

Makine Öğrenmesinde Temel Kavramlar

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, bilgisayarları açıkça programlanmaksızın görevleri nasıl gerçekleştirebileceklerini keşfetmeyi içerir. Belirli görevleri yerine getirmeleri için verilerden öğrenen algoritmaları içerir.

Makine Öğrenmesi uygulamasında, programcıların gereken her adımı belirtmesinden ziyade makinenin kendisinin algoritmaları geliştirmesine yardımcı olmaktadır.

Makine öğrenmesi (ML), yapay zekanın bir alt kümesi olarak görülür.

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**

Özellik (Öznitelik) Vektörleri

Makine öğrenmesinde, özellik vektörleri, bir nesnenin özellikler adı verilen sayısal veya sembolik özelliklerini matematiksel olarak, kolayca analiz edilebilir bir şekilde temsil etmek için kullanılır. Makine öğreniminin ve kalıp işlemenin birçok farklı alanı için önemlidirler. Makine öğrenimi algoritmaları, algoritmaların işleme ve istatistiksel analiz yapabilmesi için tipik olarak nesnelerin sayısal bir temsilini gerektirir. Özellik vektörleri, doğrusal regresyon gibi istatistiksel prosedürlerde kullanılan açıklayıcı değişkenlerin vektörlerinin eşdeğeri.

Aşına olabileceğiniz bir özellik vektörüne örnek RGB (kırmızı-yeşil-mavi) renk açıklamalarıdır. Bir renk, içinde ne kadar kırmızı, mavi ve yeşil olduğu ile tanımlanabilir. Bunun için bir özellik vektörü, renk = [R, G, B] olacaktır.

Özellik vektörleri, makine öğrenmesinde sınıflandırma problemlerinde, yapay sinir ağlarında ve k en yakın komşu algoritmalarında kullanılır.

Eğitim Modelleri

Veri setlerini tanımlamak için kullanılan farklı matematiksel modeller arasından seçme işlemi “Eğitim Model Seçimi” olarak bilinir. Model seçimi istatistik, makine öğrenimi ve veri madenciliği alanlarına uygulanmaktadır.

Bir modeli eğitmek, tüm ağırlıklar ve etiketli örneklerden eşik seviye değeri bulabilmek için iyi değerleri öğrenmek (belirlemek) anlamına gelir. Denetimli öğrenmede, bir makine öğrenimi algoritması birçok örneği inceleyerek ve kaybı en aza indiren bir model bulmaya çalışarak bir model oluşturur; bu sürece ampirik risk minimizasyonu denir.

Doğrusal regresyon modelleri için karesel kayıp adı verilen bir kayıp fonksiyonu kullanır.

Örnek:

Makine öğrenmesinde bir modeli eğitmenin amacı, tüm örneklerde ortalama olarak düşük

kayıplı bir eşik değeri bulmaktır. Bu eşik değeri ortalama kareler hatası ile bulunur.

$i=1...5$

örnek değerler, $y_i = 1, 2, 3, 4, 5$;

modelin örnekleme tahmin değerleri $y_{ti}=1, 2, 7, 2, 3$

$MSE=((1-1)^2+(2-2)^2+(3-7)^2+(4-2)^2+(5-3)^2)/5= (0+0+16+4+4)/5=24/5=4.8$

örnek değerler, $y_i = 1, 2, 3, 4, 5$;

modelin örnekleme tahmin değerleri $y_{ti}=1, 2, 4, 3, 5$

$MSE=((1-1)^2+(2-2)^2+(3-4)^2+(4-3)^2+(5-5)^2)/5= (0+0+1+1+0)/5=2/5=0.4$

İkinci örnek ortalama düşük kayıplı olduğu görülmektedir.

Toplu öğrenme:

Belirli bir hesaplama programını çözmek için sınıflandırıcılar veya uzmanlar gibi birden çok model stratejik olarak oluşturulur ve birleştirilir. Bu süreç toplu öğrenme olarak bilinir. Toplu öğrenme, bir modelin sınıflandırmasını, tahminini, işlev yaklaşımını geliştirmek için kullanılır.

Topluluk öğrenme, daha doğru ve birbirinden bağımsız bileşen sınıflandırıcılar oluşturduğunuzda kullanılır.

Toplu yöntemler:

Topluluk yöntemlerinin iki paradigması şunlardır:

- * Sıralı topluluk yöntemleri
- * Paralel topluluk yöntemleri

Entropi (Bilgi Kazancı)

Rassal bir değişkenin belirsizlik ölçütü olarak bilinen Entropi, bir süreç için tüm örnekler tarafından içerilen enformasyonun beklenen değeridir. Enformasyon ise rassal bir olayın gerçekleşmesine ilişkin bir bilgi ölçütüdür. Eşit olasılıklı durumlar yüksek belirsizliği temsil eder. Shannon'a göre bir sistemdeki durum değiştiğinde entropideki değişim kazanılan enformasyonu tanımlar. Buna göre maksimum belirsizlik durumundaki değişim muhtemelen maksimum enformasyonu sağlayacaktır.

Shannon'a göre entropi, iletilen bir mesajın taşıdığı enformasyonun beklenen değeridir. Shannon Entropisi (H) adıyla anılan terim, tüm x_i durumlarına ait $P(x_i)$ olasılıklarına bağlı bir değerdir.

$$H(X) = E(I(X)) = \sum_{1 \leq i \leq n} P(x_i) I(x_i) = \sum_{i=1}^n P(x_i) \log_2 \frac{1}{P(x_i)} = - \sum_{i=1}^n P_i \log_2 P_i$$

Örnek: Bir paranın havaya atılması olayı, rassal X sürecini temsil etsin. Yazı (1/2) ve tura (1/2) gelme olasılıkları eşit olduğu için X sürecinin entropisi 1 olan para atma olayı (X) gerçekleştiğinde 1 bitlik bilgi kazanılacaktır.

$$H(X) = - \sum_{i=1}^2 P_i \log_2 P_i = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$$

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.

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

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

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.

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.

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.

Veri tabanları kullanım amaçlarına göre farklı isimler alır:

- * İlişkisel veritabanları
- * İşlemsel veritabanı
- * Zaman serisi veritabanı

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 sistematığı olarak tanımlanabilir.”

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.

