\_LogReg/SPSS\_M9\_LogReg.htm adresinden edinilmiştir.
* Hosmer D.W. ve Lemeshow, S. (2000). Applied lojistic regression. John WileySons, Inc.:Canada. http://books.google.com.tr/ adresinden edinilmiştir.
* Kalaycı, Ç. (Ed.). (2009). SPSS uygulamalı çok değişkenli istatistik teknikleri. Ankara: Asil Yayın Dağıtım Ltd.Şti.
* Kılıç, S. (2000). Lojistik regresyon analizi ve pazarlama araştırmalarında bir uygulama. (Yüksek lisans tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul)
* Kleinbaum, D.G. ve Klein, M. (2010). Logistic regression, a self–learning text. Springer: New York.
* Mertler, C. A. ve Vannatta, R. A. (2005). Advanced and multivariate statistical methods: practical application and interpretation. Glendale, CA: Pyrczak Publishing.
* Sümbüloğlu, K. ve Akdağ, B. (2007). Regresyon yöntemleri ve korelasyon analizi. Ankara: Hatipoğlu Yayınları.
* Şentürk, E. (2011). Mutluluk düzeyinin sosyodemografik özelliklerle lojistik regresyon analizi aracılığıyla incelenmesi ve türkiye için bir uygulama . (Yüksek Lisans Tezi, Marmara Üniversitesi, Ekonometri Anabilim Dalı, İstatistik Bilim Dalı, İstanbul). http://tez2.yok.gov.tr/ adresinden edinilmiştir.
* Tabachnick, B. G. & Fidell, L. S. (1996). Using multivariate statistics. New York, USA: Harper Collins College Publishers.