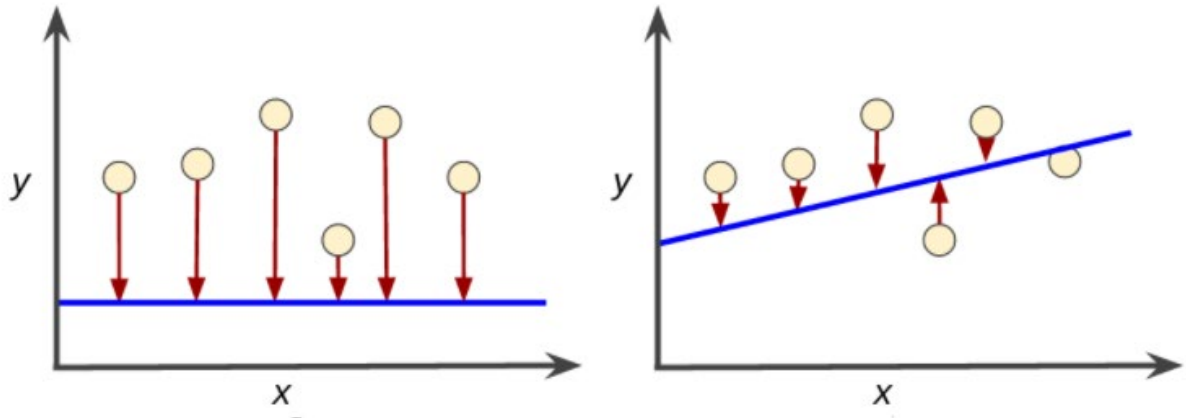
2.3. Eğitim Modelleri

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.

Makine öğrenmesi modelleri iyi performans gösterebilmeleri için çok fazla veri gerektirir. Bir makine öğrenmesi modeli eğitilirken, temsili veri örneklerinin toplanması gerekir. Eğitim setindeki veriler, bir metin topluluğuna, bir resim koleksiyonuna ve bir hizmetin tek tek kullanıcılarından toplanan verilerine kadar değişebilir.

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.

Makine öğrenmesinde kayıp, kötü bir tahminin cezasıdır. Kayıp, modelin tahmininin ne kadar kötü olduğunu gösteren bir sayıdır. Modelin tahmini mükemmelse, kayıp sıfırdır; aksi takdirde kayıp daha büyüktür. Makine öğrenmesinde bir modeli eğitmenin amacı, tüm örneklerde ortalama olarak düşük kayıplı bir eşik değer bulmaktır. Örneğin, aşağıdaki şekilde, solda yüksek kayıplı bir modeli ve sağda düşük kayıplı bir modeli gösterilmektedir. Şekilde okların uzunluğu kaybı temsil eder. Mavi çizgiler tahminleri temsil eder.



Şekil. Sol modelde yüksek kayıp; doğru modelde düşük kayıp.

Soldaki grafikteki okların sağdaki grafikteki benzerlerinden çok daha uzun olduğuna dikkat edin. Açıkçası, sağdaki çizimdeki çizgi, soldaki grafikteki çizgiden çok daha iyi bir tahmin modelidir. Kayıpları anlamlı bir şekilde bir araya getirecek bir matematiksel fonksiyon - bir kayıp fonksiyonu – kullanılır.

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

Tek bir örnek için kayıpların karesi = *etiket ve tahmin arasındaki farkın karesi*

= *(gözlem - tahmin (x))2*

= *(y - y ') 2*

Ortalama kareler hatası (MSE), tüm veri kümesi boyunca örnek başına düşen ortalama kare kayıptır. MSE'yi hesaplamak için, tek tek örnekler için tüm kayıpların karesi toplanır ve ardından örneklerin sayısına bölünür:

$$MSE = \frac{1}{N}\sum_{(x,y)\in D}(y - yt)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - yt_i)^2$$

burada:

(x,y) bir örnektir

x, modelin tahminlerde bulunmak için kullandığı özellikler kümesidir.

y, örneğin etiketidir.

Tahmin(x), yt: özellikler kümesi ile birlikte ağırlıkların ve önyargının bir fonksiyonudur.

D, (x,y) çiftler olan birçok etiketli örneği içeren bir veri kümesidir.

N, D içindeki örneklerin sayısıdır.

MSE, makine öğreniminde yaygın olarak kullanılmasına rağmen, ne tek pratik kayıp işlevi ne de tüm koşullar için en iyi kayıp işlevi değildir.

Örnek:

Makine öğrenmesinde bir modeli eğitmenin amacı, tüm örneklerde ortalama olarak düşük kayıplı bir eşik değer bulmaktır. Bu eşik değer ortalama kareler hatası ile bulunur.

i=1...5

örnek değerler, yi= 1, 2, 3, 4, 5 ;

modelin örnekleme tahmin değerleri yti=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, yi= 1, 2, 3, 4, 5 ;

modelin örnekleme tahmin değerleri yti=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

Bir topluluk yönteminin genel ilkesi, tek bir modele göre sağlamlığı artırmak için belirli bir öğrenme algoritması ile oluşturulmuş birkaç modelin tahminlerini birleştirmektir. Torbalama, istikrarsız tahmin veya sınıflandırma şemalarını iyileştirmek için topluluk içinde bir yöntemdir. Arttırma yöntemi, kombine modelin yanlılığını azaltmak için sırayla kullanılır. Artırma ve Torbalama, varyans terimini azaltarak hataları azaltabilir.

Bir öğrenme algoritmasının beklenen hatası, önyargı ve varyansa ayrıştırılabilir. Bir önyargı terimi, öğrenme algoritması tarafından üretilen ortalama sınıflandırıcının hedef işlevle ne kadar yakından eşleştiğini ölçer. Varyans terimi, öğrenme algoritmasının tahmininin farklı eğitim setleri için ne kadar dalgalandığını ölçer.

Toplulukta Artımlı Öğrenme algoritması, bir algoritmanın, sınıflandırıcı halihazırda mevcut olan veri kümesinden oluşturulduktan sonra mevcut olabilecek yeni verilerden öğrenme yeteneğidir.