Makine Öğrenmesi Temel Kavramlar ve Tanımlar

Bilgi toplama gibi öğrenme, kesin olarak tanımlanması zor olan çok çeşitli süreçleri kapsar. Öğrenmenin sözlük anlamı, bilgi edinmek, çalışmak, deneyimlemek suretiyle anlama, beceri kazanma gibi ifadelerin ve deneyimle davranışsal eğilimlerin değiştirilmesini içerir. Makine öğrenmesi alanında araştırmacılar tarafından keşfedilen kavram ve tekniklerin biyolojik öğrenmenin bazı yönlerini aydınlatabileceği de muhtemel görünüyor. Öte yandan biyolojik öğrenme metotları da makinelerin öğrenmesine inanılmaz katkı vereceği öngörülmektedir.

Yapay zekanın en aktif olarak kullanıldığı alan kuşkusuz robot teknolojileridir. Yapay zekanın gelişmesi robot teknolojilerinin gelişimini de doğrudan etkiledi. Robotlarda gerçekleşen performans problemlerini kolay bir şekilde algılayabilen yapay zeka, ihtiyaç halinde sorunları giderebiliyor. Böylece robotlar kendini yenileyebiliyor.

Makine Öğrenmesi Tanımlar:

• Makine öğrenimi, deneyimle otomatik olarak öğrenmek ve gelişmek için sistem programlamayla ilgilenen bir bilgisayar bilimi dalıdır. Örneğin: Robotlar, sensörlerden topladıkları verilere göre görevi yerine getirebilecek şekilde programlanmıştır. Verilerden programları otomatik olarak öğrenir.

• Makine Öğrenmesi: veri yığınından öğrenen ve otonom davranış sergileyen algoritmaların ve matematiksel modellerin oluşturulmasıdır.

• Veri yığınından tahmin etmeye ya da karar vermeye yönelik otonom davranış paternleri geliştiren algoritmalar ve matematiksel modellerin oluşturulmasıdır. • Veri yığınından kendi kendine öğrenen matematiksel modeller ve algoritmalar ile insandan bağımsız otonom davranış geliştirilmesidir.

• Yapısal işlev olarak veri yığınından öğrenebilen ve veriler üzerinden karar vermeye yönelik tahmin yapabilen algoritmaların çalışma ve inşalarını araştıran bir sistemdir.

Makine öğrenmesi (ML) Yapay zekanın bir alt kümesi olarak görülür. Makine öğrenmesi algoritmalardan oluşur. Açık bir şekilde tahminler veya kararlar vermek için "öğrenme verileri" olarak bilinen verilere dayalı kendi kendine öğrenen bir matematiksel model oluşturulmasıdır. Makine öğrenmesi, bilgisayarları açıkça programlanmaksızın görevleri nasıl gerçekleştirebileceklerini keşfetmeyi içerir. Belirli görevleri yerine getirmeleri için verilerden öğrenen algoritmaları içerir. Atanan basit görevler için, makineyi eldeki sorunu çözmek için gereken tüm adımların nasıl yürütüleceğini söyleyen algoritmaları programlamak mümkündür; bilgisayar tarafında öğrenmeye gerek yoktur. Daha gelişmiş görevler için, bir insanın gerekli algoritmaları manuel olarak oluşturması zor olabilir. Makine Öğrenmesi uygulamasında, programcıların gereken her adımı belirtmesinden ziyade makinenin kendisinin algoritmaları geliştirmesine yardımcı olmaktadır.

Makine öğrenmesinin omurgasını oluşturan disiplinler

: • Makine öğrenmesi, tahmin yapmaya odaklandığından istatistiksel hesaplama ile yakından ilgilidir.

• Matematiksel optimizasyon çalışması, makine öğrenmesi alanına yöntemler, teori ve uygulama alanları sağlar.

• Veri madenciliği, denetimsiz öğrenim yoluyla keşifsel veri analizine odaklanan ilgili bir çalışma alanıdır.

• Uygulamalı Matematik

• Bilgisayar sistemleri ve yazılımlar Makine öğrenmesi (ML), yapay zekanın bir alt kümesi olarak görülür. Makine öğrenmesi algoritmalardan oluşur. Açık bir şekilde programlanmadan tahminler veya kararlar vermek için "öğrenme verileri" olarak bilinen örnek verilere dayalı bir matematiksel model oluşturulmasıdır. Makine öğrenmesi algoritmaları, gerekli görevleri yerine getirmek için geleneksel algoritmalar geliştirmenin zor veya mümkün olmadığı filtreleme ve katsayıları değişken fonksiyonların çok çeşitli uygulamalarında kullanılır.

Veri madenciliği, denetimsiz öğrenim yoluyla keşifsel veri analizine odaklanan ilgili bir çalışma alanıdır. Makine öğrenmesi, bilgisayarları açıkça programlanmaksızın görevleri nasıl gerçekleştirebileceklerini keşfetmeyi içerir. Belirli görevleri yerine getirmeleri için sağlanan verilerden öğrenen algoritmaları içerir. Makine öğrenmesi tam olarak tatmin edici bir algoritmanın bulunmadığı görevleri yerine getirmelerini öğrenmek için çeşitli yaklaşımlar kullanır. Çok sayıda potansiyel cevap bulunduğu durumlarda, doğru cevapların bazılarını geçerli olarak etiketlemektir.

1.Özellik Vektörü

Makine öğrenmesinde, özellik vektörleri, bir veri örneğinin temsil edilmesi için kullanılan matematiksel bir vektördür. Bu vektör, veri örneğinin çeşitli özelliklerini (features) içerir ve bu özelliklerin sayısal değerlerle ifade edildiği bir yapıdır. Özellik vektörleri, makine öğrenmesi modellerine girdi olarak verilir ve bu modeller, bu özellik vektörlerini kullanarak öğrenme ve tahminleme yapar.

Özellik vektörleri şu temel özelliklere sahiptir:

1. **Boyut:** Özellik vektörünün boyutu, içerdiği özellik sayısına eşittir. Örneğin, bir evin fiyatını tahmin etmek için kullanılan özellik vektörü, o evle ilgili bilgileri içerebilir: alanı, oda sayısı, konumu, vb.
2. **Özellikler ve Değerler:** Her bir özellik vektörü özelliği, bir öznitelik (feature) ve onunla ilişkilendirilmiş bir değer içerir. Örneğin, bir ev fiyatını tahmin ederken, "alan" özelliği ve bu özelliğin değeri olan evin metrekare cinsinden alanı.
3. **Sayısal Temsil:** Makine öğrenmesi modelleri genellikle sayılar üzerinde işlem yapar, bu nedenle özellik vektörleri genellikle sayısal değerler içerir. Ancak, bazı özellikler kategorik verileri de içerebilir ve bu durumda bu kategoriler sayısal bir temsile dönüştürülerek kullanılır.

Örnek bir özellik vektörü şu şekilde olabilir:

Ev Özellikleri=[alan, oda sayısı, konum, bina yas]

Bu vektörde her bir özellik, evin alanını temsil eden bir sayı, oda sayısını temsil eden bir sayı, konumu temsil eden bir sayı ve bina yaşını temsil eden bir sayı olabilir. Bu özellik vektörü, ev fiyatını tahmin etmek için bir makine öğrenmesi modeline girdi olarak kullanılabilir.

2.Entropi

Makine öğrenmesinde ve özellikle karar ağaçları gibi algoritmaların eğitiminde kullanılan entropi, bir veri kümesinin düzensizliğini veya belirsizliğini ölçen bir kavramdır. İstatistik ve bilgi teorisi alanlarında önemli bir rol oynayan entropi, bir veri kümesindeki bilgi düzeyini ifade eder.

Entropi, bir veri kümesindeki farklı sınıfların ya da etiketlerin ne kadar homojen ya da heterojen olduğunu ölçer. Daha yüksek entropi, daha fazla belirsizlik anlamına gelirken, daha düşük entropi daha fazla düzen veya tahmin edilebilirlik anlamına gelir.

Entropi, şu formülle ifade edilir:

$H(S) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$

Burada:

- $H(S)$, entropiyi temsil eder.
- n, farklı sınıfların veya etiketlerin sayısını temsil eder.
- $P(x_i)$, veri kümesindeki i i. sınıfa ait örneklerin oranını temsil eder.

Bu formül, her bir sınıfın olasılığını alır, bunları logaritmaya sokar, çarpar ve negatifini alarak entropiyi hesaplar. Eğer bir veri kümesinde tüm örnekler aynı sınıfa aitse (entropi düşük), entropi sıfırdır. Ancak, farklı sınıfların oranları daha dengesizse (entropi yüksek), entropi daha yüksek olacaktır.

Makine öğrenmesinde, özellikle karar ağaçları gibi algoritmalar, entropi kullanarak veri kümesini bölerek (split) bilgi kazanmayı hedefler. Böylece, her bir bölme sonucunda daha homojen alt kümeler elde edilir ve bu sayede model, daha doğru tahminler yapabilir.

3. Kayıp Veri

Eğer veride bazı örneklerin bazı özellikleri kayıpsa izlenecek iki yol vardır:

¬ Kayıp özelliklere sahip örnekler veriden tamamen çıkartılır.

¬ Kayıp verilerle çalışabilecek şekilde algoritma düzenlenir.

Eğer kayıplı örneklerin sayısı birinci seçenek uygulanamayacak kadar çoksa ikinci seçenek uygulanmalıdır. Kayıp bilgiye sahip özellik vektörü için kazanç hesaplanırken kayıplı örnekler hariç tutularak bilgi kazancı normal şekilde hesaplanır ve daha sonra F katsayısıyla çarpılır. F, kayıpsız verinin tamamına oranıdır.

$IG(X) = F.(H(X) - H(V, X))$.

Kayıp bilgiye sahip özellik vektörü içinde en sık tekrarlanan değerin kayıp bilgi yerine yazılması da önerilen yöntemlerdendir.

Eksik değerler ortaya çıktığında veri noktalarını kolayca atamayız. Sınıflandırılacak bir test noktası da eksik değişkenlere sahip olabilir. Sınıflandırma ağaçlarının, eksik değerlerini

tamamlamanın güzel bir yolu vardır. Sınıflandırma ağaçları, bir yedek ayrım bularak sorunu çözer. Başka bir değişkene dayalı başka bir ayrım bulmak için, sınıflandırma ağaçları diğer tüm değişkenleri kullanarak tüm bölünmelere bakar ve optimum bölünmeye en çok benzeyen eğitim veri noktaları bölümünü veren birini arar. En iyi bölünmenin sonucunu tahmin etmeye çalışırlar.

4. Veri Madenciliği

Veri madenciliği, denetimsiz öğrenim yoluyla keşifsel veri analizine odaklanan ilgili bir çalışma alanıdır. Verilerdeki (geçmiş) bilinmeyen özelliklerin keşfedilmesine odaklanır. Bu veri tabanlarında bilgi keşfi analizinin bir adımıdır.

Veri Madenciliği ve Makine öğrenimi arasındaki fark nedir?

Makine öğrenimi, bilgisayarlara açıkça programlanmadan öğrenme yeteneği veren algoritmaların incelenmesi, tasarımı ve geliştirilmesi ile ilgilidir. Veri madenciliği, yapılandırılmamış verilerin bilgiyi veya bilinmeyen ilginç kalıpları çıkarmaya çalıştığı süreç olarak tanımlanabilir. Veri madenciliği işlemi sırasında, öğrenme algoritmaları kullanılır.

Makine öğrenmesi ve veri madenciliği genellikle aynı yöntemleri kullanır ve önemli ölçüde örtüşür, ancak makine öğrenmesi öğrenme verilerinden öğrenilen öngörüye odaklanırken, veri madenciliği verilerde (önceden) bilinmeyen özelliklerin keşfine odaklanır (bu veritabanlarında bilgi keşfinin analiz basamağı). Veri madenciliği birçok makine öğrenme yöntemi kullanır, ancak farklı hedefleri vardır.

Makine öğrenmesi ile genellikle aynı yöntemleri kullanır ve önemli ölçüde örtüşür, ancak makine öğrenmesi öğrenme verilerinden öğrenilen bilinen özelliklere dayanarak öngörüye odaklanırken, veri madenciliği verilerde (önceden) bilinmeyen özelliklerin keşfine odaklanır. Veri tabanlarında bilgi keşfinin ilk basamağıdır.

Veri Madenciliği, büyük miktardaki veri yığını içerisinden desenlerin, ilişkilerin, önemli bilgilerin keşfedilmesi tekniğidir. Kötüye kullanım tespiti ve anormallik tespiti yöntemleri kullanılırken yapay sinir ağları, bayes ağlar ve KNN gibi sınıflandırma yöntemleri; bölünmeli, çizge tabanlı ve hiyerarşik demetleme yöntemleri, karar ağaçları ve genetik algoritmalar da mevcuttur. Ayrıca farklı yöntemlerin birleştirilmesi ile oluşturulan hibrit yöntemler de kullanılmaktadır.

Karar Destek Sistemleri

Karar Destek Sistemleri, değişik kaynaklardan toplanan bilgilerin düzenlenerek, karar modellenerek, bilgiler analiz edilerek ve değerlendirme sonuçlarını karar vericiye sunan bilgisayar tabanlı sistemlerdir. Bir karar verici için verilen kararın doğruluğu, onun yeteneklerine, deneyimine ve bilgi birikimine olduğu kadar sahip olduğu veri kümesinin yeterliliğine de bağlıdır. Kararın başarısında, verilerin doğru depolanması, doğru sınıflanması, doğru ayıklanıp işlenmesi ve doğru yorumlanması çok önemli bir rol oynar. Bu nedenle, veri madenciliği, Karar Destek Sistemleri için etkili araçlardan biridir.

5. Veri Tabanı Yönetimi

Veri tabanları: Elde edilen verilerin tutulduğu alanlardır. Bir veri tabanı sistemi, birbiri ile ilişkili verilerin birikimini içeren, veriye erişimi sağlayarak veriyi yönetmeye yardımcı olan yazılım programları kümesidir.

Makine öğrenimi projelerindeki en kritik bileşenlerden biri veri tabanı yönetim sistemidir. Bu sistemin yardımıyla çok sayıda veri sıralanabilir ve bunlardan anlamlı içgörüler elde edilebilir.

2019 Stack Overflow Survey raporuna göre Redis en çok sevilen veri tabanı, MongoDB ise en çok aranan veri tabanı.

Veri tabanları kullanım amaçlarına göre farklı isimler alır:

İlişkisel veri tabanları, her biri farklı isimler alan tablolardan oluşur. Her tabloda her bir kaydın özelliklerinin değerlerini tutan alanlar ve her kayda ait bir tekil anahtar bulunur. Bir üniversitenin veri tabanını ilişkisel veri tabanına örnek olarak verebiliriz. Zira her bir kişi için ayırt edici bir öğrenci numarası, hangi yılda kayıt yaptırdığı, hangi bölümde okuduğu gibi alanlar ile öğrenciye ait bilgiler saklanır. Buradan çeşitli sorgular ile hangi bölümde kaç öğrencinin okuduğu, geçtiğimiz yıl kaç kişinin belli bir bölüme kayıt yaptırdığı gibi soruların cevapları bulunabilir.

İşlemsel veri tabanında her bir kaydın bir işlem olduğu varsayılır. Bir marketin veri tabanını düşünecek olursak, her an bir satış yapıldığını ve her bir satışın işlemsel veri tabanında bir kayıt olarak göründüğü varsayılabilir. Bu veri tabanından, bugün, ilgilenilen üründen kaç tane satıldığı sorusunun cevabına ulaşılabilir.

Zaman serisi veri tabanı düzenli zaman aralıkları ile elde edilmiş (yıllık, haftalık, günlük) verilerin tutulduğu alanlardır. Örnek olarak borsa verilerinin, stok kontrolleri sonucu alınan verilerin, sıcaklık ölçümlerinden elde edilen verilerin depolanması gösterilebilir. Veri Ambarları: Veri Ambarları: “Veri ambarları, tüm operasyonel işlemlerin en alt düzeydeki verilerine kadar inebilen, etkili analiz yapılabilmesi için özel olarak modellenen ve tarihsel derinliği olan veri depolama sistematiği olarak tanımlanabilir.” Günlük işlemler sonucu, farklı kaynaklardan toplanan veriler, temizleme dönüştürme, birleştirme gibi işlemlerden geçirilerek, daha önce inşa edilmiş veri ambarının yapısına uygun hale getirilerek veri ambarına aktarılır.

Veri ambarları, üzerinde, verilerin yüklenmesi ve erişimi dışında herhangi bir işlem yapılmasına izin vermez. Veri ambarları belirli aralıklar ile güncellenirler. Mimari açıdan veri ambarları üç farklı şekilde olabilir. İlki, işletmelerin farklı kaynaklardan (işletmenin kendi işlemsel veri tabanı sistemleri ve dış kaynaklar dâhil olmak üzere) aldıkları tüm verilerin tutulduğu “işletme ambarları”, ikincisi veri üzerinde çalışma yaparak karar alan kişiler için belirli kurallara göre oluşturulmuş “veri pazarları” , sonuncusu ise işlemsel veri tabanlarının görsel hali olan “ görsel ambarlar” "dır.

Veri madenciliğinde kullanılan modeller:

Tahmin Edici Modeller : Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç tahmin edilmesi amaçlanmaktadır.

Tanımlayıcı Modeller : Tanımlayıcı modellerde, veri kümesinde bulunan gizli örüntülerin (olayların ve nesnelerin ortaya çıkardığı davranış değişikliklerinin desenleri) tanımlanması amaçlanmaktadır.

Veri madenciliği süreci dört aşama ile tanımlanabilir.

• İlk aşamada problem tanımlanarak veri kaynakları değerlendirilir.

• İkinci aşamada veriler kullanıma uygun hale getirilmek için hazırlanır.

• Arkasından model kurulur ve

• nihai aşamada model değerlendirilerek kullanıma hazır hale getirilir