Yapay Öğrenme Ders Notları
Hazırlayan: Barış Baykant Alagöz
İnönü Üniversitesi, Bilgisayar Mühendisliği, 2023-2024 Bahar Dönemi Lisans Dersi
Sorularınız ve düzeltme önerileriniz için e-mail: baykant.alagoz@inonu.edu.tr

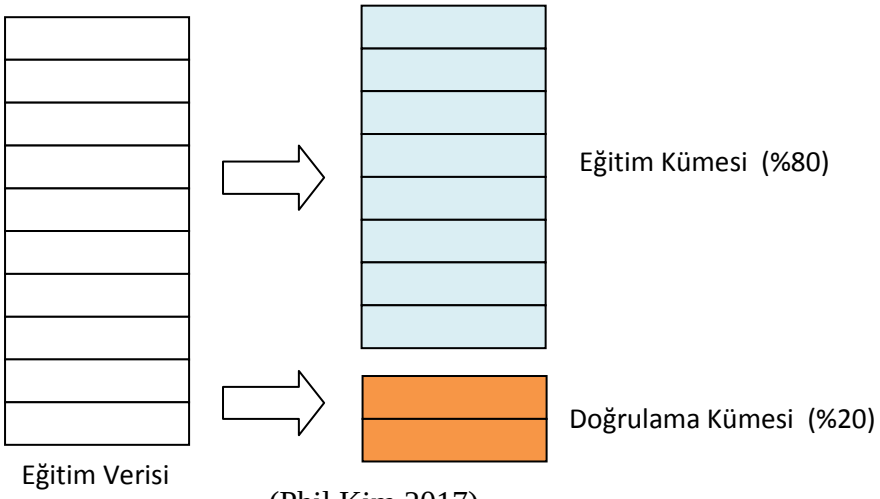
Öğrenme Felsefesi, Günümüzde Makine Öğrenmesi ve Temel Kavramları, Değerlendirme Metrikleri (Ölçütleri) III

Ezberleme probleminin çözümü için önerilen en basit yöntem: Veri Kümesinin Bölümlenmesi

Veri kümesi, eğitim kümesi (Training set) ve doğrulama kümesi (Validation set) olarak iki kısma ayrılır. Doğrulama kümesinin, veri kümesinden rastgele veriler ile oluşturulması ezberleme sorunun tespiti için daha etkilidir.

* Eğitim kümesi ile makine öğrenme algoritması eğitimi gerçekleştirir ve öğrenme modelini optimize eder.

* Doğrulama kümesi ile eğitimi bitmiş modelin performansı ölçülür. Eğer ezberleme gerçekleşmişe, eğitim kümesi performansı çok iyi olmasına rağmen doğrulama kümesi ile yapılan teste performans düşer. Bu durumda, makine öğrenmesi algoritması konfigürasyonu yeniden ayarlanır ve tekrar eğitim yapılır. Eğitim sonunda eğitim kümesi performansı ve doğrulama kümesi performansı birlikte yeterince iyi elde edilirse eğitim sonlandırılır. Genelde bir eğitim verisinin 8/10 eğitim kümesi ve 2/10 doğrulama kümesi olarak kullanılması tavsiye edilir. (Phil Kim,2017)



(Phil Kim,2017)

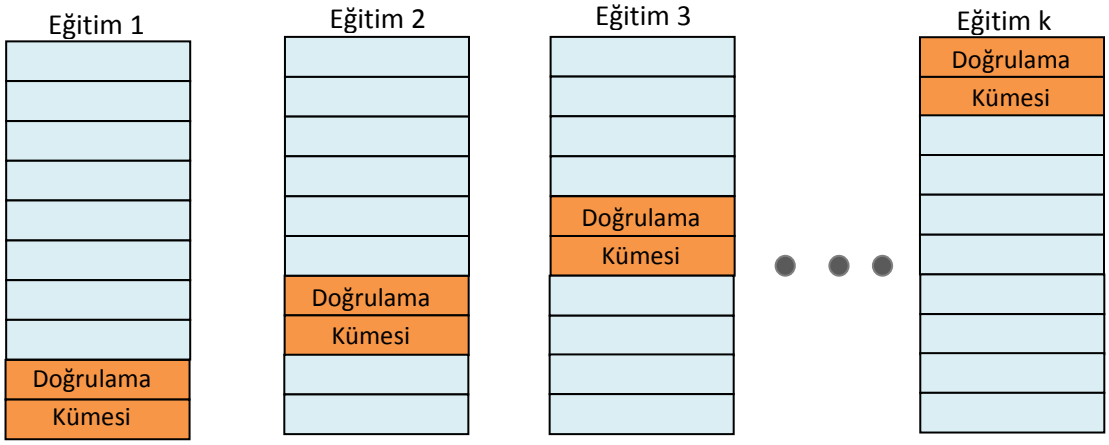
Eğitim-doğrulama süreci ile performans ayarlaması şu algoritma ile ifade edilmiştir (Phil Kim,2017).

Adım 1: Eğitim verisini 8:2 oranında eğitim kümesi ve doğrulama kümesine bölünüz.

Adım 2: Eğitim kümesi yardımı ile makine öğrenmesi modelini eğitiniz.

Adım 3: Eğitilmiş modelin, doğrulama kümesi ile performansını test ediniz. Eğer hem eğitim hem de doğrulama kümesi performansları yeterli ise eğitimi sonlandırınız. Eğer doğrulama kümesi için sistemin performansı yetersiz ise makine öğrenmesi sistemini yeniden ayarlayınız (Yapay sinir ağları için, ağın konfigürasyonu ayarlanabilir) ve tekrar Adım 2'ye dönünüz.

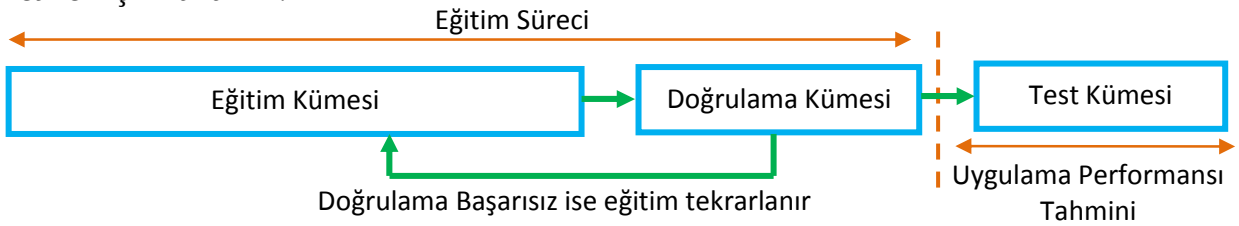
Çapraz Doğrulama Yöntemi (Cross Validation): Eğitim-doğrulama sürecinin, doğrulama kümesi değiştirilerek tekrarlamasıdır. Böylece bütün veri kümesi için doğrulama yapılabilir ve doğrulama performansı garanti edilmiş olur. Çapraz doğrulama, bir makine öğrenmesi algoritmasının öğrenme kabiliyetinin bir veri seti için ne kadar yeterli olduğunu görmek için bir test ortamı sağlar. k-defa eğitim ve doğrulama performans sonuçlarında öğrenme performansı yeterince iyi görülürse bu algoritma bu veri setinin öğrenilmesi için yeterlidir sonucuna varılabilir. Buna k-kat (k-fold) çapraz doğrulama adı verilir. Farklı makine öğrenmesi algoritmalarının bir veri kümesi üstünde başarımlarının karşılaştırılması için çapraz doğrulama tekniği yaygın olarak kullanılır. Çoklu test (k adet test) performans analizi için istatistiki veri sağlayabilir. k-kat doğrulamada veri kümeleri bütün eğitim kümesini tarayacak şekilde bölgesel seçilir. Bu nedenle doğrulama kümesi oranı (1/k) ve eğitim kümesi oranı (1-k)/k olur.



(Phil Kim,2017)

Ancak, öğrenen sistemin uygulamada ortaya koyabileceği performansı hakkında fikir sahibi olabilmek için eğitim aşamasında hiç bulunmamış (eğitimden bağımsız) veriler ile test edilmesi gerekmektedir. Çünkü uygulamada(sahada) eğitim seti dışında veriler üzerinde çalışması olağandır. Bu nedenle eğitim verileri günümüzde üç kümeye ayrılırlar: Eğitim kümesi, doğrulama kümesi, test kümesi. Eğitim ve doğrulama kümesi tekrarlı eğitimler nedeni ile eğitim sürecine karışırlar. Burada farklı olarak, test kümesi eğitime hiçbir şekilde katılmayan, eğitim-doğrulama süreçleri tamamlandıktan sonra öğrenen sistemin eğitimden bağımsız veriler ile test edilmesini sağlayan veri kümesidir. Dolayısı ile öğrenen sistemin gerçek performansını ölçmek için daha uygundur.

Eğitilmiş makine öğrenmesi algoritmasının gerçek(saha, uygulama) performansını tahmin etmek için eğitim verileri üç alt kümeye ayrılır. Eğitim ve doğrulama kümesi ezberleme sorunu çözümü ve iyi bir genelme için eğitim kalitesini iyileştirmek için kullanılır. Eğitim sürecinde hiç kullanılmayan bir test kümesi ise öğrenen modelin saha performansını tahmin etmek için kullanılır.

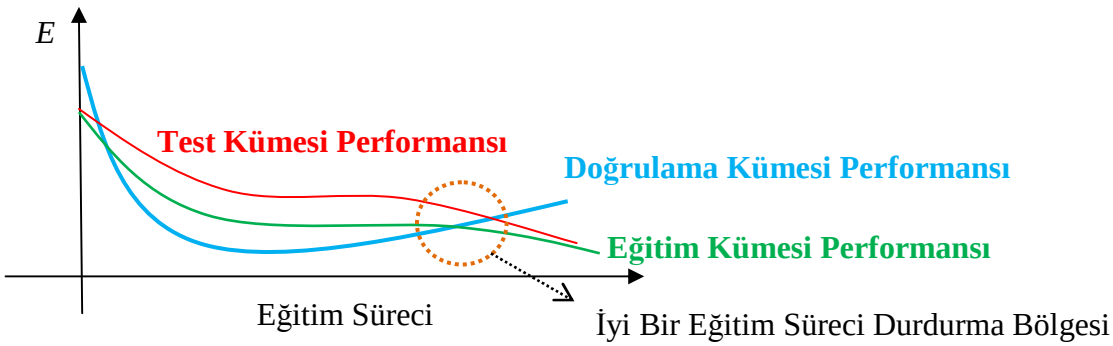


Eğitim Kümesi: Makine öğrenmesi algoritması bu kümeyi modelin eğitiminde kullanır. (Veri kümesinin %70'ı oranında olabilir.)

Doğrulama Kümesi: Eğitilmiş makine öğrenmesi algoritmasının performansının doğrulaması için kullanılır. Doğrulama kümesi performansı iyileşene kadar eğitim kümesi ile eğitim tekrar eder. Her yeni eğitim döngüsünde eğitim süreci ayarları yapılır. (Veri kümesinin %15'ı oranında olabilir.)

Test Kümesi: Eğitimsen bağımsız veri (eğitime girmeyen veri) ile makine öğrenmesi algoritmasının uygulamada performansını tahmin için ayrılan veridir. Hiçbir zaman eğitim kümesinde veya doğrulama kümesinde kullanılmaz. (Veri kümesinin %15'ı oranında olabilir.)

Ayrıca, bazı makine öğrenmesi araçları eğitim, doğrulama ve test kümesi performanslarını eğitimi otomatik durdurma için kullanır. İyi bir eğitim performansı için eğitim süreci aşağıdaki grafikteki işaretlenen bölgede durdurularak elde edilebilir. Grafikte, eğitim kümesi, doğrulama kümesi ve test kümesi performansları yeterince küçük ve birbirine yakın olduğu zaman eğitim durdurulur.



Aşırı öğrenme(ezberleme) probleminin çözümü için önerilmiş pek çok yaklaşım vardır. Bunlar eğitimin erken durdurulması, sıyırma(drop-out), düzleştirme (regularization) gibi. Bu yöntemlerin yaygın kullanılanlarına yeri geldikçe değinilecektir.

Örnek: T={x1,x2,x3,x4,x5,x6,x7,x8,x9,x10} veri seti için 5-kat çapraz doğrulama için örnek eğitim ve doğrulama setleri oluşturunuz. Eğitim kümesi ve doğrulama kümesi ortalama performansını ifade ediniz. (Pei: i. eğitim için eğitim kümesi performansı. Pdi: i. eğitim için doğrulama kümesi performansı.

Genelde 5 kat doğrulama 5 test sonucunda bütün veriler için doğrulama gerektirir. Bu nedenle doğrulama kümesinde 10/5=2 veri bulunmalı. Kalanlar 10-2=8 adet ise eğitim kümesine ayrılmalıdır. (k-kat doğrulamada doğrulama kümesi oranı (1/k=1/5=0.2 olur. Buda 2 veri olması demektir.)

Doğrulama kümesi Ds ve eğitim seti Es ile gösterilsin. Rastgele örnekleme kuralı ile kümeleri oluşturalım. Bunun için Ds kümelerini rastgele seçmemiz yeterli olur. Kalan veriler Es eğitim kümesini oluşturur.

1. doğrulama için: Ds1={x1,x9}, Es1=T-Ds1={x2,x3,x4,x5,x6,x7,x8,x10} bu eğitim için performanslar eğitim performansı Pe1 ve doğrulama verisi için performans Pd1 olsun.

2. doğrulama için: Ds2={x5,x7}, Es2=T-Ds2={x1,x2,x3,x4,x6,x8,x9,x10} bu eğitim için performanslar eğitim performansı Pe2 ve doğrulama verisi için performans Pd2 olsun.

3. doğrulama için: Ds3={x3,x10}, Es3= T-Ds3={{x1,x2,x4,x5,x6,x7,x8,x9} bu eğitim için performanslar eğitim performansı Pe3 ve doğrulama verisi için performans Pd3 olsun.

4. doğrulama için: Ds4={x4,x6}, Es4= T-Ds4={{x1,x2,x3,x5,x7,x8,x9,x10} bu eğitim için performanslar eğitim performansı Pe4 ve doğrulama verisi için performans Pd4 olsun.

5. doğrulama için: Ds5={x2,x8}, Es5= T-Ds5={{x1,x3,x4,x5,x6,x7,x9,x10} bu eğitim için performanslar eğitim performansı Pe5 ve doğrulama verisi için performans Pd5 olsun.

Bu 5 test ile makine öğrenmesi algoritması 5 farklı eğitim ile bütün veri kümesi üzerinde doğrulama performansı ölçülür. Bu ölçümler istatistiki olarak değerlendirilerek (ortalama, standart sapma, maksimum ve minimum değer) algoritmanın bu veri seti üzerinde çoklu test performansları gösterilebilir.

Ortalama eğitim performansı Pe_ort=1/5*(Pe1+ Pe2+ Pe3+ Pe4+ Pe5)

Ortalama doğrulama performansı Pd_ort=1/5*(Pd1+ Pd2+ Pd3+ Pd4+ Pd5)

Pe_ort ve Pd_ort birbirine yakın değerde ve yeterince iyi çıkması ise bu makine öğrenmesi algoritmasının ve modelinin bu verinin eğitimi için yeterince uygun olduğunu ifade eder.

Makine Öğrenmesi İle Çözülen Sınıflama Problemlerinde Performans Değerlendirmesi

Karışıklık Matrisi (Confusion Matrix):

Karışıklık matrisi sınıflama işleminde, veri kümesi üzerinde sınıflamanın doğru ve yanlış tahmin sayılarının verildiği tablodur. Örneğin ikili sınıflama işleminde durumlar pozitif (sınıfa dahil) ve negatif (sınıfa dahil değil) olsun. Bu iki sınıf üzerinde doğru ve yanlış sınıflama 4 farklı durum oluşturur. (1) Pozitif örnekler doğru sınıflanabilir (Doğru pozitif), (2) pozitif örneklerin yanlış sınıflanması olabilir (Yanlış pozitif), (3) Negatif örneklerin doğru sınıflanması (Doğru negatif) ve (4) negatif örneklerin yanlış sınıflanması (Yanlış negatif). Aşağıda bu 4 farklı sınıflama performansını ifade eden karışıklık matrisi gösterilmiştir.

		Gerçek Sınıflar (Örneklerin Doğru Etiketi)	
		Pozitif (P)	Negatif (N)
Makine öğrenmesi algoritması sınıflaması (tahmin etiketi)	Pozitif	**Doğru Pozitif (True Positive) Sayısı (TP)**	**Yanlış Pozitif (False Positive) Sayısı (FP)**
	Negatif	**Yanlış Negatif (False Negative) Sayısı (FN)**	**Doğru Negatif (True Negative) Sayısı (TN)**

* Pozitif örnek sayısı P olsun. P=TP+FN yazılabilir.

* Negatif örnek sayısı N olsun. N=TN+FP yazılabilir.

* Toplam örnek sayısı için P+N= TP+ FN + TN+ FP

Makine Öğrenmesi Algoritmalarının Sınıflama Performansını Değerlendirme Metrikleri(Ölçütleri):

Yukarıda verilen karışık matrisi değerlerine göre sınıflama işlemi performans değerlendirmesi için tanımlanmış bazı önemli metrikler aşağıdaki tabloda özetlenmiştir.

Değerlendirme Metrik	Hesaplama Formülü	Tanımı
Doğruluk (accurarcy)	$\frac{TP+TN}{FP+FN+TP+TN}$	Algoritmanın doğru olarak tahmin oranı. (Bütün örnekler içinde doğruluk pozitif ve negatif tahminlerinin oranı)
Hassasiyet (Precision)	$\frac{TP}{TP+FP}$	Algoritmanın pozitif tahminleri içinde doğru pozitif tahmini oranı. (pozitif tahminlerinin içinde doğruluk oranı)
Duyarlılık (Sensitivity) (Recall (anma), Doğru Pozitif Oranı (True Positive Rate- TPR) denir)	$\frac{TP}{TP+FN}$	Algoritmanın pozitif örnekler içinde doğru pozitif tahmin oranı. (Pozitif sınıfa örnekler içinde doğru pozitif tahmin oranı)
Özgünlük (Specificity)	$\frac{TN}{TN+FP}$	Algoritmanın negatif örnekler içinde doğru negatif tahmin oranı.
F1-Skor	$2\frac{Hassasiyet \times Duyarlıuya}{Hassasiyet + Duyarlıuya}$	Hassasiyet ve duyarlılığı birlikte değerlendirmek için önerilen bir skor.
Yanlış Pozitif Oranı (False Pozitif Rate- FPR) denir	$1-\frac{TN}{TN+FP}=\frac{FP}{TN+FP}$	(1- Özgünlük) ile hesaplanır. ROC eğrisi çiziminde kullanılır.

Örnek: Bir medikal makine öğrenmesi uygulamasında 100 adet pozitif ve 90 adet negatif örnek üzerinde yapılan test sonucunda doğru ve yanlış sınıflama oranları şöyle elde edilmiştir. TP=90, TN=85 elde edilmiştir.

a) Bu bilgiler ile karmaşıklık matrisini oluşturunuz.

b) Doğruluk, hasiyet, duyarlılık ve özgünlük, F1 skor değerlerini hesaplayınız.

c) Bu test için ROC eğrisine üzerine konacak noktanın koordinatlarını bulunuz.

a) Test deki toplam pozitif örnek sayısı P=100

P=TP+FN => FN=P-TP=100-90=10 adet.

Test deki toplam negatif örnek sayısı N=90

N=TN+FP => FP=N-TN=90-85=5 adet.

		Gerçek Sınıf (Doğru Etiket)	
		Pozitif (P=100)	Negatif (N=90)
Tahmin edilen sınıf (sınıflanan etiket)	Pozitif	**TP=90**	**FP=5**
	Negatif	**FN=10**	**TN=85**

b) Doğruluk= $\frac{TP+TN}{FP+FN+TP+TN}=\frac{90+85}{10+5+90+85}=0.92$

Hassasiyet= $\frac{TP}{TP+FP}=\frac{90}{90+5}=0.94$

Duyarlılık (TPR, recall)= $\frac{TP}{TP+FN}=\frac{90}{90+10}=0.90$

Özgünlük= $\frac{TN}{TN+FP}=\frac{85}{85+5}=0.94$

F1 skore= $2\frac{Hassasiyet\times Duyarlıuya}{Hassasiyet+Duyarlıuya}=2\frac{0.94*0.90}{0.94+0.90}=0.91$

Makine Öğrenmesi Algoritmalarının Regresyon Performansını Değerlendirme Metrikleri(Ölçütleri):

Aşağıdaki tabloda regresyon problemi çözümü için elde edilen modelin performansını değerlendirmek için sık kullanılan bazı performans ölçütleri tanıtılmıştır.

Burada p adet veriden oluşan veri kümesi $T=\{(x_1,d_1),(x_2,d_2),(x_3,d_3),...,(x_p,d_p)\}$ ile temsil edilsin. Burada i. giriş verisi x_i için modelin çıkışı y_i ile gösterilsin. Öğrenen model $y_i=f(x_i)$ ile gösterilsin. Bu giriş için istenen değer(etiket) d_i olsun.

Değerlendirme Metrik	Hesaplama Formülü	Tanımı
Ortalama Karesel Hata (OKH) (Mean Square Error)	$OKH = \frac{1}{p}\sum_{i=1}^{p}(d_i - y_i)^2$	OKH hatanın karesini dikkate aldığı için 1'den büyük hataları güçlendirir 1'den küçük hataları zayıflatır. Dolayısı ile 1'den büyük hatalara daha fazla duyarlıdır. Optimizasyon sürecinin küçük hataları daha az önemsemesine yol açabilir.
Ortalama Mutlak Hata (OMH) (Mean Absolute Error)	$OMH = \frac{1}{p}\sum_{i=1}^{p}\lvert d_i - y_i\rvert$	OMPH hatanın büyüklüğünü (genliğini), dikkate alır. Dolaysı ile hataları kendi değerleri ile değerlendirmek için uygundur.
Ortalama Mutlak Yüzde Hata (OMYH) (Mean Absolute Percentage Error)	$OMYH = \frac{\%100}{p}\sum_{i=1}^{p}\frac{\lvert d_i - y_i\rvert}{\lvert d_i\rvert}$	OMYH hatanın büyüklüğünü (genliğini) gerçek veri değerine oranla değerlendirir. Hata yüzde (%) olarak verilir. Böylece, verinin değer büyüklüğünün hataya ölçümüne etkisini kompanze eder. Farklı büyüklük değerlerine sahip veri kümelerinin hatalarını karşılaştırmada kullanılabilir.
R^2-skore (Coefficient of determination)	$R^2 = 1 - \sum_{i=1}^{p}\frac{(d_i - y_i)^2}{(d_i - y_{ort})^2}$, $y_{ort} = \frac{1}{p}\sum_{i=1}^{p} d_i$	Modelin verilere ne ölçüde uyduğunu ifade eder. Modelin geçerliliği konusunda fikir verir. Model çıkışı ile istenen değer bire bir aynı ise 1 sonucunu verir. Bu %100 uyum ifade eder. Değeri sıfıra doğru azatlıkça modelin veriler ile uyumluluğu artar.

Örnek: (x_i, d_i) formatında verilen $T = \{(1,0),(2,1),(3,4)\}$ eğitim kümesi ile eğitim sonucunda $y_i = x_i - 1$ regresyon modeli elde eden bir algoritma için ortalama karesel hata, ortalama mutlak hata, ortalama mutlak yüzde hata ve R^2-skore değerini hesaplayınız.

Bu soruyu daha kolay çözebilmek için tablo oluşturalım.

x_i	d_i	$y_i = x_i - 1$	Veri Başına Hata: $d_i - y_i$	Veri Başına Karesel Hata $(d_i - y_i)^2$	Veri Başına Mutlak Hata $\lvert y_i - d_i \rvert$	Veri Başına Mutlak Yüzde Hata $\frac{\lvert d_i - y_i \rvert}{\lvert d_i \rvert}$
1	0	0	0	0	0	Sıfıra bölüm var. Hesaba katılmaz
2	1	1	0	0	0	0
3	4	2	2	4	2	0.5

Eğitim kümesinde veri sayısı $p = 3$

$$OKH = \frac{1}{p}\sum_{i=1}^{p}(d_i - y_i)^2 = \frac{1}{3}\left((0-0)^2 + (1-1)^2 + (4-2)^2\right) = \frac{1}{3}(0+0+4) = \frac{4}{3}$$

$$OMH = \frac{1}{p}\sum_{i=1}^{p}\lvert d_i - y_i \rvert == \frac{1}{3}(\lvert 0-0 \rvert + \lvert 1-1 \rvert + \lvert 4-2 \rvert) = \frac{1}{3}(0+0+2) = \frac{2}{3}$$

OMYH hesaplamasında ilk veride sıfıra bölüm hatası var. Çünkü $\frac{\lvert d_i - y_i \rvert}{\lvert d_i \rvert} = \frac{\lvert 0-0 \rvert}{\lvert 0 \rvert}$. Bu veri için hesaplanamaz. Bu nedenle bu veri OMYH hesaplamasında dikkate alınmadı. ($p = 2$)

$$OMYH = \frac{\%100}{p}\sum_{i=1}^{p}\frac{\lvert d_i - y_i \rvert}{\lvert d_i \rvert} = \frac{1}{2}\left(\frac{\lvert 1-1 \rvert}{\lvert 1 \rvert} + \frac{\lvert 4-2 \rvert}{\lvert 4 \rvert}\right) = \frac{1}{2}(0+0.5) = 0.25$$

R^2-skore hesabı

$$y_{ort} = \frac{1}{p}\sum_{i=1}^{p} d_i = \frac{1}{3}(0+1+4) = \frac{5}{3}$$

$$R^2 = 1 - \sum_{i=1}^{p}\frac{(d_i - y_i)^2}{(d_i - y_{ort})^2} = 1 - \left(\frac{(0-0_i)^2}{\left(0-\frac{5}{3}\right)^2} + \frac{(1-1_i)^2}{\left(1-\frac{5}{3}\right)^2} + \frac{(4-2_i)^2}{\left(4-\frac{5}{3}\right)^2}\right) = 1 - \left(0+0+\frac{4}{(2.33)^2}\right) = 1 - 0.73 = 0.27$$