1.6 Test Etme ve Tahminlerin Doğruluğunu Ölçme

Bir modelin yeni gelen örnek verilerde ne kadar iyi genelleşebileceğini bilmenin tek yolu yeni veriler üzerinde denemektir. Verilerinizi iki gruba ayırırsınız; **eğitim seti** ve **test seti**. Eğitim setini kullanarak modelinizi eğitirsiniz ve test setini kullanarak test edersiniz. Test setinde hata oranınızı hesaplarsınız ve bu değer, modelinizin daha önce hiç görmediği örneklerde ne kadar iyi performans göstereceğini size gösterir.

Şimdi bazı temel kavramlara bir bakalım.

1.6.1 Doğruluk, Hassasiyet, Hata

Hassasiyet, tekrarlanabilirliği ifade ederken, **doğruluk** ölçülen değerin gerçek değere ne kadar yakın olduğunu gösterir.

Doğruluk ve Kesinlik(Hassasiyet)

Doğruluk, bir ölçümün o ölçüm için doğru değere ne kadar yakın olduğudır. Bir ölçüm sisteminin hassasiyeti, tekrarlanan ölçümler arasında ölçümlerin birbirine ne kadar yakın olduğunu ifade eder (aynı koşullar altında tekrarlanma). Ölçümler hem doğru hem kesin, doğru ama kesin değil, kesin ama doğru değil, ya da hiçbiri olmayabilir

Yüksek Doğruluk, Düşük Hassasiyet

Bu hedef tahtasına göre, vuruşların hepsi merkeze yakın, ancak hiçbiri birbirine yakın değil; Bu hassas olmayan bir doğruluk örneği.

Düşük Doğruluk, Yüksek Hassasiyet

Bu hedef tahtasına göre, isabetlerin hepsi birbirine yakın, ancak hedefe yakın değil; Bu doğruluğu olmayan bir hassasiyet örneği.

Hassasiyet bazen şu şekilde ayrılır:

Tekrarlanabilirlik – Aynı araç ve operatörü kullanarak koşulları sabit tutmak ve ölçümleri kısa bir süre tekrarlamak için her türlü çabanın gösterilmesi sonucunda ortaya çıkan varyasyon.
Yinelenebilirlik – Farklı ölçüm cihazları ve operatörler arasında aynı ölçüm sürecini kullanarak ve daha uzun sürelerle ortaya çıkan varyasyon.

Hata

Tüm ölçümler, sonucun belirsizliğine katkıda bulunan hataya tabidir. Hatalar insan hatası veya teknik hata olarak sınıflandırılabilir.

Teknik hata iki kategoriye ayrılabilir: **rastgele hata** ve **sistematik hata**. İsimlerinden de anlaşılacağı üzere, rastgele hata, periyodik olarak tanınabilir bir desen olmadan ortaya çıkar. Sistematik hata, cihazla ilgili bir sorun olduğunda ortaya çıkar. Örneğin, bir ölçek yanlış kalibre edilmiş olabilir ve üzerinde hiçbir şey olmayan 0,5 g okunabilir. Bu nedenle

tüm ölçümler 0,5 g ile fazla tahmin edilecektir. Ölçümünüzde bunu hesaba katmadığınız sürece ölçümünüzde bir miktar hata oluşacaktır.

Doğruluk, Hassaslık ve Hata birbirleriyle nasıl ilişkilidir?

Rastgeleli hata, daha doğru bir aletle (ölçümler daha ince artışlarla yapılır) ve daha fazla tekrarlanabilirlik veya yinelenebilirlik ile daha küçük olacaktır.

Alınan ölçümler ne kadar çok olursa, bir miktarın gerçek değerini o kadar yakından öğrenebiliriz. Çoklu ölçümlerle (tekrarlar ile), sonuçların hassaslığını değerlendirip, sistemde herhangi bir sistematik hata olmasaydı ortalama değerin gerçek değere ne kadar yakın olacağını tahmin etmek için basit istatistikleri uygulayabiliriz. Ölçümlerin sayısı arttıkça ortalama, *"gerçek değer"* den daha az sapma gösterecektir.

1.6.2 Çapraz Doğrulama

Verileri Test Etme

Bir makine öğrenmesi modelinin performansını değerlendirirken bir sonraki önemli soru, model performansını değerlendirmek için hangi veri kümesinin kullanılması gerektiğidir. Makine öğrenmesi modeli, eğitim seti kullanılarak basitçe test edilemez, çünkü çıktı, önyargılı olur, çünkü makine öğrenmesi modeli eğitimi boyunca, tahmin edilen sonucu önceden eğitim veri setine ayarlamıştır. Dolayısıyla, genelleme hatasını tahmin edebilmek için modelin henüz görmediği bir veri kümesini test etmesi gerekir; test veri kümesi.

Bu nedenle, modelin test edilmesi amacıyla, etiketli bir veri kümesi gerekir. Bu, eğitim veri setini eğitim veri setine ve test veri setine bölerek başarılabilir. Bu, k-kat çapraz doğrulama, jackknife yeniden örnekleme ve önyükleme gibi çeşitli tekniklerle başarılabilir. A / B testi gibi teknikler, gerçek kullanıcı etkileşiminden gelen tepki karşısında üretimdeki makine öğrenmesi modellerinin performansını ölçmek için kullanılır.

Çapraz Doğrulama, bir makine öğrenmesi modelinde yapılan testin hatasını daha iyi tahmin edebilmek için model seçiminde kullanılan bir tekniktir. Çapraz doğrulamanın arkasındaki fikir, eğitim verileri setinden doğrulama kümeleri olarak bilinen örnek gözlem bölümlerini oluşturmaktır. Bir modeli eğitim verilerine yerleştirdikten sonra, performansı, her yeni doğrulama kümesine karşı ölçülür ve daha sonra, yeni gözlemleri öngörmek istenildiğinde modelin nasıl performans göstereceğine ilişkin daha iyi bir değerlendirme elde edilir. Yapılacak bölüm sayısı, örnek veri kümesindeki gözlem sayısına ve önyargı varyansı dengelemesine ilişkin kararın, daha fazla bölünmenin daha küçük bir yanlılığa yol açmasına ve daha fazla varyansa bağlı olarak değişmesine bağlıdır.

Holdout yöntemi çapraz doğrulamanın en basit çeşididir. Veri seti, eğitim seti ve test seti olarak adlandırılan iki gruba ayrılmıştır. İşlev yaklaşımcısı, yalnızca eğitim setini kullanarak bir işleve uyar. Sonra, fonksiyon yaklaşımından test setindeki verilerin çıkış değerlerini tahmin etmesi istenir (daha önce bu çıkış değerlerini hiç görmemiş). Yaptığı hatalar, modeli değerlendirmek için kullanılan ortalama mutlak test kümesi hatasını vermek için daha önce olduğu gibi biriktirilir. Bu yöntemin avantajı artık yöntemin tercih edilmesi ve hesaplamanın artık gerekmemesidir. Bununla birlikte, değerlendirmesi çok değişken olabilir. Değerlendirme, hangi veri noktalarının eğitim setine girdiğine ve hangi test grubuna dönüştüğüne bağlı olabilir ve bu nedenle değerlendirme, bölümün nasıl yapıldığına bağlı olarak önemli ölçüde farklılık gösterebilir.

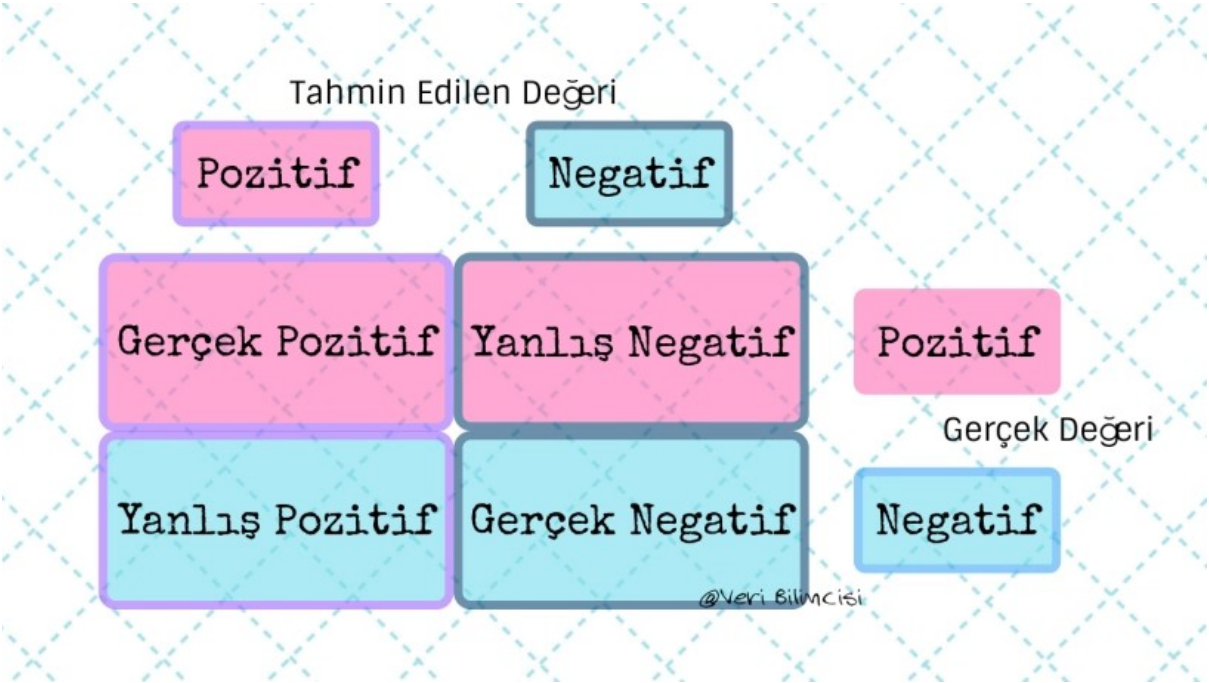
K katlamalı çapraz doğrulama, holdout yöntemini geliştirmenin bir yoludur. Veri kümesi k alt küme altına bölünmüştür ve bekletme yöntemi k kez tekrarlanmaktadır. Her defasında, k alt kümelerinden biri test kümesi olarak kullanılırken diğer k-1 alt kümeleri bir eğitim kümesi oluşturmak üzere bir araya getirilir. Ardından, tüm k denemelerindeki ortalama hatası hesaplanır. Bu yöntemin avantajı, verilerin nasıl bölündüğünü daha az önemsemektir. Her veri noktası tam olarak bir kez test kümesine girer ve k-1 kez bir eğitim setine girer. Sonuç tahmini varyansı, k arttıkça azaltılır. Bu yöntemin dezavantajı, eğitim algoritmasının sıfırdan k kere tekrarlanması gerektiğidir, yani değerlendirme yapmak için k kere kadar hesaplama gerektirir. Bu yöntemin bir varyantı, verileri rastgele bir test ve

eğitim setine k farklı zamanlara bölmektir. Bunu yapmanın avantajı, her bir test kümesinin ne kadar büyük olduğunu ve kaç denemenizin bağımsız olduğunu seçebilmenizdir.

Tek-çıkışlı çapraz doğrulama, K katlı çapraz doğrulamasının mantıksal uç noktasına getirilmesidir; K, setteki veri noktalarının sayısı N'ye eşittir. Bunun anlamı N'nin ayrı zamanlarda, fonksiyon yaklaşımcısının bir nokta dışındaki tüm veriler üzerinde eğitilmesi ve bu nokta için bir tahmin yapılmasıdır. Daha önce olduğu gibi ortalama hata hesaplandır ve modeli değerlendirmek için kullanılır. Tek-çıkışlı çapraz doğrulama hatası (leave-one-out cross validation error)(LOO-XVE) ile verilen değerlendirme iyidir, ancak ilk geçişte hesaplanması çok pahalı görünmektedir. Neyse ki, yerel olarak ağırlıklandırılmış öğreniciler, LOO tahminlerini normal tahminler yaparken olduğu kadar kolay yapabilir. Bu, LOO-XVE'nin hesaplanması artık hatanın hesaplanmasından daha fazla zaman alması anlamına gelmez ve modelleri değerlendirmek için çok daha iyi bir yoldur.

1.6.3 Karışıklık Matrisi ve Performans Ölçütleri

Karışıklık Matrisi



Bir karışıklık matrisi, gerçek değerlerin bilinmekte olduğu bir dizi test verisi üzerinde bir sınıflandırma modelinin performansını tanımlamak için sıklıkla kullanılan bir tablodur (veya

“sınıflandırıcı”). Karmaşıklık matrisini anlamak kolaydır, ancak terminolojisi kafa karıştırıcı olabilir.

Bir ikili sınıflandırıcı için bir örnek karışıklık matrisiyle başlayalım (ancak birden fazla sınıfın olması durumunda kolayca genişletilebilir):

İkili sınıflandırıcı için örnek karışıklık matrisi

n=165	Tahmini Değeri: HAYIR	Tahmini Değeri: EVET
Gerçek Değeri: HAYIR	50	10
Gerçek Değeri: EVET	5	100

Bu Matristen Ne Öğrenebiliriz?

Tahmini iki sınıf vardır: “evet” ve “hayır”. Örneğin, bir hastalığın varlığını önceden tahmin edersek “evet” hastalığa yakalanmış oldukları anlamına gelir ve “hayır” hastalığa yakanılmadığı anlamına gelebilir.
Sınıflandırıcı toplam 165 tahmin yapmış.
Bu 165 olayın içinde sınıflandırıcı 110 kez “evet” ve 55 kez “hayır” tahmin etmiş.

Şimdi en temel terimleri tanımlayalım:

Gerçek Pozitifler (TP): Bunlar evet tahmin ettiğimiz vakalardır (hastalıkları vardır) ve hastalıkları vardır.
Gerçek Negatifler (TN): Hayır öngördük ve hastalığı yok.
Yanlış Pozitifler (FP): Evet tahmin ettik, ancak aslında hastalığı yok. (“Tip I hatası” olarak da bilinir)
Yanlış Negatifler (FN): Hayır öngördük ama aslında hastalığa sahipler. (“Tip II hata” olarak da bilinir.)

İkili sınıflandırıcı için örnek karışıklık matrisi

n=165	Tahmini Değeri: HAYIR	Tahmini Değeri: EVET	
Gerçek Değeri: HAYIR	TN = 50	FP = 10	60
Gerçek Değeri: EVET	FN = 5	TP = 100	105
	55	110	

Bu, genellikle bir ikili sınıflandırıcı için bir karışıklık matrisinden hesaplanan oranlar listesidir:

Doğruluk (Accuracy Rate): Genel olarak, sınıflayıcı ne sıklıkta düzeltilir?
(TP + TN) / toplam = (100 + 50) / 165 = 0.91
Yanlış Sınıflandırma Oranı (Misclassification Rate): Genel olarak, ne sıklıkta yanlış?
(FP + FN) / toplam = (10 + 5) / 165 = 0.09
1 eksi doğrulukla eşdeğerdir.
“Hata Oranı” olarak da bilinir.(“Error Rate”)
Doğru Olumlu Hız (True Positive Rate): Aslında evet olduğunda, evet ne sıklıkta öngörülebilir?
TP / gerçek evet = 100/105 = 0.95
“Hassasiyet” veya “Hatırlama” olarak da bilinir.(“Sensitivity” or “Recall”)
Yanlış Olumlu Hız (False Positive Rate): Aslında hayır olduğunda, evet ne sıklıkta öngörülebilir?
FP / gerçek no = 10/60 = 0.17
Özgüllük (Specificity:): Aslında hayır olduğunda, ne kadar sıklıkla hayır diyeceksiniz?
TN / gerçek no = 50/60 = 0.83
1 eksi yanlış Pozitif Orana eşdeğer
Hassasiyet (Precision): Evet tahmininde bulunduğunda, ne sıklıkta düzeltilir?
TP / tahmini evet = 100/110 = 0.91
Yaygınlık (Prevalence): Örneklemimizde evet durumu ne sıklıkta ortaya çıkmaktadır?

Gerçek evet / toplam = 105/165 = 0.64

Olumlu Tahmin Edici Değer (Positive Predictive Value): Bu, yaygınlık dikkate alınması dışında hassasiyetle çok benzerdir. Sınıfların mükemmel şekilde dengelenmesi durumunda (yaygınlık %50 olduğu anlamına gelir), pozitif öngörme değeri (PPV) hassasiyetle eşdeğerdir.
Null Hata Oranı (Null Error Rate): Çoğunluk sınıfını hep öngördüyseniz, bu sıklıkta yanlış olur. (Örneğimizde, boş hata oranı 60/165 = 0.36 olur, çünkü her zaman evet tahmin ettiyseniz, yalnızca 60 “hayır” vakası için yanlıştır.) Bu, sınıflandırıcınızı karşılaştırmak için yararlı bir temel metrik olabilir. Bununla birlikte, Doğruluk Paradoksu tarafından gösterildiği gibi, belirli bir uygulama için en iyi sınıflandırıcı bazen boş hata oranından daha yüksek bir hata oranına sahip olacaktır.
Cohen’s Kappa: Sınıflandırıcının aslında ne kadar iyi performans gösterdiğinin bir ölçüsüdür. Diğer bir deyişle, doğruluk ve boş hata oranı arasında büyük bir fark varsa, bir modelin yüksek bir Kappa puanı olacaktır.
F Puan (F Score): Bu, gerçek olumlu oranın (recall) ve hassasiyetin ağırlıklı ortalamasıdır.
ROC Eğrisi (ROC Curve): Bu, sınıflandırıcının tüm olası eşikler üzerinde performansını özetleyen sık kullanılan bir grafiktir. Belirli bir sınıfa gözlem atanması eşiğini değiştirdiğinizde Yanlış Olumlu Orana (x ekseni) karşı Gerçek Olumlu Oranı (y ekseni) çizerek oluşturulur.

1.6.4 Metrikler

Makine öğrenmesinde tahminlerdeki hataları ölçmek için çeşitli istatistiksel ölçüler kullanılır. Makine öğrenmesi sisteminin maliyetini ölçmek için kullanılan metriklerden bazılarını aşağıda bulabilirsiniz.

Ortalama Mutlak Hata — Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|x_t - \hat{x}_t\right|$$

Ortalama Kare Hata — Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Kök Kareler Karesi — Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{n}}$$

Ortalama Mutlak Yüzde Hatası — Mean Absolute Percent Error (MAPE)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i - \widehat{Y}_i\right|}{Y_i}$$