

**T.C.**

**FENERBAHÇE ÜNİVERSİTESİ**

**MÜHENDİSLİK VE MİMARLIK FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**CROSS VALIDATION**

**MAKALE RAPORU**

**Hazırlayan**

**Mehmet Çolak 190301022**

**ÖZET**

Çapraz doğrulama, modellerin gerçek tahmin hatasını tahmin etmek ve model parametrelerini ayarlamak için en yaygın kullanılan veri yeniden örnekleme yöntemlerinden biridir. Bu makale en yaygın çapraz doğrulama türlerine ve bunların ilgili veri yeniden örnekleme yöntemlerine bir giriş sağlamaktadır.

Anahtar kelimeler: çapraz doğrulama, model

**İÇİNDEKİLER**

[**Şekil Listesi 2**](#_Toc123582372)

[**1. Giriş 3**](#_Toc123582373)

[**2. Temel Kavram ve Gösterimler 4**](#_Toc123582374)

[**2.1 Tek bekletmeli rasgele alt örnekleme (Single hold-out random subsampling) 5**](#_Toc123582375)

[**2.2 k-katlı rasgele alt örnekleme 5**](#_Toc123582376)

[**2.3 k katlı çapraz doğrulama 6**](#_Toc123582377)

[**3. Sonuç 6**](#_Toc123582378)

[**KAYNAKLAR 7**](#_Toc123582379)

# Şekil Listesi

[Şekil 1: En Küçük Kareler 3](file:///C:\Users\colak\Desktop\MEHMET_COLAK-MAKINE_OGRENMESI_RAPORU.docx#_Toc123582311)

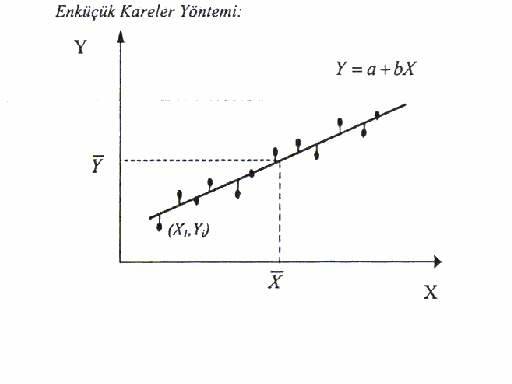
[Şekil 2: 10 katlı çapraz doğrulama Veri seti, her biri (yaklaşık olarak) verinin %10'unu içeren on ayrık alt kümeye rastgele bölünür. Model, eğitim seti üzerinde eğitilir ve daha sonra doğrulama setine uygulanır. 6](file:///C:\Users\colak\Desktop\MEHMET_COLAK-MAKINE_OGRENMESI_RAPORU.docx#_Toc123582312)

# Giriş

Cross validation, öngörü modellerinin genelleme yeteneğini değerlendirmek ve ezberi önlemek için, bir veri yeniden örnekleme yöntemidir. Bu rapor, çapraz doğrulamaya ve bununla ilgili yeniden örnekleme yöntemlerine giriş yapan makaleyi özetlemek için hazırlanmıştır.

N etiketli örneklerden (veya veriler), xi, i = 1 ... n'den oluşan bir veri seti D'yi düşünün. Her veri bir dizi özellik (veya öznitelik) tarafından tanımlanır. Her xi verisinin herhangi bir yi sınıfına ait olduğunu varsayıyoruz. Biyoinformatikten tipik bir örnek, DNA mikro dizi verilerine dayanan bir gen ekspresyonu veri setidir, burada her bir öznitelik bir gen ifade profili tarafından tarif edilen etiketli bir tümör örneğini temsil eder. Yaygın zorluklardan biri, ifade profillerine göre yeni, görünmeyen tümör örnekleri sınıfını güvenilir bir şekilde tahmin edebilecek sınıflandırıcının gelişimi ile ilgilidir. Kavramsal olarak, öngörü bir modeli olan ƒ (), D veri kümesine dayalı bir duruma ait sınıf etiketi atamak için bir kuraldır, yani ƒ (x, D) = ýi, burada ý, durum xi için öngörülen sınıf etiketidir. Makine öğrenmesinde, böyle bir modelin yapımı denetimli öğrenme olarak söylenir. Denetimli öğrenmede merkezi bir soru, ortaya çıkan modelin doğruluğu ile ilgilidir. Burada ana problem ezberlemedir. Eldeki veri kümesine mükemmel şekilde uyarlanmış bir model oluşturmak çok kolaydır, ancak daha sonra gelecek yeni veriler için bu durumu genelleştirmek mümkün olmayacaktır. Örneğin, bağımsız değişken y’yi, i = 1..n ile n gözlemlerine (xi, yi) dayanan bağımsız değişkenden tahmin etmek istediğimiz tek değişkenli bir regresyon probleminiz olduğunu düşünün. Bir eğriyi bu noktalara mükemmel bir şekilde uydurmak için n - 1 derecesinin bir polinomunu kullanabilir ve daha sonra yeni bir değer olan xi+1 için yi+1 değerini tahmin etmek için eğriyi kullanabiliriz. Ancak, bu eğrinin eldeki verileri ezberlemesi muhtemeldir. Sadece bağımlı ve bağımsız değişken arasındaki ilişkiyi yansıtmakla kalmaz, aynı zamanda veri kümesindeki doğal gürültüyü de modeller.

Öte yandan, en küçük kareler (Least squares) yöntemi gibi daha basit bir model, doğal gürültüden daha az etkilenir, ancak değişkenler arasındaki ilişkiyi iyi yakalayamayabilir.

Böyle bir modelin yeterince beslenmediği söylenir. Ne ezberleyen ne de yeterince beslenmeyen modelin iyi genelleme yapması beklenmez. Buradaki en büyük zorluk ezberlemeyle eksik beslenme arasındaki dengeyi bulabilmektir.

Şekil : En Küçük Kareler

Bir modelin genelleme yeteneğini nasıl değerlendirebiliriz? İdeal olarak, modeli, modeli oluşturmak için kullandığımız verilerle aynı popülasyonda bulunan kaynaklardan yeni veriler besleyerek değerlendiririz. Uygulamada, “yeni bağımsız” doğrulama çalışmaları genellikle mümkün değildir. Ayrıca, harici bir doğrulama için zaman ve diğer kaynaklara yatırım yapmadan önce, öngörücü performansını tahmin etmeniz önerilir. Bu, genellikle çapraz doğrulama gibi veri yeniden örnekleme yöntemleriyle yapılır. Bu rapor, çapraz doğrulamanın birkaç ana alt tipini ve bunların ilgili yeniden örnekleme yöntemleri içermektedir.

# Temel Kavram ve Gösterimler

Bir öngörü modeli oluşturmak ve değerlendirmek için mevcut olan veri setinde, öğrenme seti Değitim olarak adlandırılır. Bu veri setinin ilgi çekici bir nüfustan bir örnek olduğu varsayılmaktadır. Öğrenme seti için Değitim ve test seti için Dtest rastgele alt örnekleme yöntemleri kullanılarak oluşturulur. Model daha sonra eğitim setleri kullanılarak oluşturulur (veya eğitilir) ve test kümelerinde test edilir. Çeşitli rastgele alt örnekleme yöntemleri, eğitim ve test setlerinin nasıl üretildiğine göre farklılık gösterir. “Eğitim” teriminin, öğrenme algoritmasını verilerin bir alt kümesine uyguladığımızı ima ettiğini unutmayın. Ortaya çıkan model, ƒ (x, Değitim), aynı öğrenme işlevini tüm öğrenme setine uygulanmasından sonrasında elde edilecek modelin sadece bir tahminidir, ƒ (x, Döğrenim). Tekrarlanan alt örneklemeye dayanan model değerlendirmesi, birkaç veri alt kümesine bir öğrenme fonksiyonunun uygulandığı ve sonuçta ortaya çıkan modellerin, daha sonra eğitim sırasında kullanılmayan diğer alt kümelerde (yani test setleri veya doğrulama setleri) değerlendirildiği anlamına gelir. Modellerin bu alt kümelerde elde ettiği performansın ortalaması, nihai model ƒ (x, Döğrenim) performansının bir tahminidir. Her durumda sadece yi etiketin ilişkilendirildiğini varsayalım. Sınıflandırma durumunda, yi ayrı bir sınıf etiketidir. Sınıflandırıcı, ayrı bir sınıf etiketi atayan tahmine dayalı bir modelin özel bir durumudur. Regresyon görevlerinde, hedef genellikle gerçek bir değerdir, yi ∈ R. Tahmine dayalı bir model ƒ (), xi durumunun hedef yi 'sini ƒ (xi) = ýi olarak tahmin eder. Kayıp fonksiyonu, £(yi, ýi), tahmin hatasını nicelleştirir. Örneğin, bir sınıflandırmada 0-1 kayıp fonksiyonunu kullanarak, yi ≠ ýi ise kayıp 1, aksi halde 0'dır. Kayıp fonksiyonu ile artık iki farklı hatayı hesaplayabiliriz, (1) eğitim hatası ve (2) test hatası. Eğitim hatası (veya yeniden yerine koyma) bize eğitim set(ler)ine adaptasyon hakkında bir şeyler söylerken, test hatası gerçek tahmin hatasının bir tahminidir. Bu tahmin, modelin genelleştirme yeteneğini nicelleştirir. Modeli eğitmek için kullanılan aynı veriler modeli değerlendirmek için yeniden kullanıldığından, eğitim hatasının gerçek tahmin hatasını hafife alma eğiliminde olduğuna dikkat edin.

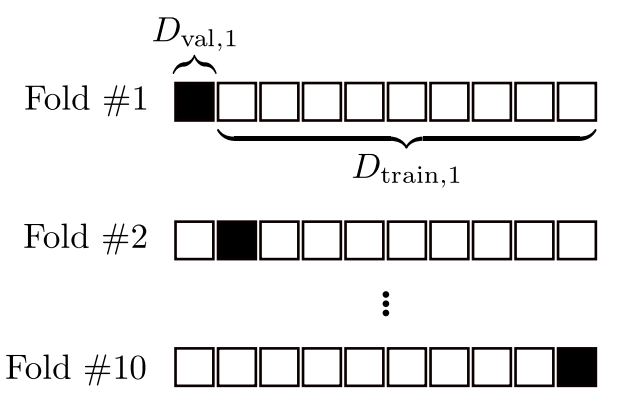
## Tek bekletmeli rasgele alt örnekleme (Single hold-out random subsampling)

Çeşitli veri yeniden örnekleme stratejileri arasında en basitlerinden biri, test seti için öğrenme setinden bazı vakaları rastgele örnekleyen, geri kalan vakaları ise eğitim setini oluşturan tek bekletme yöntemidir. Genellikle, test seti mevcut veri setinin yaklaşık %10 ila %30'unu içerir ve eğitim seti veri setinin yaklaşık %90 ila %70'ini içerir. Öğrenme kümesi yeterince büyükse ve sonuç olarak hem eğitim hem de test kümeleri büyükse, o zaman gözlemlenen test hatası yeni yani görünmeyen durumlar için modelin gerçek hatasının güvenilir bir tahmini olabilir.

## 2.2 k-katlı rasgele alt örnekleme

K-katlı rasgele alt deneyde, tek uzatma yöntemi k kez tekrarlanır, böylece k çift Değitim,j ve Dtest,j , j = 1..k üretilir. Öğrenme fonksiyonu her bir eğitim setine uygulanır ve ortaya çıkan model daha sonra karşılık gelen test setine uygulanır. Performans, tüm k test setlerinin ortalaması olarak tahmin edilmektedir. Herhangi bir eğitim ve test kümesi çiftinin ayrık olduğuna dikkat edin, yani kümelerin herhangi bir ortak durumu yoktur, Değitim,j ∩ Dtest,j = ∅. Ancak, herhangi iki eğitim seti veya iki test seti elbette çakışabilir.

# k katlı çapraz doğrulama

Çapraz doğrulama, tekrarlanan rasgele alt örnekleme yöntemine benzer, ancak örnekleme, iki test seti çakışmayacak şekilde yapılır. K-katlı çapraz doğrulamada, mevcut öğrenme seti, yaklaşık olarak eşit boyutta k ayrık altkümeye bölünür. Burada "k", elde edilen alt kümelerin sayısını ifade eder. Bu bölümleme, öğrenme setinden yer değiştirmeden rastgele örnekleme vakaları ile gerçekleştirilir. Model, birlikte eğitim setini temsil eden k - 1 alt küme kullanılarak eğitilir. Daha sonra model, doğrulama seti olarak adlandırılan kalan alt kümeye uygulanır ve performans ölçülür. Bu prosedür, k adet alt kümenin her biri doğrulama kümesi olarak görev yapana kadar tekrarlanır. k doğrulama setindeki k performans ölçümünün ortalaması, çapraz doğrulama performansıdır. Şekil 2, bu işlemi k = 10, yani 10 kat çapraz doğrulama için göstermektedir. Birinci katta, ilk alt küme Dval,1 doğrulama kümesi olarak hizmet eder ve geri kalan dokuz alt küme Dtrain,1 eğitim kümesi olarak hizmet eder. İkinci katta, ikinci alt küme doğrulama kümesidir ve geri kalan alt kümeler eğitim kümesidir ve bu işlem tüm alt kümeler doğrulama kümesi olana kadar devam eder.

Şekil : 10 katlı çapraz doğrulama Veri seti, her biri (yaklaşık olarak) verinin %10'unu içeren on ayrık alt kümeye rastgele bölünür. Model, eğitim seti üzerinde eğitilir ve daha sonra doğrulama setine uygulanır.

# Sonuç

Çapraz doğrulama, tahmine dayalı bir modelin genelleştirme yeteneğini değerlendirmek ve ezberi önlemek için en yaygın kullanılan veri yeniden örnekleme yöntemlerinden biridir. Gelecekteki gerçek durumların tahmini için nihai modeli oluşturmak için, öğrenme fonksiyonu (veya öğrenme algoritması) f genellikle tüm öğrenme setine uygulanır. Bu son model çapraz doğrulamaya tabi tutulamaz. Model oluşturma aşamasındaki çapraz doğrulamanın amacı, bu son modelin yeni veriler üzerindeki performansı için bir tahmin sağlamaktır.

# KAYNAKLAR

Berrar, D. (2019). Cross-Validation.