VERİ BİLİMİ İLE İLGİLİ BİLİNMESİ GEREKEN TEMEL BİLGİLER

SÜHEYL ÇAVUŞOĞLU

Şubat, 2023

İçindekiler

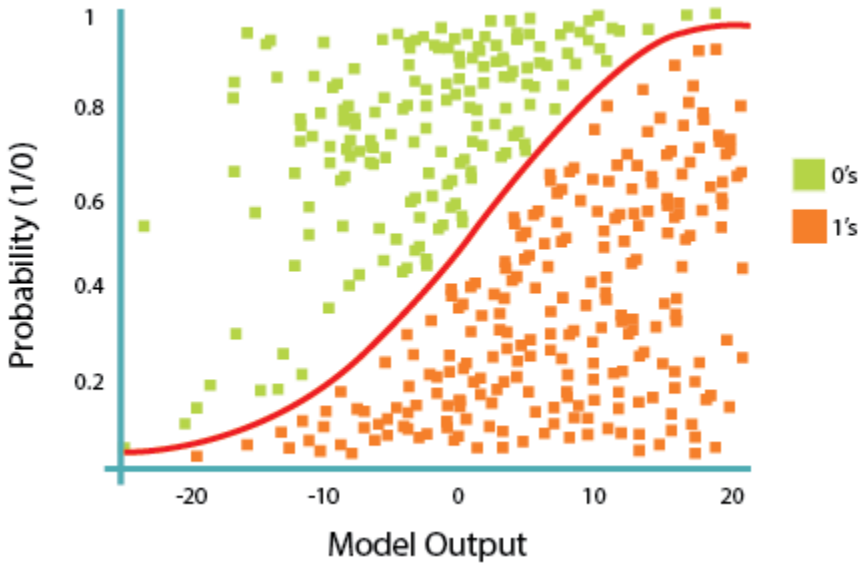
Şekiller Tablosu

1. GİRİŞ

Bu çalışmada Veri Bilimi ile ilgili bilinmesi gereken temel bilgiler yer almaktadır. Veri Bilimci pozisyonlarına başvuranlara yöneltilen, 23 farklı soru örneğine [1] açıklamalar getirilmiştir. Soruların referans alındığı adrese gidilerek, konularla ilgili İngilizce olarak hazırlanmış videolardan daha fazla bilgi edinilebilir. Bu çalışmanın amacı, İngilizce konusunda kendisinin yeterli olmadığını düşünenler için, Veri Bilimi ile ilgili bilinmesi gereken temel bilgileri, özet şeklinde, Türkçe olarak aktarmaktır. Bu konular ile ilgili daha detaylı bilgilere ulaşmak isteyenler, konu hakkında daha yetkin olan kişilerin hazırladığı kaynaklardan faydalanabilirler.

2. Logistic Regression Nedir?

Lojistik regresyon, bir sınıflandırma probleminde kullanılan bir istatistiksel yöntemdir. Amacı, bir veri kümesindeki bağımsız değişkenlerin belirli bir sonucu (yani bağımlı değişkeni) etkileyip etkilemediğini belirlemektir. Örneğin, bir müşterinin bir ürünü satın alıp almayacağı gibi bir evet/hayır sonucu veren bir sınıflandırma problemi için lojistik regresyon kullanılabilir. Lojistik regresyon, doğrusal regresyona benzer şekilde, bir dizi bağımsız değişkeni kullanarak bağımlı değişkeni tahmin etmeye çalışır. Ancak, lojistik regresyonun çıktısı, bir olasılık değeri (0 ile 1 arasında) olarak verilir. Bu olasılık, bir eşik değeri belirleyerek (genellikle 0.5), bir sınıflandırma sonucuna dönüştürülebilir.



Şekil 2.1 Lojistik Regresyon [2]

Lojistik regresyon, birçok endüstride yaygın olarak kullanılan bir yöntemdir. Örneğin, tıp alanında, bir hastanın belirli bir hastalığa yakalanma riskini tahmin etmek için lojistik regresyon

kullanılabilir. Ayrıca, pazarlama, finans ve diğer alanlarda müşteri davranışı, kredi riski ve diğer sınıflandırma problemleri için de kullanılır.

Lojistik regresyon, birçok farklı varyasyonu olan bir yöntemdir. Örneğin, çoklu lojistik regresyon, birden fazla bağımsız değişkenin kullanıldığı durumlar için kullanılır. Ayrıca, lojistik regresyon, makine öğrenimi algoritmaları ile birlikte kullanılarak, daha karmaşık sınıflandırma problemlerini çözmeye yardımcı olabilir.

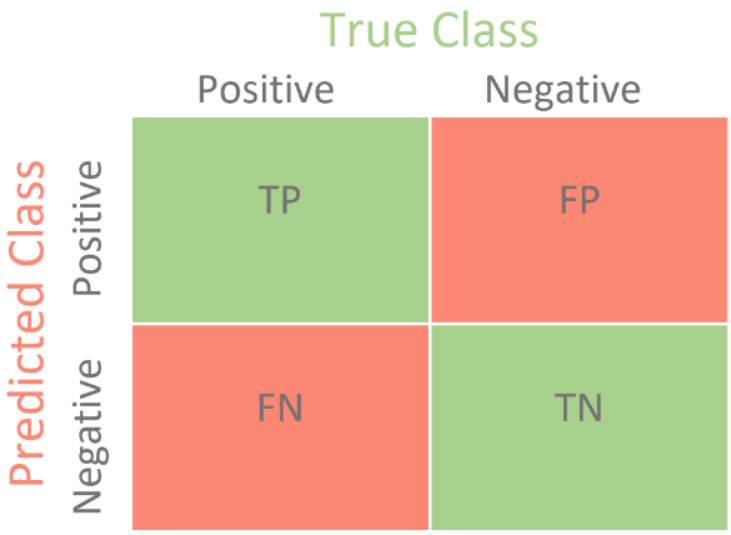
3. Confusion Matrix nedir?

Confusion Matrix (Karmaşıklık Matrisi), sınıflandırma problemlerinde modelin performansını değerlendirmek için kullanılan bir tablodur. Confusion Matrix, bir test setinde gerçek sınıf ve tahmin edilen sınıf arasındaki ilişkiyi gösterir. Tablo, dört farklı kategoriye ayrılmıştır: gerçek pozitif (TP), gerçek negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN).

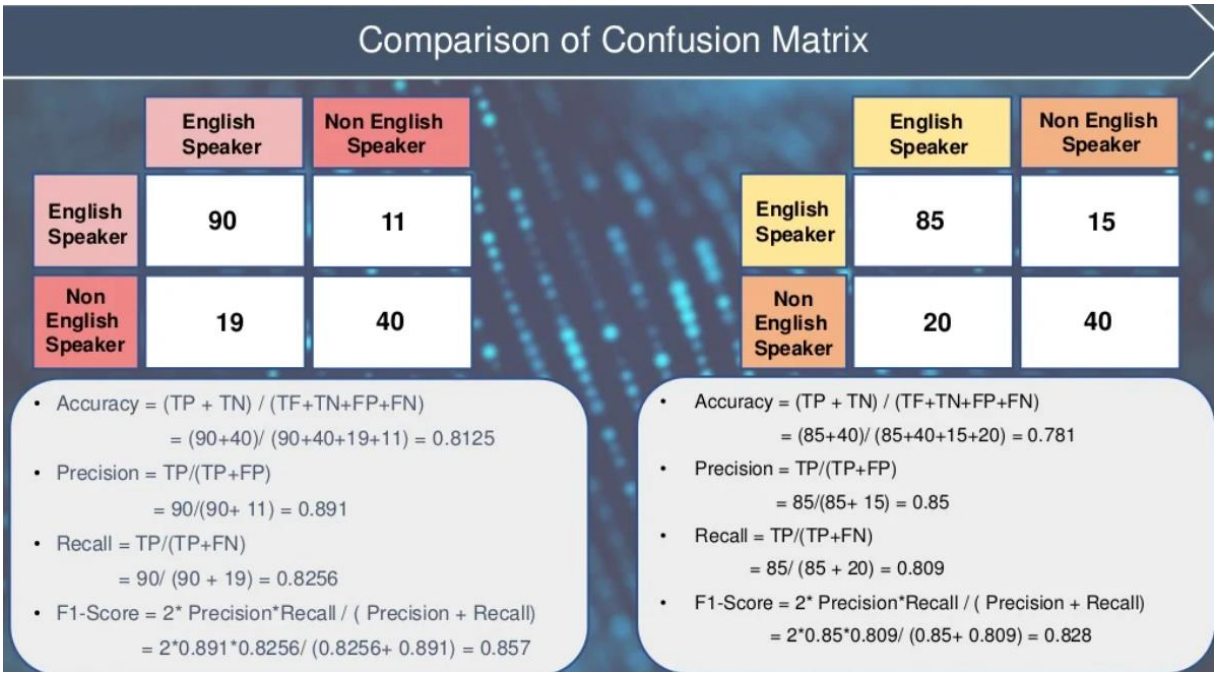
Gerçek pozitifler (TP), gerçek sınıfı pozitif olarak tahmin edilen örnek sayısını temsil eder. Gerçek negatifler (TN), gerçek sınıfı negatif olarak tahmin edilen örnek sayısını temsil eder. Yanlış pozitifler (FP), negatif sınıfı pozitif olarak tahmin edilen örnek sayısını temsil eder. Yanlış negatifler (FN), pozitif sınıfı negatif olarak tahmin edilen örnek sayısını temsil eder.

Bu matris, sınıflandırma modelinin doğruluğunu ölçmek için kullanılabilir. Örneğin, bir tıbbi test sonucunu değerlendirmek için kullanılan bir modelin performansını değerlendirmek istediğimizi varsayalım. Gerçek pozitifler, gerçek hastaların doğru bir şekilde tespit edildiği sayıdır. Gerçek negatifler, sağlıklı insanların doğru bir şekilde tespit edildiği sayıdır. Yanlış pozitifler, sağlıklı insanların yanlış bir şekilde hasta olarak tespit edildiği sayıdır. Yanlış negatifler, gerçek hastaların yanlışlıkla sağlıklı olarak tespit edildiği sayıdır.

Confusion Matrix, sınıflandırma modelinin performansını ölçmenin yanı sıra, performansı iyileştirmek için de kullanılabilir. Örneğin, modelin yanlış pozitiflerini azaltmak, hastalara yanlış bir şekilde teşhis konulması riskini azaltabilir.



Şekil 3.1 Confusion Matrix [3]



Şekil 3.2 İki farklı Confusion Matrix'in karşılaştırılması [4]

Yukarıdaki şekilde iki farklı Confusion Matrix'in karşılaştırılması yer almaktadır. Şekilde yer alan formüller ile Accuracy Score ve F-1 Score değerleri hesaplanıp karşılaştırma yapılabilir.

4. Veri Bilimi ile Geleneksel Programlama Arasındaki Farklar Nelerdir?

Veri bilimi ve geleneksel programlama arasındaki fark, problem çözme yaklaşımlarında ve programların oluşturulmasında yatmaktadır.

Geleneksel programlama, belirli bir işlevi yerine getirmek üzere tasarlanmıştır. Bir program, belirli bir girdi alır ve belirli bir çıktı üretir. Örneğin, bir program, kullanıcının girilen iki sayıyı toplamasını sağlayabilir. Bu programın amacı, doğru bir şekilde iki sayıyı toplamaktır. Programın amacı, belirli bir veri kümesi üzerinde işlem yapmak değil, belirli bir işlevi yerine getirmektir. Geleneksel programlama, önceden tanımlanmış adımlar ve kararlarla sınırlıdır.

Veri bilimi, verileri anlama, keşfetme ve analiz etme sürecidir. Veri bilimi, önceden tanımlanmış adımlara bağlı değildir. Veri bilimi, verilerin keşfedilmesine, temizlenmesine, analiz edilmesine ve sonuçların yorumlanmasına dayanır. Veriler üzerinde yapılan işlemlerden elde edilen sonuçlarla ilgili yargılar oluşturmayı ve öngörüler yapmayı amaçlar.

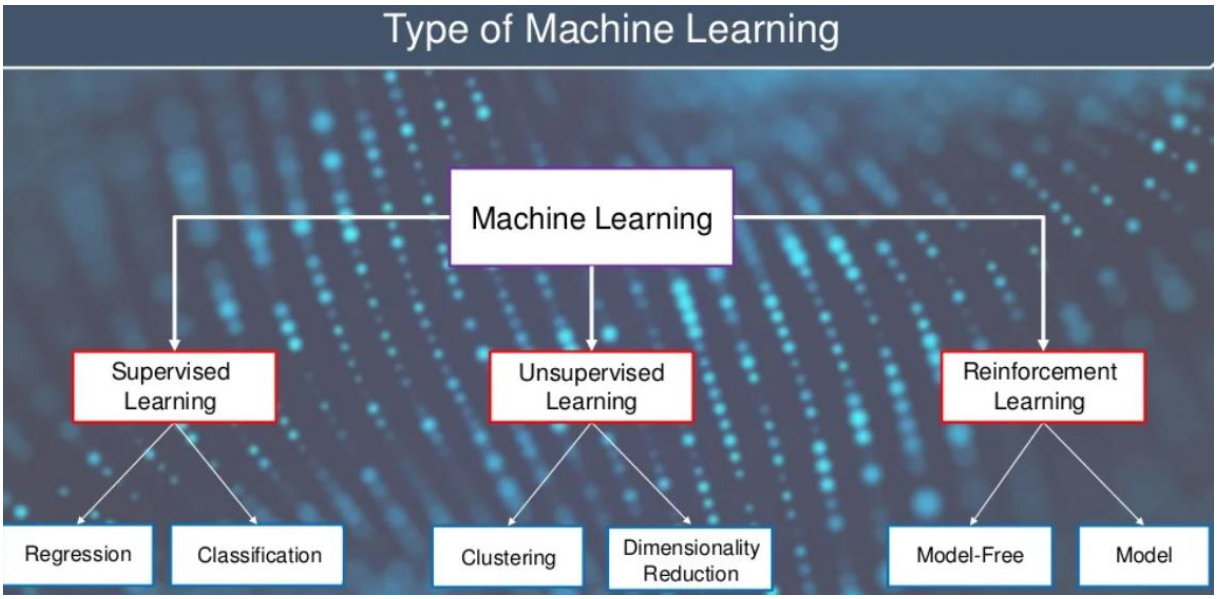
Veri bilimi ve geleneksel programlama arasındaki farkları örneklendirmek için bir örnek verecek olursak, bir programın, kullanıcının girdiği sayılara göre ortalama hesaplaması için yazıldığını düşünelim. Bu program, kullanıcının girdiği sayıları alacak, toplamını hesaplayacak ve ardından girdi sayısına bölerek ortalama değeri hesaplayacaktır. Programın adımları önceden belirlenmiştir ve program verilerin kendisiyle değil, veriler üzerinde işlem yapar.

Veri bilimi örneğinde ise, bir şirketin satış verilerini analiz etmek istediğimizi varsayalım. Bu durumda, öncelikle verilerin temizlenmesi, kategorize edilmesi ve daha sonra analiz edilmesi gerekmektedir. Analiz aşamasında, verilerin trendleri, müşteri segmentasyonu ve gelecekteki satışların tahmini gibi konular ele alınabilir. Bu işlem adımları önceden belirlenmemiştir ve veriler üzerinde işlem yapılır.

Bu örneklerden de anlaşılacağı gibi, veri bilimi ve geleneksel programlama arasındaki fark, problem çözme yaklaşımları ve programların oluşturulmasında yatmaktadır. Geleneksel programlama, belirli bir işlevi yerine getirmek üzere tasarlanmıştır ve önceden tanımlanmış adımlara bağlıdır. Veri bilimi ise, veriler üzerinde yapılan işlemlerden elde edilen sonuçlarla ilgili yargılar oluşturmayı ve öngörüler yapmayı amaçlar ve önceden belirlenmemiş adımlara bağlıdır.

Veri bilimciler, programlama becerileri kadar matematik, istatistik ve makine öğrenimi gibi alanlarda da uzmanlık sahibi olmalıdırlar. Bu sayede, verileri analiz edebilir, yorumlayabilir ve sonuçlara dayalı öngörüler yapabilirler.

5. Denetimli ve Denetimsiz Öğrenme Arasındaki Farklar Nelerdir?



Şekil 5.1 Makine Öğrenmesi Türleri [4]

Makine öğrenmesinde, iki ana kategori vardır (bir de pekiştirmeli öğrenme kategorisi vardır ancak denetimli ve denetimsiz öğrenme daha yaygın olduğu için örnekler bu ikisi üzerinden verilmiştir): denetimli öğrenme ve denetimsiz öğrenme. İkisi arasındaki temel fark, eğitim verilerinin nasıl işlendiğidir. Denetimli öğrenme, etiketli veriler kullanarak bir modelin öğrenmesini içerir. Bu etiketler, verilerin hedef çıktılarıdır ve modelin bu çıktıları tahmin etmesi beklenir. Örneğin, bir evin satış fiyatını tahmin etmek için kullanılabilecek veriler, evin özellikleri (oda sayısı, banyo sayısı, evin yaşı vb.) ve satış fiyatıdır. Bu örnekte, satış fiyatı hedef çıktısıdır ve model, evin özelliklerine dayanarak bu çıktıyı tahmin etmeye çalışır. Denetimli öğrenme, sınıflandırma ve regresyon gibi görevler için kullanılabilir. Denetimsiz öğrenme ise, etiketlenmemiş verileri kullanarak modelin öğrenmesini içerir. Bu verilerde, herhangi bir hedef çıktısı yoktur ve model, verilerin özelliklerini belirleyerek veriler arasındaki ilişkileri keşfetmeye çalışır. Örneğin, bir perakende mağazasının müşterilerinden aldığı veriler (müşteri yaşları, alışveriş sepetleri, satın alma sıklıkları vb.) denetimsiz öğrenme için kullanılabilir. Bu örnekte, herhangi bir hedef çıktısı yoktur ve model, müşteriler arasındaki benzerlikleri ve farklılıkları bulmaya çalışır. Denetimsiz öğrenme, kümeleme ve boyut azaltma gibi görevler için kullanılabilir. Özetlemek gerekirse, denetimli öğrenme, etiketli veriler kullanarak belirli bir hedefi tahmin etmek için bir modelin öğrenmesini içerirken, denetimsiz öğrenme etiketlenmemiş verileri kullanarak veriler arasındaki ilişkileri ve yapıları keşfetmek için bir modelin öğrenmesini içerir.

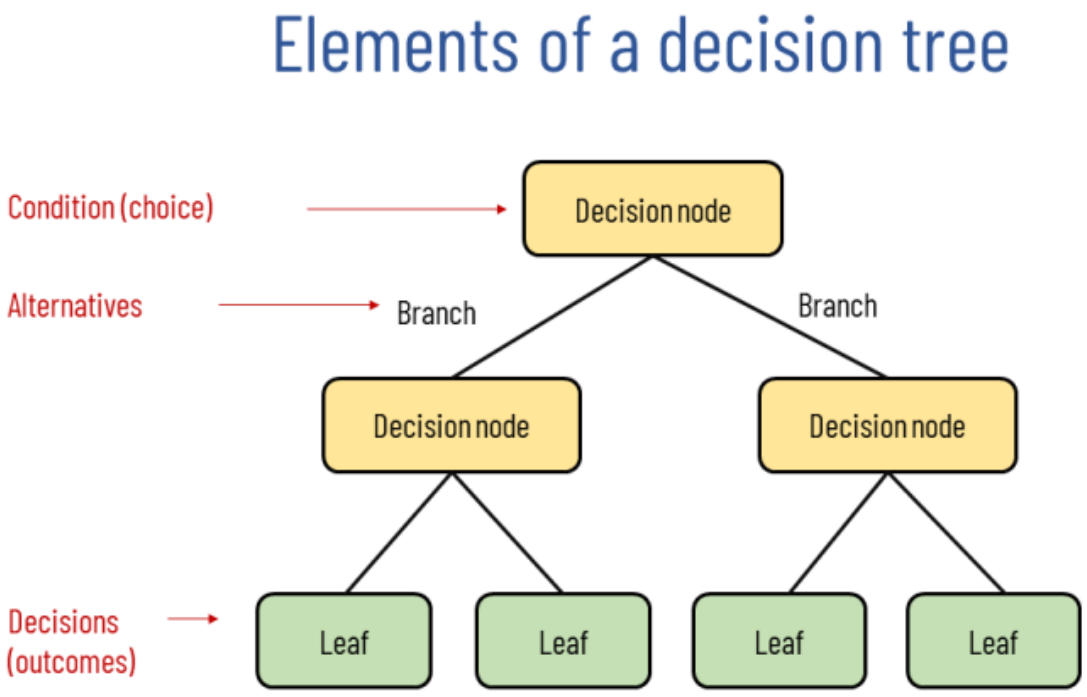
6. Karar Ağacı Nedir?

Karar ağacı, veri madenciliği ve makine öğrenmesi alanlarında sınıflandırma veya regresyon problemlerini çözmek için kullanılan bir yöntemdir. Karar ağacı, veri kümesindeki özelliklerin ve hedef çıktıların yapısını öğrenmek için bir ağaç yapısı oluşturur.

Karar ağacı, birçok basit karar ağacından oluşan bir hiyerarşik yapıdır. Her karar ağacı, bir özellik veya özellikler kümesi üzerindeki bir karar kuralını temsil eder. Karar kuralı, bir özellik değeri üzerinde bir karşılaştırma yaparak "doğru" veya "yanlış" olarak sonuç verir. Bu şekilde, veriler ağaçtaki bir düğümdeki karar kuralına göre sınıflandırılır. Ağacın yaprakları, sınıflandırılmış verilerin sonuçlarına karşılık gelir.

Karar ağacı, veri kümesindeki özelliklerin önem sıralamasını elde etmek için kullanılabilir. Ayrıca, eğitim verileri üzerinde doğru sonuçlar veren bir model oluşturmak için kullanılabilir. Yeni veriler, oluşturulan ağaçtaki karar kuralına göre sınıflandırılabilir.

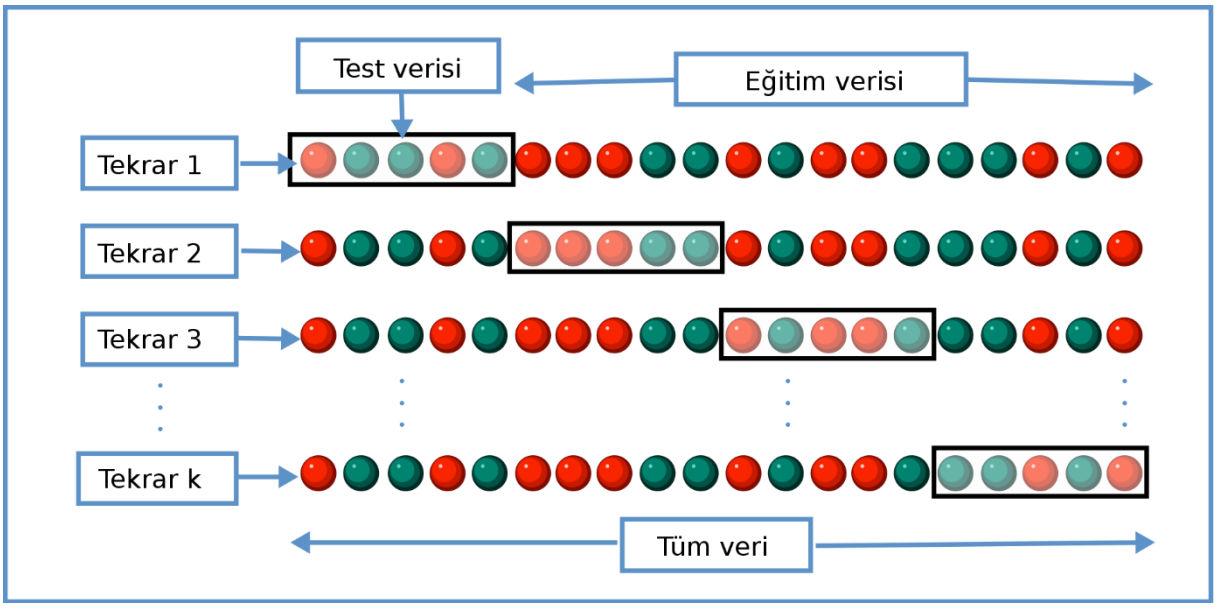
Karar ağacı, basit bir yapısı olduğu için anlaşılması ve yorumlanması kolaydır. Bununla birlikte, ağaç yapısının doğru bir şekilde oluşturulması, özellik seçiminin önemi ve ağacın aşırı uydurması gibi bazı zorluklar da vardır.



Şekil 6.1 Karar Ağacının Elemanları [5]

7. Çapraz Doğrulama Nedir?

Çapraz doğrulama, makine öğrenmesi modelinin performansını ölçmek için kullanılan bir yöntemdir. Modelin performansının ölçülmesi, modelin gerçek dünya verilerinde nasıl çalışacağı hakkında bir fikir verir. Çapraz doğrulama, veri kümesinin bir kısmını eğitim verisi ve geri kalanını test verisi olarak ayırarak başlar. Model eğitim verisi üzerinde eğitilir ve test verisindeki performansı ölçülür. Bu, modelin performansını ölçmek için bir kez yapılır. Ancak, tek seferlik bir test, modelin gerçek dünya verilerinde nasıl çalışacağı hakkında tam bir fikir vermez. Bu nedenle, çapraz doğrulama, model performansının daha kesin bir tahminini sağlamak için kullanılır. Çapraz doğrulama, veri kümesinin farklı parçalarını eğitim ve test verisi olarak kullanarak tekrarlanır. Örneğin, k-fold çapraz doğrulama, veri kümesini k eşit parçaya böler ve her seferinde bir parça test verisi olarak kullanılırken geri kalanları eğitim verisi olarak kullanılır. Bu işlem k kez tekrarlanır ve sonuçlar ortalama alınır. Böylece, modelin performansının farklı veri parçaları için nasıl değiştiği ölçülebilir. Çapraz doğrulama, modelin performansını ölçmek için kullanılan yaygın bir yöntemdir ve overfitting gibi sorunları tespit etmeye yardımcı olur. Ayrıca, farklı modellerin performansını karşılaştırmak için de kullanılabilir.



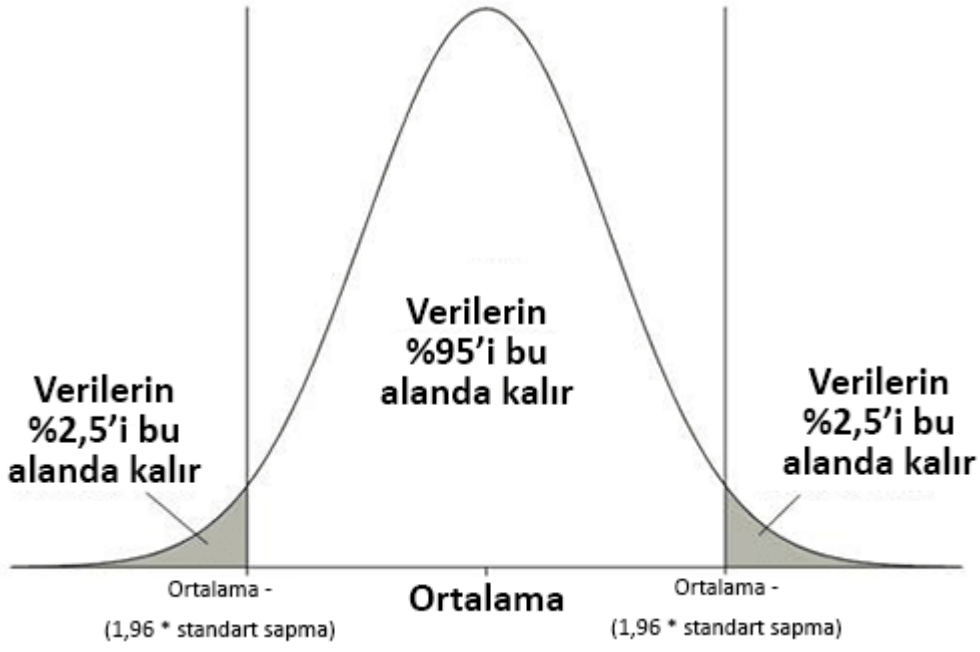
Şekil 7.1 Çapraz Doğrulama Gösterimi

8. Normal (Gauss) Dağılım Nedir?

Normal dağılım, istatistikte ve olasılık kuramında sıkça kullanılan bir olasılık dağılımıdır. Normal dağılımın diğer adı Gauss dağılımıdır. Normal dağılım, birçok farklı veri setinde

gözlenen doğal varyasyonların modellenmesi için kullanılır. Simetrik ve tek tepe noktası olan bir dağılımdır. Normal dağılımın belirleyici özellikleri şunlardır:

- Merkezi Limit Teoremi: Normal dağılım, merkezi limit teoremi tarafından açıklanan bir davranışın sonucudur. Merkezi limit teoremi, birçok bağımsız rassal değişkenin toplamının, özellikle de büyük sayıdaki bağımsız değişkenlerin toplamının, normal bir dağılım göstermesini öngörür.
- Bell Şekilli Olma: Normal dağılımın grafiği, zirve noktası ortasında olan bir çan şeklini andırır.
- Ortalama ve Standart Sapma: Normal dağılımın ortalama değeri, simetrik dağılımın merkezindeki en yüksek noktaya denk gelir. Normal dağılımın yayılması, standart sapma tarafından ölçülür ve standart sapmanın artması, dağılımın yayılmasını arttırır. Normal dağılım, birçok farklı alanda kullanılır. Örneğin, istatistiksel testlerde, varsayımların doğrulanması için normal dağılım önemlidir. Ayrıca, normal dağılım, sosyal bilimlerde, ekonomide, mühendislikte ve doğa bilimlerinde kullanılan birçok farklı ölçümün dağılımını modellemek için de kullanılır.



Şekil 8.1 Normal (Gauss) Dağılımı [6]

9. Random Forest Algoritması Nedir?

Random Forest, bir makine öğrenmesi algoritmasıdır ve aynı zamanda topluluk öğrenmesi (ensemble learning) yöntemlerinden biridir. Random Forest, birden çok karar ağacını

kullanarak veri kümesini eğitir ve sonuçları bir araya getirerek daha doğru sonuçlar üretir. Random Forest, veri setlerinin yanı sıra hem sınıflandırma hem de regresyon problemlerini çözebilir. Random Forest algoritması, her bir ağacın örnekleme verilerinde rastgele özellikler seçerek ve bu özellikler üzerinde ayrıntılı bir karar ağacı oluşturarak çalışır. Bu rastgele özellik seçimi, her bir ağacın özelliklerinin aynı ölçüde önemli olduğunu varsayar. Bu, yüksek boyutlu veri setleri için yararlıdır çünkü bu tür veri setleri genellikle birkaç önemli özellik içerir. Random Forest algoritması, ağaçlar arasında değişkenlik yaratarak overfitting (aşırı öğrenme) sorununu da azaltır. Aynı zamanda, birden çok ağacı bir araya getirerek, bireysel ağaçların doğru sonuçlar vermeyi başaramayacakları durumlarda bile doğru sonuçlar elde etmek için birbirlerini tamamlarlar. Bu nedenle, sınıflandırma ve regresyon problemleri için oldukça popüler bir makine öğrenmesi algoritmasıdır.

10. Tek Değişkenli, İki Değişkenli ve Çok Değişkenli Analizler Nedir?

Tek değişkenli analiz, yalnızca tek bir bağımsız değişkenin etkilerini inceleyen bir veri analizi yöntemidir. Örneğin, bir şirketin yıllık gelirini analiz etmek istiyorsanız, yalnızca gelir değişkenini dikkate alırsınız. Tek değişkenli analiz, veri setindeki değişkenlerin özelliklerini, dağılımlarını, merkezi eğilimleri ve dağılım genişliklerini tanımlamak için istatistiksel teknikler kullanır. İki değişkenli analiz, iki değişken arasındaki ilişkiyi inceleyen bir veri analizi yöntemidir. Örneğin, bir şirketin satışlarını artırmak istediğinde, satışların hava durumu gibi diğer faktörlere nasıl bağlı olduğunu inceleyebilirsiniz. Bu analiz türü, bir değişkenin diğerine göre nasıl değiştiğini ve bu değişimin ne kadar güçlü olduğunu incelemek için korelasyon, regresyon ve varyans analizleri gibi istatistiksel teknikler kullanır. Çok değişkenli analiz, birçok bağımsız değişkenin birbirleriyle nasıl ilişkili olduğunu inceleyen bir veri analizi yöntemidir. Örneğin, bir şirketin satışları artırılmak istendiğinde, müşteri profilleri, ürün özellikleri, pazarlama stratejileri, fiyatlandırma politikaları ve diğer faktörleri dikkate alarak analiz yapılabilir. Bu analiz türü, birçok bağımsız değişkenin etkilerini ayrı ayrı veya bir arada inceleyebilmek için çoklu regresyon, lojistik regresyon, faktör analizi, diskriminant analizi ve veri madenciliği gibi teknikleri kullanır. Bu analiz türü, verilerin daha kapsamlı bir şekilde incelenmesini ve çoklu değişkenlerin birbirleriyle nasıl ilişkili olduğunun daha iyi anlaşılmasını sağlar.

11. Eksik Veriler Nasıl Ele Alınır?

Veri analizinde eksik veriler oldukça yaygındır. Eksik veriler, veri setindeki belirli bir değişkenin bazı gözlemlerinde eksik olması nedeniyle ortaya çıkar. Bu, veri analizi ve modelleme sürecinde problem yaratabilir. Ancak, eksik verileri ele almak için bazı teknikler mevcuttur. Bu tekniklerden bazıları aşağıda verilmiştir:

- Verilerin yeniden toplanması: Eksik verilerin en iyi çözümü verilerin yeniden toplanmasıdır. Ancak, bu genellikle pratik değildir.
- Basit Değer Atama: Eksik verilerin yerine ortalamalar, medyanlar veya modlar kullanarak basit bir değer atama yapılabilir. Ancak, bu yöntem bazen yanıltıcı sonuçlar verebilir ve veri setinin dağılımını bozabilir.
- Tahmine Dayalı Değer Atama: Eksik verilerin yerine, diğer değişkenlerin kullanılmasıyla tahmin edilen değerler atanabilir. Bu yöntem genellikle daha doğru sonuçlar verir, ancak veri setindeki diğer değişkenlerin doğru bir şekilde modellenmesi gereklidir.
- Model Tabanlı Değer Atama: Eksik verilerin yerine, özellikle regresyon modelleri gibi modeller kullanarak tahmin edilen değerler atanabilir. Bu yöntem, verilerin modelleme sürecinde doğru anlanmasını gerektirir.
- Eksik Veri Silme: Veri setinden eksik gözlemler tamamen kaldırılabilir. Ancak, bu yöntem genellikle çok fazla veri kaybına neden olur ve sonuçta model performansını olumsuz etkileyebilir.

Hangi yöntemin kullanılacağı veri setinin özelliklerine ve eksik verilerin nedenine bağlıdır.

12. Boyutluluk Azaltmanın Faydası Nedir?

Boyutluluk azaltma, veri setindeki değişken sayısını azaltarak veri setini daha anlaşılır ve işlenebilir hale getiren bir yöntemdir. Boyutluluk azaltmanın birkaç faydası vardır:

- Bellek Kullanımını Azaltır: Yüksek boyutlu veri setleri, işlemek için büyük miktarda bellek gerektirebilir. Boyutluluk azaltma, bellek kullanımını azaltarak veri işleme sürecini hızlandırabilir.
- Veri Görselleştirme: Boyutluluk azaltma, veri setindeki değişken sayısını azaltarak verilerin görselleştirilmesini daha kolay hale getirir. Daha az boyutta veri setleri, örneğin 2D veya 3D grafikler oluşturmak için daha uygun hale gelir.

- Karmaşıklığı Azaltır: Yüksek boyutlu veri setleri daha karmaşık ve anlaşılması zor hale gelebilir. Boyutluluk azaltma, verilerin anlaşılmasını kolaylaştırarak veri setindeki ilişkileri daha açık hale getirir.
- Veri Analizini Kolaylaştırır: Boyutluluk azaltma, veri setlerindeki yüksek boyutluluğun neden olduğu gürültü ve gereksiz bilgiyi azaltarak veri analizini daha kolay hale getirir.
- Model Performansını Artırır: Yüksek boyutlu veri setleri, makine öğrenimi modellerinin performansını olumsuz etkileyebilir. Boyutluluk azaltma, veri setindeki gürültüyü ve gereksiz bilgiyi azaltarak model performansını artırabilir.

Sonuç olarak, boyutluluk azaltma, yüksek boyutlu veri setlerinde veri işleme ve analiz süreçlerini daha hızlı ve verimli hale getirir. Ayrıca, veri setlerindeki gereksiz bilgiyi ve gürültüyü azaltarak, daha iyi ve doğru sonuçlar elde etmek için makine öğrenimi modellerinin performansını da artırabilir.

13. Aykırı Değer ile Nasıl Başa Çıkılır?

Aykırı değerler, veri setindeki diğer gözlemlerden önemli ölçüde farklı olan nadir gözlemlerdir. Aykırı değerler, veri analizinde çeşitli sorunlara neden olabilir. Bu sorunları çözmek için birkaç yöntem vardır:

- Aykırı değerlerin nedenlerini anlamak: Aykırı değerler, yanlış veri girişleri, ölçüm hataları, doğal varyasyonlar veya gerçekten özel durumlar gibi çeşitli nedenlerden kaynaklanabilir. Öncelikle aykırı değerlerib neden oluştuklarını anlamak, uygun çözüm yöntemlerinin belirlenmesine yardımcı olabilir.
- Aykırı değerleri kaldırmak: Aykırı değerleri veri setinden çıkarmak, analiz sonuçlarını bozabilir. Ancak, bazı durumlarda aykırı değerleri kaldırmak, analizlerin doğruluğunu artırabilir. Örneğin, veri setinde birkaç aykırı değer varsa ve bu değerler veri setindeki diğer gözlemlerden önemli ölçüde farklıysa, bu aykırı değerlerin çıkarılması, model performansını iyileştirebilir.
- Aykırı değerleri değiştirmek: Aykırı değerleri değiştirmek, veri setinin doğal dağılımını korumak için daha iyi bir seçenek olabilir. Bu işlem, aykırı değerleri, veri setindeki diğer gözlemlere daha yakın bir değere değiştirmeyi içerir. Bu değişiklikler, veri setinin doğal varyasyonlarını koruyarak, analizlerin doğruluğunu artırabilir.
- Aykırı değerleri bölümlere ayırmak: Veri setindeki aykırı değerler, veri setindeki diğer gözlemlerden önemli ölçüde farklıysa ve bu değerler gerçekten özel durumları yansıtıyorsa, aykırı değerleri ayrı bir gruba koymak en iyi seçenek olabilir. Bu, aykırı

değerleri işleme almadan önce diğer gözlemlerden ayrı bir şekilde analiz etmenizi sağlar.

Aykırı değerler, veri analizi sırasında önemli sorunlar yaratabilir, ancak yukarıda belirtilen yöntemlerle bu sorunların üstesinden gelinebilir. Önemli olan, aykırı değerlerin neden oluştuğunu anlamak ve uygun çözüm yöntemlerini kullanarak veri setinin doğal varyasyonlarını korumaktır.

14. Toplu Öğrenme Nedir?

Toplu öğrenme (ensemble learning), birden fazla makine öğrenimi modelini bir araya getirerek daha iyi sonuçlar elde etmeyi amaçlayan bir tekniktir. Genellikle tek bir modelin performansı yetersiz kalıyorsa, birçok farklı modelin bir araya getirilerek performansının artırılması hedeflenir.

Toplu öğrenme teknikleri, veri setinin farklı alt kümeleri veya farklı özellikleri kullanarak oluşturulan birden fazla modelin bir araya getirilmesiyle gerçekleştirilir. Bu tekniklerin temelinde, her modelin farklı bir şekilde öğrenme yapması ve farklı hatalar yapması yatar. Bu hataların bir araya getirilerek, daha doğru bir sonuç elde edilmesi hedeflenir.

Örnek olarak, Random Forest, Gradient Boosting ve AdaBoost gibi teknikler, toplu öğrenme yöntemlerine örnek olarak gösterilebilir. Bu teknikler, farklı alt kümeler veya farklı ağırlıklarla bir araya getirilen birden fazla karar ağacı modelinin kullanılması ile gerçekleştirilir.

15. Makine Öğrenimi ile Derin Öğrenme Arasındaki Farklar Nelerdir?

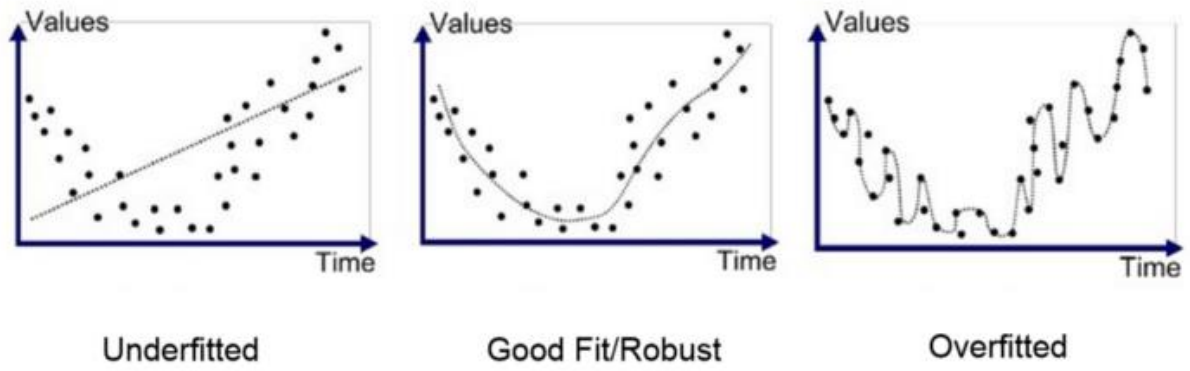
Makine öğrenimi ve derin öğrenme, yapay zeka alanında sıkça kullanılan iki kavramdır. Makine öğrenimi, bir veri setindeki kalıpları ve ilişkileri öğrenmek için bilgisayar algoritmalarının kullanılmasıdır. Derin öğrenme ise, makine öğrenimi alanının alt kategorilerinden biridir ve büyük veri kümelerindeki karmaşık yapıları anlamak için çok katmanlı yapay sinir ağları kullanır.

Makine öğrenimi, bir veri kümesindeki kalıpları, özellikleri ve ilişkileri öğrenmek için çeşitli algoritmalar kullanır. Bu algoritmalar, veri kümesindeki özelliklerin veya girdilerin matematiksel modellerini oluşturarak, veri seti üzerinde tahmin yapmak, sınıflandırmak veya çıkarımda bulunmak için kullanılabilir. Makine öğrenimi algoritmaları, karar ağaçları, doğrusal regresyon, k-means kümeleme, destek vektör makineleri (SVM) gibi birçok farklı yöntem içerir.

Derin öğrenme ise, yapay sinir ağları kullanarak veri kümesindeki daha karmaşık kalıpları ve yapıları öğrenir. Derin öğrenme algoritmaları, büyük veri kümelerindeki görsel, işitsel veya metin verileri gibi yüksek boyutlu verileri işleyebilir ve öğrenebilir. Bu nedenle, derin öğrenme, görüntü tanıma, ses tanıma, doğal dil işleme ve özerk araçlar gibi alanlarda kullanılır.

Özetle, makine öğrenimi genellikle verilerin özelliklerini matematiksel modellerle eşleştirerek bir çıktı elde etmek için kullanılan bir dizi algoritmadır. Derin öğrenme ise, büyük veri kümelerindeki karmaşık yapıları anlamak için çok katmanlı sinir ağlarının kullanılmasıdır.

16. Overfitting ve Underfitting Arasındaki Farklar Nelerdir?



Şekil 16.1 Underfitting – Overfitting [7]

Overfitting ve Underfitting, makine öğrenimi ve istatistiksel modelleme gibi alanlarda yaygın olarak kullanılan terimlerdir. Bu terimler, bir modelin eğitim verileri üzerinde ne kadar iyi performans gösterdiğine bağlı olarak modelin genelleştirilebilirliği hakkında bilgi verir. Overfitting, bir modelin eğitim verilerine çok fazla uyum sağlaması durumudur. Bu durumda, model aşırı karmaşık hale gelebilir ve eğitim verilerindeki her bir özellik veya desene aşırı derecede duyarlı hale gelebilir. Sonuç olarak, model, eğitim verileri üzerinde mükemmel bir performans sergilese de, gerçek dünya verilerinde hatalı sonuçlar üretebilir. Overfitting, modeldeki aşırı karmaşıklığın azaltılması ve genelleştirilebilirliğin artırılması yoluyla ele alınabilir. Underfitting, bir modelin eğitim verilerine yeterince uyum sağlayamaması durumudur. Bu durumda, model yetersiz veya basit hale gelebilir ve eğitim verilerindeki önemli özellikleri veya desenleri kaçırabilir. Sonuç olarak, model, eğitim verileri üzerinde düşük bir performans sergileyebilir ve gerçek dünya verilerinde daha doğru sonuçlar elde edilebilir. Eksik yerleştirme, modelin daha karmaşık hale getirilmesi veya daha fazla veri toplanması yoluyla ele alınabilir. Genel olarak, bir modelin iyi bir performans sergilemesi için, hem Overfitting

hem de Underfitting problemlerinden kaçınılmalı ve optimum bir denge bulunmalıdır. Bu nedenle, model eğitiminde kullanılan verilerin uygun şekilde bölünmesi, doğru bir özellik seçimi, doğru hiperparametre ayarlamaları ve gerektiğinde farklı modellerin denenmesi gibi stratejiler kullanılabilir.

17. Düzenleme (Regularisation) Nedir? Neden Faydalıdır?

Düzenleme, Overfitting'i azaltmak ve daha iyi genelleme yapmak için kullanılan bir yöntemdir. Overfitting, bir modelin eğitim verilerinde çok iyi performans göstermesine rağmen, test verilerinde kötü performans göstermesine neden olabilir. Bu, modelin öğrenme verilerindeki gürültüyü veya özelikleri ezberlemesinden kaynaklanabilir. L1 düzenleme, regresyon problemlerinde kullanılan bir yöntemdir ve Lasso olarak da bilinir. Bu yöntem, modelin kayıp fonksiyonuna, modeldeki her bir özelliğin katkısının mutlak değerinin toplamını ekleyerek bir düzenleme terimi ekler. Bu, bazı özelliklerin modeldeki katkısını sıfıra indirerek, modeldeki değişken sayısını azaltır. Bu nedenle, L1 düzenleme, modele seyrekliği (sparse) kazandırarak, özellik seçimi (feature selection) yapmak için kullanılabilir. L2 düzenleme, regresyon ve sınıflandırma problemlerinde kullanılan bir yöntemdir ve Ridge olarak da bilinir. Bu yöntem, modelin kayıp fonksiyonuna, modeldeki her bir özelliğin katkısının karesinin toplamının yarısını ekleyerek bir düzenleme terimi ekler. Bu, tüm özelliklerin modelde katkısını azaltarak modelin genelleştirme yeteneğini artırır. Bu yöntem, L1 düzenlemeye kıyasla daha yumuşak bir etkiye sahiptir ve genellikle tüm özellikleri modelde bırakır. Düzenleme, overfitting'i azaltmanın yanı sıra, veri setindeki gürültü ve anormalliklerin etkisini azaltarak daha güvenilir sonuçlar elde etmemizi sağlar.

18. Seçim Önyargısı (Selection Bias) Nedir?

Seçim önyargısı, bir veri kümesi veya popülasyonun belirli bir örneklemini seçerken, yanlı veya hatalı bir seçim yapma eğilimini ifade eder. Bu önyargı, örneklemin seçim kriterleri veya örneklemleme yöntemi nedeniyle ortaya çıkabilir.

Örneğin, bir çalışma yalnızca üniversite mezunlarından oluşan bir örneklem kullanarak bir işletme okulundaki öğrencilerin performansını ölçmek istiyor. Ancak bu, işletme okuluna kayıtlı olmayan veya üniversite mezunu olmayan öğrencilerin performansını ölçmeyi engeller, bu da sonuçların yanlış yorumlanmasına neden olabilir. Bu, örneklem seçimindeki bir önyargı örneğidir.

Seçim önyargısı, bir çalışmanın sonuçlarının yanıltıcı olmasına neden olabilir ve bu nedenle araştırmacılar tarafından dikkate alınması gereken bir önemli bir sorundur.

19. Doğrulama Seti ile Test Seti Arasındaki Farklar Nelerdir?

Makine öğrenmesi modellerinin performansını ölçmek için genellikle veriler bölünür ve ayrı ayrı kullanılır. Bu bölünme genellikle eğitim, doğrulama ve test setleri arasında yapılır. Eğitim seti, modelin eğitimini yapmak için kullanılan veriler kümesidir. Model, eğitim setindeki örnekleri kullanarak özellikleri ve hedefleri arasındaki ilişkileri öğrenir. Doğrulama seti, eğitim setinden ayrılmış bir veri kümesidir. Modelin eğitimi tamamlandıktan sonra, doğrulama seti genellikle modelin performansını değerlendirmek için kullanılır. Doğrulama seti, modelin eğitiminde kullanılan verilerden farklı olmalıdır, ancak aynı zamanda modelin genelleştirilebilirliğini ve performansını ölçmek için yeterli miktarda örnek içermelidir. Test seti, son olarak, modelin performansını kesin olarak ölçmek için kullanılır. Test seti, hem eğitim setinden hem de doğrulama setinden farklı bir veri kümesidir ve modelin hiçbir zaman görmediği örnekleri içermelidir. Test seti, modelin gerçek dünya verilerine ne kadar iyi uyarlandığını ve genelleştirilebilirliğini değerlendirmek için kullanılır. Doğrulama seti ve test seti arasındaki ana fark, amaçlarıdır. Doğrulama seti, eğitim setindeki hiperparametrelerin ayarlanmasına yardımcı olmak için kullanılırken, test seti modelin performansını nihai olarak ölçmek için ayrılır. Bu şekilde, modelin performansının gerçek dünya verilerine ne kadar iyi uyarlandığı ve genelleştirilebilirliği daha güvenilir bir şekilde ölçülebilir.

20. Regresyon ve Sınıflandırma Makine Öğrenmesi Teknikleri Arasındaki Fark Nedir?

Regresyon ve sınıflandırma, makine öğrenimi alanında yaygın olarak kullanılan iki farklı tekniktir. Her iki teknik de girdi verilerine dayalı olarak bir çıktı üretir, ancak amaçları farklıdır.

Regresyon, bir bağımlı değişkenin bağımsız değişkenlere bağlı olarak nasıl değiştiğini anlamak için kullanılır. Yani, regresyon modelleri bir sürekli değişkenin tahmin edilmesinde kullanılır. Örneğin, bir evin fiyatını etkileyen değişkenlerin (oda sayısı, evin yaşı, konumu vb.) analizi için regresyon kullanılabilir. Öte yandan, sınıflandırma, girdi değişkenlerine dayalı olarak bir örneğin bir sınıfa veya kategoriye ait olup olmadığını tahmin etmek için kullanılır.

Sınıflandırma modelleri genellikle verilerin etiketli olduğu durumlarda kullanılır. Örneğin, bir müşterinin bir ürünü satın alıp almayacağını tahmin etmek için sınıflandırma kullanılabilir.

Bu nedenle, regresyon modelleri çıktı olarak bir sayısal değer üretirken, sınıflandırma modelleri çıktı olarak bir kategori veya sınıf üretir.

21. Yapay Sinir Ağları Nedir?

Yapay sinir ağları, biyolojik sinir ağlarından esinlenerek tasarlanmış matematiksel modellerdir. Bu algoritma, makine öğrenimi ve yapay zeka gibi konular için temel bir tekniktir. Yapay sinir ağları, girdileri alır, bu girdileri işler, ağırlıklandırır ve çıktılar üretir. Bu nedenle, birçok farklı işlem için kullanılabilirler, örneğin görüntü işleme, doğal dil işleme, ses işleme ve tahmin yapma gibi alanlarda kullanılabilirler. Yapay sinir ağları, birçok farklı katman türü ve düğüm türü kullanır. Örneğin, giriş katmanı, girdileri kabul eder ve işler. Gizli katmanlar, girdileri ağırlıklandırır ve çıktı katmanı son çıktıları üretir. Bu ağlarda, ağırlıklar ve eşikler genellikle rastgele başlatılır ve ardından ağ, hata fonksiyonlarını hesaplayarak eğitilir. Eğitim süreci, geriye doğru yayılma (backpropagation) olarak adlandırılan bir yöntemle gerçekleştirilir. Yapay sinir ağları, derin öğrenme gibi tekniklerle birleştirilerek, karmaşık verileri analiz etmek ve çıktıları tahmin etmek için kullanılabilirler. Bu tekniklerin kullanımı, yapay zeka ve makine öğrenimi alanlarında hızla yayılmaktadır ve birçok uygulama için son derece faydalıdır.

22. Bir Veri Bilimcisi' nin Kullandığı Araçlar Nelerdir?

Veri bilimcisi, birçok araç ve cihazı kullanmaktadır. Bunların bazıları şunlardır:

- Programlama dilleri: Python, R, SQL gibi programlama dilleri, veri analizi, modelleme ve makine öğrenimi algoritmaları oluşturmak için temel araçlardır.
- Veri Depolama Sistemleri: Büyük veri kümelerini saklamak ve yönetmek için veri depolama sistemleri kullanılır. Bunlara örnek olarak Hadoop, Cassandra, Amazon S3 ve Redshift gibi araçlar sayılabilir.
- Veri Görselleştirme Araçları: Verilerin görselleştirilmesi, verilerin daha anlaşılır ve anlamlı hale getirilmesine yardımcı olur. Bu amaçla kullanabileceğimiz araçlar arasında Tableau, Power BI, Matplotlib ve Seaborn gibi araçlar bulunmaktadır.
- Makine Öğrenimi Kütüphaneleri: Makine öğrenimi modellerinin oluşturulması, eğitilmesi ve değerlendirilmesi için birçok kütüphane mevcuttur. Bunlar arasında Scikit-learn, TensorFlow, Keras ve PyTorch gibi kütüphaneler bulunur.
- Büyük Veri İşleme Araçları: Verilerin işlenmesi ve analiz edilmesi genellikle büyük veri kümeleri içerir. Büyük veri işleme araçları arasında Apache Spark, Hadoop, Storm ve Flink gibi araçlar bulunmaktadır.

- Bulut Bilişim Platformları: Verilerin depolanması, işlenmesi ve analiz edilmesi için bulut bilişim platformlarına başvurulur. Bu platformlar arasında Amazon Web Services (AWS), Google Cloud Platform (GCP) ve Microsoft Azure bulunmaktadır.
- İşbirliği ve Proje Yönetimi Araçları: Projelerde işbirliği yapmak ve iş akışını yönetmek için araçlara ihtiyaç duyulur. Bunlar arasında GitHub, JIRA, Trello, Asana ve Slack gibi araçlar bulunmaktadır.

Veri bilimcilerinin kullanabileceği araçlar ve cihazlar hızla değişen bir alandır. Dolayısıyla, veri bilimcileri güncel kalmak için sürekli olarak yeni araçlar ve teknolojiler öğrenmeli ve kullanmaya devam etmelidirler.

23. Doğal Dil İşleme (NLP) Nedir?

Doğal Dil İşleme (NLP), bilgisayarların insan dilini anlamasına, yorumlamasına ve üretmesine olanak tanıyan bir alan olarak tanımlanabilir. NLP, makine öğrenimi, yapay zeka ve dilbilim alanlarının kesiştiği bir alandır. NLP kullanarak doğal dilde yazılmış veya konuşulmuş verileri analiz edebilir, anlamlı bilgi çıkarabilir ve hatta bu verileri insanların anlayabileceği bir formatta sunabilirsiniz. NLP'nin gerçek hayattan bazı örnekleri şunlardır:

- Otomatik Dil Çevirisi: NLP, Google Translate gibi otomatik dil çeviri araçlarının arkasındaki teknolojidir. NLP modelleri, bir dilde yazılmış veya konuşulmuş metni, başka bir dile çevirmek için kullanılabilir.
- Sosyal Medya Analizi: Sosyal medya platformları, NLP algoritmalarını kullanarak kullanıcıların paylaşımlarını ve yorumlarını analiz edebilir. Bu analizler, müşteri memnuniyeti, ürün/hizmet kalitesi ve diğer konular hakkında fikir sahibi olmak için kullanılabilir.
- Metin Madenciliği: NLP, büyük metin veri setleri üzerinde çalışarak, özetleme, anahtar kelime çıkarma, kelime frekansı sayımı ve diğer analizler yapabilir. Bu analizler, birçok sektörde kullanılabilir, örneğin haberler, pazarlama, tıp ve hukuk.
- Chatbotlar: NLP, chatbotların arkasındaki teknolojidir. Chatbotlar, kullanıcıların doğal dilde sorduğu sorulara cevap vermek için tasarlanmıştır ve NLP modellerini kullanarak kullanıcının sorusunu anlayabilir ve buna uygun bir yanıt üretebilir.
- Metin Sınıflandırması: NLP, metin sınıflandırması için kullanılabilir. Örneğin, spam filtreleri, NLP algoritmalarını kullanarak spam mesajlarını tespit edebilir. Diğer bir örnek, haberlerin siyasi veya ekonomi gibi kategorilerde sınıflandırılmasıdır.

Bu örnekler, NLP'nin gerçek hayatta ne kadar yaygın kullanıldığını göstermektedir. NLP, insanlar ve makineler arasındaki doğal dil iletişimini geliştirerek, birçok endüstri için büyük faydalar sağlayabilir.

24. Normalizasyon Nedir? Normalleştirme ve Standardizasyon Arasındaki Fark Nedir?

Veri normalizasyonu, veri özelliklerinin farklı aralıklarda olmasından kaynaklanan problemleri çözmek için veri işleme işlemidir. Normalizasyon, verileri sabit bir aralığa sığdırmayı amaçlayarak özelliklerin ölçeklerini veya birimlerini karşılaştırılabilir hale getirir. Normalleştirme, verilerin 0 ile 1 arasında ölçeklendirilmesi işlemidir. Bu, verilerin en küçük değerlerini 0, en büyük değerlerini 1 olarak ayarlar ve geri kalan değerleri aradaki oranlara göre ölçeklendirir. Standardizasyon, verilerin ortalama değerlerinin sıfır, standart sapmalarının ise bir olarak ayarlanması işlemidir. Standardizasyon, normalleştirme gibi değerleri belirli bir aralığa sığdırmaz, ancak verilerin özelliklerinin dağılımını daha iyi korur. Örneğin, bir veri setinde bir özellik, diğer özelliklere göre çok daha büyük sayılar alabilir. Bu, bu özelliklerin analizinde diğer özelliklerin gölgede kalmasına neden olabilir. Bu durumda normalizasyon veya standardizasyon uygulanabilir. NLP uygulamalarında da metin verileri için normalizasyon ve standardizasyon sıklıkla kullanılmaktadır.

KAYNAKLAR

[1] https://www.hasanyildiz.com/23-veri-bilimi-mulakat-sorusu/#close

[2] https://developer.ibm.com/developer/default/tutorials/build-a-logistic-regression-neural-network-using-tensorflow/images/figure1.png

[3] https://miro.medium.com/v2/resize:fit:720/format:webp/1*fxiTNIgOyvAombPJx5KGeA.png

[4] https://www.slideshare.net/Simplilearn/confusion-matrix-in-machine-learning-confusion-matrix-explained-with-example-simplilearn/Simplilearn/confusion-matrix-in-machine-learning-confusion-matrix-explained-with-example-simplilearn

[5] https://why-change.com/2021/11/13/how-to-create-decision-trees-for-business-rules-analysis/

[6] http://www.analitikbeyin.com/blog/icerik/normal-dagilim-nedir

[7] https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76