

Sınıflandırma ve Kümeleme Yöntemlerinin Değerlendirilmesi

Bu sunum, veri madenciliğinde kullanılan sınıflandırma ve kümeleme yöntemlerinin değerlendirilmesini ele alacak. Performans metrikleri, karmaşıklık matrisi ve diğer önemli konular hakkında bilgi vereceğiz.



by Alper Talha Karadeniz

Performans Metrikleri Nedir?

Doğruluk (Accuracy)

Bir modelin, tüm sınıflandırma denemeleri içinde **doğru** tahmin ettiği örneklerin oranını

verir.**Avantaj:** Kolay anlaşılır, genel performans hakkında hızlı bir fikir verir.

Dezavantaj: Dengesiz veri kümelerinde (örneğin, pozitif örnek çok az olduğunda) yanıltıcı olabilir.

Duyarlılık / Kesinlik (Precision)

Modelin **pozitif** tahmin yaptığı durumlarda, bu tahminlerin ne kadarının **gerçekte** doğru (pozitif) olduğunu gösterir. Precision yüksekse, model pozitif sınıfı tahmin ederken **çoğunlukla haklı** çıkıyor (yanlış pozitif oranı düşük).

Spam filtresi örneği: “Spam” dediği e-postaların gerçekten spam olma oranını yansıtır

Geri Çağırma (Recall- Sensitivity)

Gerçek pozitiflerin içinden, modelin **pozitif** olarak yakalayabildiği (bulabildiği) paydır. Recall yüksekse, gerçekte pozitif olanları **kaçırmıyor** (FN düşük).

Kanser tespiti örneği: Gerçekten hasta olanların ne kadarını “hasta” olarak sınıflandırdığını gösterir.

F1 Puanı (F1 Score)

Precision ve **Recall** arasında **harmonik ortalama** alan, ikisini tek bir metriğe birleştiren bir ölçüt. Özellikle **dengesiz veri setlerinde** yaygın kullanılır. Precision ve Recall değerlerinin **ikisi de** yüksek olduğunda F1 da yüksek olur.

Dengesiz sınıflarda (ör. az sayıda pozitif örnek), Accuracy yanıltıcı olabilir; F1 daha dengeli bir gösterge sağlar.

Specificity (Özgüllük)

Gerçek **negatif** olanların içinden, modelin **negatif** olarak doğru sınıflandırdığı paydır. Bazen $1 - \text{False Positive Rate}$ ($1 - \text{FPR}$) olarak geçer. Negatif örneklerin ne kadarının doğru şekilde negatif bulunabildiğini gösterir.

“Yanlış alarm” (FP) hassasiyetine dikkat eden uygulamalarda önemlidir.

Karmaşıklık Matrisi (Confusion Matrix)

Karmaşıklık matrisi, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir tablodur. Gerçek pozitifler, negatifler, yanlış pozitifler ve yanlış negatifler gibi farklı hata türlerini gösterir.

| | | GERÇEK | |
|--------|---------|---------|---------|
| | | Pozitif | Negatif |
| TAHMİN | Pozitif | TP | FP |
| | Negatif | FN | TN |

| | |
|---------------------------|--|
| TP (True Pozitif) | Hem gerçekte pozitif olan hem de modelin pozitif olarak tahmin ettiği değerlerdir. |
| TN (True Negatif) | Hem gerçekte negatif olan hem de modelin negatif olarak tahmin ettiği değerlerdir. |
| FP (False Pozitif) | Gerçekte negatif olan ama modelin pozitif olarak tahmin ettiği değerlerdir. |
| FN (False Negatif) | Gerçekte pozitif olan ama makinenin negatif olarak tahmin ettiği değerlerdir. |

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity (Se)} = \frac{TP}{TP + FN}$$

$$\text{Specificity (Sp)} = \frac{TN}{TN + FP}$$

$$\text{Precision (Prec)} = \frac{TP}{TP + FP}$$

$$\text{Fscore (F - Sc)} = \frac{2TP}{2TP + FP + FN}$$

Sınıflandırma Başarısının Değerlendirilmesi

1

ROC Eğrisi

Farklı eşik değerleri için doğruluk ve geri çağırma değerlerini gösterir.

2

AUC Değeri

ROC eğrisinin altındaki alanı gösterir, yüksek AUC değeri daha iyi performansı gösterir.

3

Karmaşıklık Matrisi Analizi

Farklı sınıflar için doğruluk, hassasiyet ve geri çağırma değerlerini gösterir.

| | Frenquette | Bilecik | Chandler | Fernette | Fernor | Hardley | Howard | Kaman1 | Kaplan86 | Lara | Maya1 | Mithland | Oguzlar77 | Pedro | Sebin | Sen | Serr | Yalova3 |
|------------|------------|---------|----------|----------|--------|---------|--------|--------|----------|------|-------|----------|-----------|-------|-------|-----|------|---------|
| Frenquette | 457 | 0 | 4 | 2 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 15 | 0 | 0 | 0 | 3 | 4 | 0 |
| Bilecik | 2 | 355 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 6 | 3 | 7 |
| Chandler | 1 | 0 | 307 | 2 | 2 | 1 | 0 | 3 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Fernette | 4 | 1 | 1 | 324 | 6 | 1 | 0 | 1 | 2 | 6 | 0 | 1 | 0 | 2 | 0 | 0 | 2 | 0 |
| Fernor | 1 | 0 | 3 | 8 | 366 | 0 | 7 | 8 | 1 | 7 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 1 |
| Hardley | 0 | 6 | 10 | 1 | 1 | 342 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 7 | 2 | 2 | 1 | 1 |
| Howard | 0 | 1 | 1 | 0 | 12 | 0 | 277 | 19 | 4 | 13 | 3 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Kaman1 | 0 | 0 | 2 | 6 | 7 | 1 | 7 | 350 | 7 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Kaplan86 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 3 | 173 | 2 | 147 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Lara | 0 | 0 | 4 | 9 | 8 | 0 | 5 | 5 | 5 | 202 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| Maya1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 1 | 169 | 0 | 117 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Mithland | 10 | 0 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 393 | 0 | 0 | 0 | 2 | 9 | 0 |
| Oguzlar77 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 211 | 0 | 3 | 7 | 1 | 5 |
| Pedro | 0 | 2 | 3 | 3 | 2 | 0 | 1 | 0 | 0 | 1 | 6 | 1 | 0 | 280 | 1 | 2 | 0 | 1 |
| Sebin | 0 | 1 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 2 | 323 | 4 | 3 | 1 |
| Sen | 1 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 505 | 0 | 9 |
| Serr | 2 | 10 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 24 | 0 | 1 | 2 | 2 | 394 | 3 |
| Yalova3 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 14 | 6 | 399 |

Confusion Matrix of Quadratic SVM

| | Tahmin: Pozitif | Tahmin: Negatif |
|-----------------|-----------------|-----------------|
| Gerçek: Pozitif | TP = 50 | FN = 5 |
| Gerçek: Negatif | FP = 10 | TN = 35 |

- TP (True Positive) = 50
- FN (False Negative) = 5
- FP (False Positive) = 10
- TN (True Negative) = 35

Bu matrisin satırlarında verinin **gerçek etiketi**, sütunlarında ise **modelin tahmini** bulunur.

$$\text{Toplam örnek sayısı} = TP + FN + FP + TN = 50 + 5 + 10 + 35 = 100$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{50 + 35}{50 + 35 + 10 + 5} = \frac{85}{100} = 0.85$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{50}{50 + 10} = \frac{50}{60} = 0.8\overline{3}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{50}{50 + 5} = \frac{50}{55} = 0.90\overline{90}$$

Kullanım Alanları ve Örnekler



Müşteri Segmentasyonu

Müşterileri benzer özelliklere göre gruplandırmak için kullanılır.



Dolandırıcılık Tespiti

Dolandırıcılık şüpheli işlemleri tespit etmek için kullanılır.



Tıbbi Teşhis

Hastalıkları teşhis etmek için kullanılır.



Görüntü Tanıma

Görüntüleri sınıflandırmak için kullanılır.

- **Accuracy:** Tüm örneklerin içinde kaçını doğru bildiğinizi gösterir.
- **Precision:** Pozitif tahminlerinizin ne kadarının gerçekten pozitif olduğu.
- **Recall:** Gerçek pozitiflerin ne kadarını doğru yakaladığınız.
- **Specificity:** Gerçek negatiflerin ne kadarını doğru bulduğunuz.
- **F1-Score:** Precision ve Recall'un dengeli bir şekilde harmanlanmış hâli.
- **ROC & AUC:** Değişik karar eşikleri altında modelin TPR-FPR dengesini ve genel ayırt edici gücünü gösterir.